HHS Public Access

Author manuscript

Nat Neurosci. Author manuscript; available in PMC 2017 April 03.

Published in final edited form as:

Nat Neurosci. 2016 December ; 19(12): 1563–1565. doi:10.1038/nn.4404.

Ultra-rare disruptive and damaging mutations influence educational attainment in the general population

Andrea Ganna[#1,2,3,4], **Giulio Genovese**[#2,3,5], **Daniel P. Howrigan**[1,2,3], **Andrea Byrnes**[1,2,3], **Mitja Kurki**[1,2,3,6], **Seyedeh M. Zekavat**[2,7], **Christopher W. Whelan**[2,3,5], **Mart Kals**[8,9], **Michel G. Nivard**[10], **Alex Bloemendal**[1,2,3], **Jonathan M. Bloom**[1,2,3], **Jacqueline I. Goldstein**[1,2,3], **Timothy Poterba**[1,2,3], **Cotton Seed**[1,2,3], **Robert E. Handsaker**[2,3,5], **Pradeep Natarajan**[2,7], **Reedik Mägi**[8], **Diane Gage**[3], **Elise B. Robinson**[1,2,3], **Andres Metspalu**[8], **Veikko Salomaa**[11], **Jaana Suvisaari**[11], **Shaun M. Purcell**[12,13], **Pamela Sklar**[12,13], **Sekar Kathiresan**[2,7], **Mark J. Daly**[1,2,3], **Steven A. McCarroll**[2,3,5], **Patrick F. Sullivan**[4,14], **Aarno Palotie**[1,3,6], **Tõnu Esko**[2,8], **Christina Hultman**[4], and **Benjamin M. Neale**[1,2,3]

[1] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston 02114, MA, USA. [2] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [3] Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [4] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm 171 77, Sweden. [5] Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [6] Institute for Molecular Medicine Finland, FIMM, University of Helsinki, Helsinki FI-00014, Finland. [7] Center for Human Genetic Research and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, 02114, USA. [8] Estonian Genome Center, University of Tartu, Tartu 51010, Estonia. [9] Institute of Mathematics and Statistics, University of Tartu, Tartu 50409, Estonia [10] Department of Biological Psychology, VU University Amsterdam, Amsterdam 1081 HV, The Netherlands. [11] Department of Health, THL-National Institute for Health and Welfare, Helsinki FI-00271, Finland. [12] Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA [13] Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [14] Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, North Carolina 27599-7264, USA.

Corresponding author: Andrea Ganna, Analytic and Translational Genetics Unit, Massachusetts General Hospital, Richard B. Simches Research Center, 185 Cambridge Street, CPZN-6818, Boston, MA 02114, aganna@broadinstitute.org.

ACCESSION CODES

URLs. Swedish WES data are available through dbGAP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000473

COMPETING FINANCIAL INTERESTS

I declare that the authors have no competing interests as defined by Springer Nature, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Data availability

Swedish WES data are available trough dbGAP at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000473. The code used in this study is available at: https://github.com/andgan/URV_edu_attainment

Code availability

The code used in this study is available at: https://github.com/andgan/URV_edu_attainment

[#] These authors contributed equally to this work.

Abstract

Disruptive and damaging ultra-rare variants (URVs) in highly constrained (HC) genes are enriched in individuals with neurodevelopmental disorders. In the general population, this class of variants was associated with a decrease in years of education (YOE; −3.1 months; P-value=3.3×10^{-8}). This effect was stronger among high brain-expressed genes and explained more YOE variance than pathogenic copy number variation, but less than common variants. Disruptive and damaging URVs in HC genes influence the determinants of YOE in the general population.

Educational attainment, measured by the highest number of years of education (YOE) attained, is a complex trait influenced by public policy [1], economic resources[2] and many heritable traits, including cognitive abilities and behavior [3]. Importantly, YOE is positively associated with healthy behaviors and lower rates of chronic diseases [4].

Genome-wide association study (GWAS) meta-analyses have identified 162 genome-wide significant loci for YOE [5]. The additive heritability of YOE explained by common genetics variants has been estimated at 21% (95% confidence intervals [CI] 11-31%) [6], which is approximately half of the total heritability estimated from twin studies (40%; 95% C.I. 35-44%) [7]. It has been hypothesized that rare to ultra-rare exonic variants might account for some of the heritability currently not captured by GWAS [8].

Recent studies of intellectual disability, autism and schizophrenia have shed light on the impact of *de novo* and ultra-rare variants (URVs: variants that are observed only once (singletons) in the study and not observed in 60,706 exomes sequenced in the Exome Aggregation Consortium (ExAC) [9]) on the genetic architecture of these disorders [10-12], showing a specific enrichment in highly constrained (HC: genes intolerant to loss-of-function or missense mutations, i.e. having a probability of being loss_of_function intolerant (*pLI)* > 0.9). Moreover, emerging evidence suggests that *de novo* loss-of-function mutations are associated with reduced adaptive functioning in individuals without diagnosis of autism[13].

We tested the hypothesis that a burden of URVs in HC genes is associated with YOE in 14,133 individuals participating in four studies from three Northern European countries: Sweden, Estonia and Finland. Of these, 5,047 individuals have been diagnosed with schizophrenia.

The average numbers of YOE were 13.1, 13.6, and 11.8 in Swedish, Estonian, and Finnish participants, respectively. These differences are partially explained by different age and sex distributions, as well as by different methods used to measure educational attainment (**Supplementary Table 1**).

We observed lower YOE among men (12.8 *vs.* 13.2 years, P-value=4.8×10^{-12}) and older individuals (0.8 month less of education for each additional year of age, P-value < 1×10^{-15}) (**Supplementary Table 2)**.

We developed a new software package called *Hail* to very efficiently perform quality control, annotation and analysis of large-scale sequencing data (**Online Methods**).

We identified URVs in HC genes using whole exome sequencing (WES) data (N. individuals=11,431) and protein coding regions in high-coverage whole genome sequencing (WGS) data (N. individuals=2,702). The primary reason to focus on URVs in HC genes is to maximize the expected deleteriousness of the variants included (due to purifying selection).

Within the set of URVs in HC genes we defined variants that were: (1) disruptive: putative loss-of-function variants including premature stop codons, essential splice site mutations and frameshift indels; (2) damaging: missense variants classified as damaging by seven different *in silico* prediction algorithms (**Online Methods**) and (3) negative control: synonymous variants not predicted to change the encoded protein. We observed one or more of such mutations in 25%, 24% and 78% of individuals, respectively (**Supplementary Table 3**). Principal components of genetic data showed that individuals within each study were of similar ancestry (**Supplementary Fig. 1**).

On average (**Fig. 1**), we observed a 3.1 months reduction in YOE for each disruptive mutation (95% CI: −4.3,-2.0; P-value=3.3×10^{-8}), and similar effect for damaging mutations (2.9 months less YOE; 95% CI: −4.1,-1.7; P-value=1.3×10^{-6}). Furthermore, each additional disruptive mutation on average reduced the chance of going to college by 14% (odds ratio=0.86; 95% CI: 0.78,0.95; P-value=0.0017). These results were consistent when using a mixed linear model approach to correct for population stratification in the Finnish and Estonian samples with WGS data (2.4 months less YOE; 95% CI: −4.3, −0.95; P-value= 0.014, N=2,702).

The negative association between URVs and YOE remained consistent when we examined the control cohort and schizophrenia case cohort separately (**Supplementary Fig. 2**). Furthermore, the effect remained consistent when excluding individuals diagnosed with a neurodevelopmental disorder (i.e. schizophrenia, bipolar disorder, autism, mental retardation and Asperger's syndrome), as identified via linkage with the Swedish national inpatient registry (**Supplementary Fig. 3**). We did not observe any significant association when we restricted our analysis to synonymous variants in HC genes (P-value=0.62) or disruptive mutations in unconstrained genes (P-value=0.73).

We used gene-expression data to determine whether restricting to genes enriched for brain expression concentrated our URVs burden signal. Specifically, we used the Genotype-Tissue Expression consortium data [14] to identify the 20% top brain-expressed HC genes. The intersection between HC and brain-expressed genes (N. genes=683 and 313 for disruptive and damaging URVs, respectively) more than doubled the impact on YOE (6.5 less months of YOE per each additional disruptive variant; 95% CI: −9.6,-3.4; P-value=3.4×10^{-5}; **Fig. 2**). When using increasingly liberal thresholds for defining genes enriched for brain-expression, we saw a consistent decrease in the association (**Supplementary Fig. 4**). The association was not significant when considering non brain-enriched HC genes or all brain-enriched genes (P-value > 0.05). We further examined a subset of genes for which basal gene-expression was at least two fold higher in the brain compared to other tissues (brain-

enriched HC genes; **Supplementary Fig. 5**). Although, the impact on YOE was higher for brain-enriched HC genes than non brain-enriched HC genes, the signal was specific to disruptive variants. Overall, this approach was less effective in identifying a HC gene subset impacting YOE.

To place disruptive and damaging URVs into context, we also examined the impact of previously reported genetic influences on YOE, including a polygenic score from common variants [5], runs of homozygosity [15] and a burden of rare pathogenic copy number variants (CNVs) [16]. We sought to establish if these different forms of genetic variation act independently on YOE. For this purpose we defined four scores: (1) a polygenic score including all the independent single nucleotide polymorphisms (SNPs) with P-value for association with YOE < 1 (as this threshold has been shown to maximize variance explained in YOE) in a large GWAS consortia of YOE [5] (2) the summed length of all runs of homozygosity (3) burden of disruptive and damaging URVs in HC genes and (4) burden of self-curated list of pathogenic CNVs from the literature (**Supplementary Table 4**). The polygenic score was only calculated in the Swedish samples (N=10,644), since the other three studies were included in the original GWAS of YOE.

We first explored the association between each genetic score and YOE separately. The strongest change in YOE was observed among CNV carriers (−7.6 months less YOE; 95% CI: −13.7,-1.5; P-value= 0.015). However, these events were rare in the population (161 carriers among 11,999 individuals with CNV measured).

We then fit the four normalized scores in the same regression model to assess the relative contribution of each genetic class to YOE. All four scores were independently associated with YOE (**Fig. 3**). The polygenic score showed the strongest association in standard deviations from the mean, explaining the largest proportion of the variability in YOE (2.9% *vs* 0.4% for the ultra-rare variants, 0.2% for runs of homozygosity and 0.1% for pathogenic CNVs).

We further evaluated whether the association between the polygenic score and YOE changes in individuals with and without disruptive or damaging URVs or CNVs. We found that the polygenic score was more strongly associated with YOE in individuals without disruptive or damaging URVs or CNVs (8.2 vs. 6.2 more months of YOE for 1 standard deviation increase in the polygenic score; P-value for interaction=0.007, **Supplementary Fig. 6**).

We sought to identify individual genes driving the observed association between disruptive and damaging URVs and YOE. Using a gene-based burden test implemented in SKAT [17], and using an exome-wide significance threshold of 1×10^{-6}, we didn't identify any statistically significantly associated gene (**Supplementary Fig 7, upper panels**). Similar results were observed when we included all variants with minor allele frequency < 0.05%, rather than only URVs (**Supplementary Fig 7, lower panel**s).

In this study we focused on YOE, a phenotype that is relatively easy to collect in large samples and which has a strong genetic correlation with intelligence and cognitive function [6,18]. We integrated WGS, WES and array data on more than 14,000 individuals and described the impact of URVs disrupting HC genes on YOE. This class of variants have

been previously associated with autism[3] and schizophrenia[4], but the impact on YOE in the general population has not been described before. Here, for the first time, we show that disruptive and damaging URVs in HC genes are likely to affect factors underlying education attainment among individuals not diagnosed with psychiatric or neurodevelopmental disorders. Exploring the extent to which this association is mediated by cognitive-related determinants of YOE, or by other non-cognitive factors will require studies integrating detailed cognitive, psychological and personality measurements. Similar to the analyses of schizophrenia[10] and autism, the majority of the signal lies in genes highly expressed in brain. This observation does not exclude the existence of causal mutations outside this gene class, but suggests that strong acting mutations are heavily concentrated within these genes.

Furthermore, we show that disruptive and damaging URVs in HC genes, common variants associated with YOE, runs of homozygosity, and pathogenic CNVs, all act on cognitive function or personality traits ultimately reflected in the educational attainment of our study participants. This effect was not simply additive. We identified a modest, but significant interaction between the polygenic score and the presence of URVs or CNVs. Whether this observation is driven by the interplay of partially overlapping pathways between common and rare variants or by genotype-phenotype heterogeneity (e.g. common and rare variants impacting different subsets of individuals) will be a matter of future investigation.

We report that, on average, an additional disruptive URVs in HC genes results in a 3.1 months reduction in YOE. This effect is likely to be a mixture of variants with larger effect and variants that are not associated with YOE. The polygenic score based on common variants effect sizes estimated from a much larger cohort of 405,072 individuals explained a larger fraction of the YOE. This is not surprising, given that common variants are expected to have the largest contribution to heritable variation in most complex traits [19].

The prioritization approaches used to select variants contributing to the score from common variants and the score from rare variants are different. The former uses estimates of the association with YOE and the proportion of variance explained by the score is likely to improve once the sample size used to originate these estimates increases. The latter uses *in-silico* prediction of the variants' functional effect coupled with population genetics expectations built on the mutation rate. As with the common variant score, we expect that the score based on URVs in selected gene sets will continue to improve in predictive validity of YOE as a more precise characterization of which genes and genomic regions are associated with YOE emerges.

Our study could not detect disruptive or damaging mutations in a given gene as being unequivocally associated with YOE; however, as sample sizes increase, specific genes will emerge. Nevertheless, our proof-of-concept work shows that a wide range of genetic variation from URVs and CNVs to common variants influence determinants of YOE in the population.

ONLINE METHODS

Studies description and selection

In this study we used epidemiological studies with YOE and exome or whole-genome sequencing information available, no formal power calculation was done.

Ethical committees in Sweden, Estonia and Finland approved all procedures and all subjects provided written informed consent (or legal guardian consent and subject assent).

Sweden-WES—A total of 12,384 blood-derived DNA samples from Swedish research participants were collected from 2005 to 2013. Psychiatric cases with a diagnosis of schizophrenia were ascertained from the Swedish National Hospital Discharge Register. The register is complete from 1987 and augmented by psychiatric data from 1973-86. It contains dates and ICD discharge diagnoses (World Health Organization, 1992) for each hospitalization, and captures the clinical diagnosis made by the attending physician. Case inclusion criteria: ≥2 hospitalizations with a discharge diagnosis of schizophrenia, both parents born in Scandinavia, and age ≥18 years. Case exclusion criteria: hospital register diagnosis of any medical or psychiatric disorder mitigating a confident diagnosis of schizophrenia as determined by expert review, and included removal of 3.4% of eligible cases due to the primacy of another psychiatric disorder (0.9%) or a general medical condition (0.3%) or uncertainties in the Hospital Discharge Register (e.g., contiguous admissions with brief total duration, 2.2%). The validity of this case definition of schizophrenia is strongly supported as described in [20]. Controls were selected at random from Swedish population registers. Control inclusion criteria: never hospitalized for schizophrenia or bipolar disorder (given evidence of genetic overlap with schizophrenia), both parents born in Scandinavia, and age ≥18 years.

Estonia-WGS—Estonian-WGS samples are the subset of the Estonian Biobank of the Estonian Genome Center at the University of Tartu [21]. It is a population-based biobank, containing almost 52,000 samples of the adult population (aged ≥18 years), which closely reflects the age, sex and geographical distribution of the Estonian population. All subjects have been recruited randomly by general practitioners or physicians in hospitals throughout the country. The participants donated blood samples for DNA, white blood cells and plasma tests and filled the Computer Assisted Personal Interview (CAPI).

In total, 2,300 geographically diverse samples have whole genome sequencing data, selected randomly by county of birth.

Finnish-WES and Finnish-WGS—All of the Finnish individuals are part of the FINRISK cohort, a national survey on risk factors of chronic and non-communicable diseases in Finland [22]. The survey has been conducted every five years since 1972 in randomly selected, representative population samples from different parts of Finland. All of the samples are from FINRISK 1992, 1997, 2002 and 2007 surveys.

Finnish-WES mainly includes individuals that are part of an IBD case-control study, where controls were selected to have a high IBD polygenic risk score [23].

Finnish-WGS includes schizophrenia cases and controls selected using nationwide hospital discharge registry and/or nationwide medicine reimbursement registry where all psychosis cases or psychosis medication purchases are systematically recorded. Controls were selected to have high polygenic risk score for schizophrenia [24].

Phenotype definition

We matched the original educational categories with the International Standard Classification of Education (ISCED), as described in **Supplementary Table 1**. Thereafter we used the equivalent of United States years of schooling to obtain the YOE. Going to college was defined as having an ISCED category > 4.

To remove potential bias introduced by uncompleted education, we excluded all the individuals younger than 30 years at the time of sample collection. For the Estonian and Finnish samples, we used self-report data; whereas for the Swedish sample, we obtained YOE from the national registries. YOE was approximately normally distributed.

Sequencing procedures

Estonian WGS and Finnish WGS samples have been sequenced at Broad Institute on Illumina HiSeq X Ten machines run to 20x and 30x mean coverage (150bp paired reads), respectively. Estonian samples followed a PCR-free sample preparation. Swedish-WES and Finnish-WES samples were sequenced using either the Agilent SureSelect Human All Exon Kit or the Agilent SureSelect Human All Exon v.2 Kit. Sequencing was performed at Broad institute on Illumina GAII, Illumina HiSeq2000 or Illumina HiSeq X Ten. Mean target coverage was 90x.

All samples have been aligned against the GRCh37 human genome reference and BAM processing was carried out using BWA Picard. Genotype calling was done using GATK Haplotype Caller and was performed at Broad Institute for all studies.

Hail software

To overcome the growing computational challenge of learning from large genomic datasets, we utilized Hail, an open-source software framework for scalably and flexibly analyzing such data (https://github.com/broadinstitute/hail). Hail, under active development, includes support for data import/export, quality control, analysis of population structure, and methods for performing both common and rare variant association. Hail is written in Scala (a Java virtual machine language) and builds on open-source software for scalable distributed computing including Hadoop (http://hadoop.apache.org/) and Spark (http://spark.apache.org/). Hail achieves near-perfect scalability for many tasks and can run on thousands of nodes. Hail automates fault-tolerant distribution of data and compute, greatly simplifying distributed pipeline execution compared to traditional HPC job schedulers like LSF and Grid Engine. Pipelines written in Hail's high-level language typical require orders-of-magnitude fewer lines of code than comparable pipelines written in general purpose languages.

Samples and variants QC

Quality control was performed independently for each study using *Hail*. We excluded individuals with high proportion of chimeric reads (>5%), high contamination (>5%) or an excessive number of singletons variants not observed in ExAC (> 100 for WES and > 20,000 for WGS). We included only unrelated individuals (IBD proportion < 0.2) and those for whom the sex predicted from genetic data matched the self-reported gender. We kept only 'PASS' variants, as determined by The Genome Analysis Toolkit [25] Variant Quality Score Recalibration (VQSR) filter, are set to missing variants with GQ < 20 and allele balance > 0.8 or < 0.2. We further excluded variants with call rate < 0.8. In WGS data, we excluded low complexity regions as defined by Li [26]. In the burden test analysis we excluded variants with both Hardy-Weinberg equilibrium test P-value < 1×10^{-6} and negative inbreeding coefficient (expected heterozygosity less than observed heterozygosity).

Annotation and URVs scores definition

Annotation was performed using SnpEff 4.2 (build 2015-12-05) [27] using Ensemble gene models from database GRCh37.75. We further annotated variants with SnpSift 4.2 (build 2015-12-05) [28] using annotations from database dbNSFP 2.9 [29]. In **Supplementary Table 3** we have provided a detailed description of the criteria used for selecting variants in each score. The set of HC genes was defined separately for disruptive and damaging variants. For disruptive and synonymous mutations we defined HC genes those having a probability of being loss_of_function intolerant (*pLI)* > 0.9 (N genes=3,488). For missense damaging mutation we used a missense *z-score* > 3.09 (N genes=1,614) [30]. Both measures have been previously described [30] and available online at ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3/functional_gene_constraint. We used a version derived from The Exome Aggregation Consortium without cases of psychiatric disorders.

Principal component analysis and mixed models

We used a subset of high confidence SNPs to calculate principal components. We selected variants with minor allele frequency larger than 5%, call rate > 90%, Hardy-Weinberg equilibrium test P-value > 1×10^{-6} and we pruned for variants in linkage disequilibrium using *plink* with command line '--indep 50 5 2'.

We used a similar approach to filter variants used to generate the genetic relationship matrix (GRM). We then fit a liner mixed model including the GRM as random effect and age, sex, year of birth, (year of birth – 1950)^2, (year of birth – 1950)^3, the number of singletons synonymous variants not in ExAC and the number of URVs in HC genes as fixed effects.

Association between URVs and educational attainment

We fit a linear regression model where the dependent variable was YOE and the independent predictors were: age, sex, year of birth, (year of birth – 1950)^2, (year of birth – 1950)^3, the 10 first principal components, the number of singletons synonymous variants not in ExAC, schizophrenia status (only in studies including schizophrenic patients) and the URV score (count of disruptive, damaging or synonymous URVs). We adjust for the number of all ultra-rare synonymous variants to correct for potential technical artifacts. We observed similar

results when adjusting for number of ultra-rare synonymous variants + number of ultra-rare disruptive (or damaging) variants in HC genes.

Brain-expressed and brain-enriched HC genes analysis

Using the Genotype-Tissue Expression consortia (GTEx) data [14], we ranked gene-expression levels (in RPKM) in brain tissues and defined the top 20% HC genes as "brain-expressed" (N. genes=683 and 313 for disruptive and damaging, respectively). Conversely, we defined "non brain expressed" the bottom 20% of the HC genes (N. genes=683 and 313 for disruptive and damaging, respectively).

We also compute estimated fold-change in the brain as follows. Suppose samples *1, 2, ..., N_b* are brain samples and samples *(N_b+1), (N_b+2), ... , N* are the samples from other tissues. Denote with *xij* the expression of gene *j* and sample *i*, in reads per kilobase of transcript per million (RPKM). We compute fold-change (FC):

$$FC_j = \frac{\frac{1}{N_b}\sum_{i=1}^{N_b} x_{ij}}{\frac{1}{N}\sum_{k=1}^{N} x_{kj}} = \frac{mean\,(x_j|brain)}{mean\,(x_k)}$$

We label the genes *j*, such that $FC_j > 2$ as "brain-enriched genes" and $FC_j < 0.5$ as "non-brain-enriched genes". The number of brain-enriched HC genes was 447 and 287 for disruptive and damaging mutations, respectively. The number of non brain-enriched HC genes was 2,225 and 935 for disruptive and damaging mutations, respectively.

Polygenic score, CNVs and runs homozygosity

The polygenic score for YOE was obtained from array data in the Swedish WES study (quality control for the array data have been previously described [20]) and directly from WGS data in the Finnish-WGS and Estonian-WGS studies. We included all the independent markers with P-value < 1 in largest GWAS of educational attainment [5] and obtained the polygenic score as weighted sum of risk alleles using the --*score* command in Plink [31].

CNVs for the Swedish WES study were called as part of a separate project [32] using a composite pipeline comprising the CNV callers PennCNV, iPattern, Birdsuite and C-Score organized into component pipelines. We considered only rare CNVs by filtering out all CNVs that present at ≥ 1% allele frequency. CNVs < 20kb or having fewer than 10 probes were also excluded. We used the plink --*cnv-intersect* function with a value of 0.5 to determine the overlap between detected CNVs and the list of pathogenic CNVs reported in **Supplementary Table 4**.

CNVs in Finnish WGS and Estonian WGS were genotyped according to the methods described in [33] and implemented in Genome STRiP 2.0. Briefly, read depth information was collected from WGS data, excluding regions of the genome that are not uniquely alignable or have low sequence complexity, and adjusted for GC content bias. Each CNV reported in **Supplementary Table 4** was directly genotyped using Genome STRiP's genotyping

module, which examines the read depth across all samples and fits a constrained Gaussian mixture model with components representing each possible diploid copy number and sample-specific variance terms to account for differences in sequencing depth.

The summed runs of homozygosity were determined using the same pipeline described in [15]. Specifically we used *plink* with command line '--homozyg --homozyg-window-snp 35 --homozyg-snp 35 --homozyg-kb 1500 --homozyg-gap 1000 --homozyg-density 250 --homozyg-window-missing 5 --homozyg-window-het 1'.

Gene-based burden test

We first extracted from each dataset variants falling within UCSC known genes and merged the four datasets using *plink*. If a variant was not present in all cohorts, we forced it as homozygous reference across the remaining cohorts (using "--fill-missing-a2" option in *plink*). We then computed principal components for the combined dataset after further merging with 1000 Genomes project samples as described in (Genovese et al, *jointly submitted).* To test the hypotheses that disruptive URVs in individual genes were associated with YOE and college status, we performed a burden test [34] using the SKAT software [35] using default parameters (method=davies, impute.method=bestguess, r.corr=1.0), adjusting for age, sex, year of birth, (year of birth – 1950)^2, (year of birth – 1950)^3, the first 10 principal components, schizophrenia status and number of URVs identified in coding regions. We used a python wrapper to run the SKAT software (available at https://github.com/freeseek/gwaspipeline).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENT

We want to thank Dr. Raymond Walters for the insightful discussion. Dr. Andrea Ganna is supported by the Knut and Alice Wallenberg Foundation (2015.0327) and the Swedish Research Council (2016-00250). This study was supported by grants from the National Human Genome Research Institute (U54 HG003067, R01 HG006855), the Stanley Center for Psychiatric Research, the Alexander and Margaret Stewart Trust, the National Institute of Mental Health (R01 MH077139 and RC2 MH089905), and the Sylvan C. Herman Foundation. Michel G. Nivard is supported by Royal Netherlands Academy of Science Professor Award (PAH/6635) to Dorret I. Boomsma. Veikko Salomaa was supported by the Finnish Foundation for Cardiovascular Research.

REFERENCES

1. McLendon MK, Perna LW. State Policies and Higher Education Attainment. The ANNALS of the American Academy of Political and Social Science. 2014; 655:6–15.
2. Haveman R, Wolfe B. The Determinants of Children's Attainments: A Review of Methods and Findings. Journal of Economic Literature. 1995; 33:1829–1878.
3. Krapohl E, et al. The high heritability of educational achievement reflects many genetically influenced traits, not just intelligence. Proc Natl Acad Sci U S A. 2014; 111:15273–8. [PubMed: 25288728]
4. Cutler DM, Lleras-Muney A. Education and Health: Evaluating Theories and Evidence. National Bureau of Economic Research Working Paper Series No. 12352. 2006
5. Okbay A, et al. Genome-wide association study identifies 74 loci associated with educational attainment. Nature advance online publication. 2016

6. Marioni RE, et al. Molecular genetic contributions to socioeconomic status and intelligence. Intelligence. 2014; 44:26–32. [PubMed: 24944428]
7. Branigan AR, McCallum KJ, Freese J. Variation in the Heritability of Educational Attainment: An International Meta-Analysis. Social Forces. 2013
8. Zuk O, et al. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A. 2014; 111:E455–64. [PubMed: 24443550]
9. Lek M, et al. Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv. 2015
10. Genovese G. Nature Neuroscience. 2016
11. Gilissen C, et al. Genome sequencing identifies major causes of severe intellectual disability. Nature. 2014; 511:344–7. [PubMed: 24896178]
12. Iossifov I, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014; 515:216–21. [PubMed: 25363768]
13. Robinson EB, et al. Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. Nat Genet. 2016
14. Consortium G. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013; 45:580–5. [PubMed: 23715323]
15. Joshi PK, et al. Directional dominance on stature and cognition in diverse human populations. Nature. 2015; 523:459–62. [PubMed: 26131930]
16. Stefansson H, et al. CNVs conferring risk of autism or schizophrenia affect cognition in controls. Nature. 2014; 505:361–6. [PubMed: 24352232]
17. Wu MC, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; 89:82–93. [PubMed: 21737059]
18. Davies K. What is effective intervention?--using theories of health promotion. Br J Nurs. 2006; 15:252–6. [PubMed: 16607253]
19. Yang J, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet. 2015; 47:1114–20. [PubMed: 26323059]

METHODS-ONLY REFERENCES

20. Ripke S, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet. 2013; 45:1150–9. [PubMed: 23974872]
21. Leitsalu L, et al. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int J Epidemiol. 2015; 44:1137–47. [PubMed: 24518929]
22. Vartiainen E, et al. Thirty-five-year trends in cardiovascular risk factors in Finland. Int J Epidemiol. 2010; 39:504–18. [PubMed: 19959603]
23. Jostins L, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012; 491:119–24. [PubMed: 23128233]
24. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014; 511:421–7. [PubMed: 25056061]
25. DePristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43:491–8. [PubMed: 21478889]
26. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014; 30:2843–51. [PubMed: 24974202]
27. Cingolani P, et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet. 2012; 3:35. [PubMed: 22435069]
28. Cingolani P, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012; 6:80–92. [PubMed: 22728672]
29. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat. 2013; 34:E2393–402. [PubMed: 23843252]
30. Samocha KE, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014; 46:944–50. [PubMed: 25086666]

31. Chang CC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015; 4:7. [PubMed: 25722852]

32. Marshall C, et al. A contribution of novel CNVs to schizophrenia from a genome-wide study of 41,321 subjects. bioRxiv. 2016

33. Handsaker RE, et al. Large multiallelic copy number variations in humans. Nat Genet. 2015; 47:296–303. [PubMed: 25621458]

34. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009; 5:e1000384. [PubMed: 19214210]

35. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet. 2013; 92:841–53. [PubMed: 23684009]

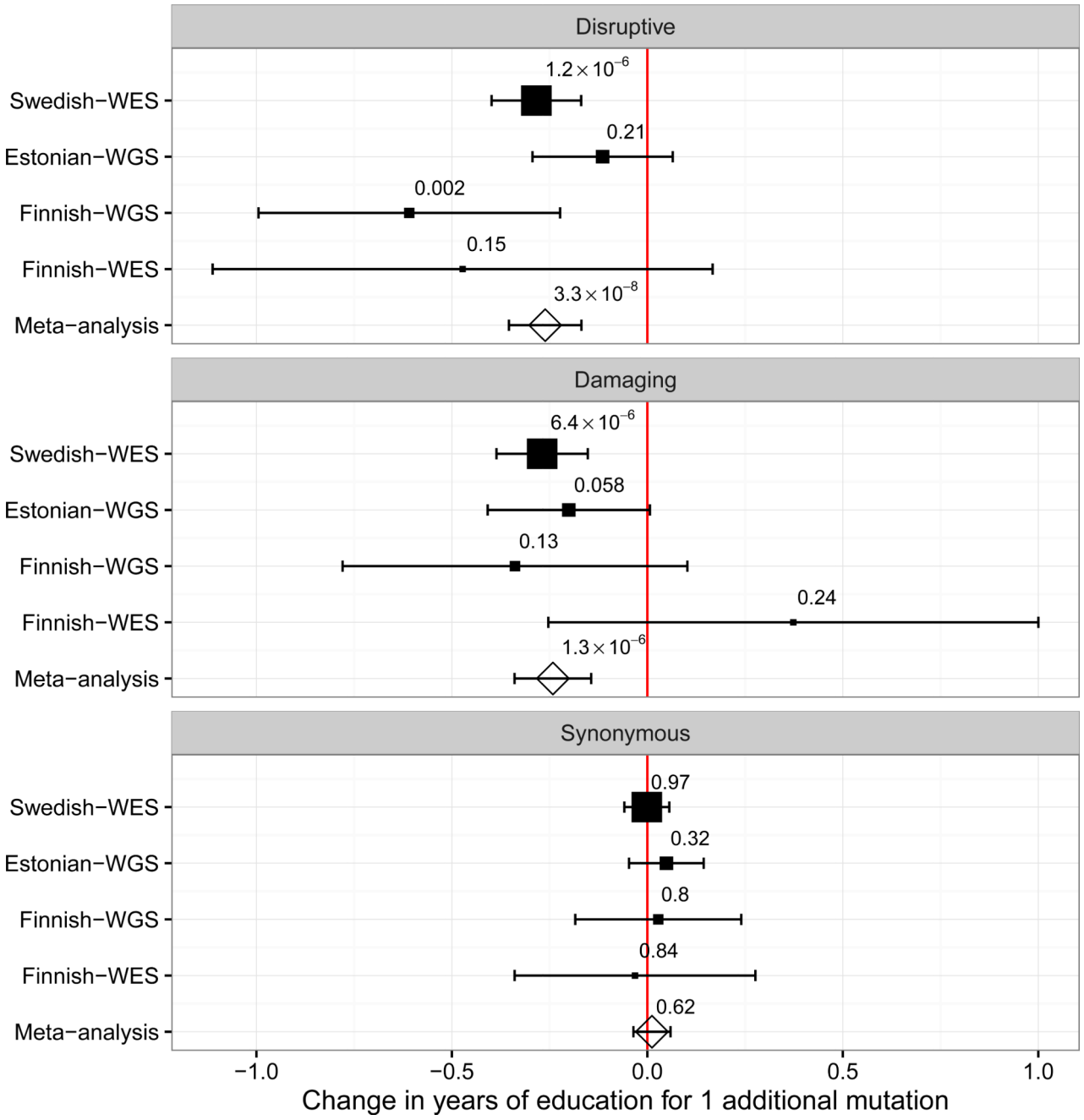

Figure 1.

Association between number of disruptive, damaging, and synonymous URVs in HC genes and YOE. Disruptive and damaging, but not synonymous URVs are significantly associated with reduced YOE. The size of the squares is proportional to the size of the study. The horizontal bars represent 95% confidence intervals. All the estimates are obtained from a linear regression model. Meta-analysis results are obtained using a fixed-effect approach.

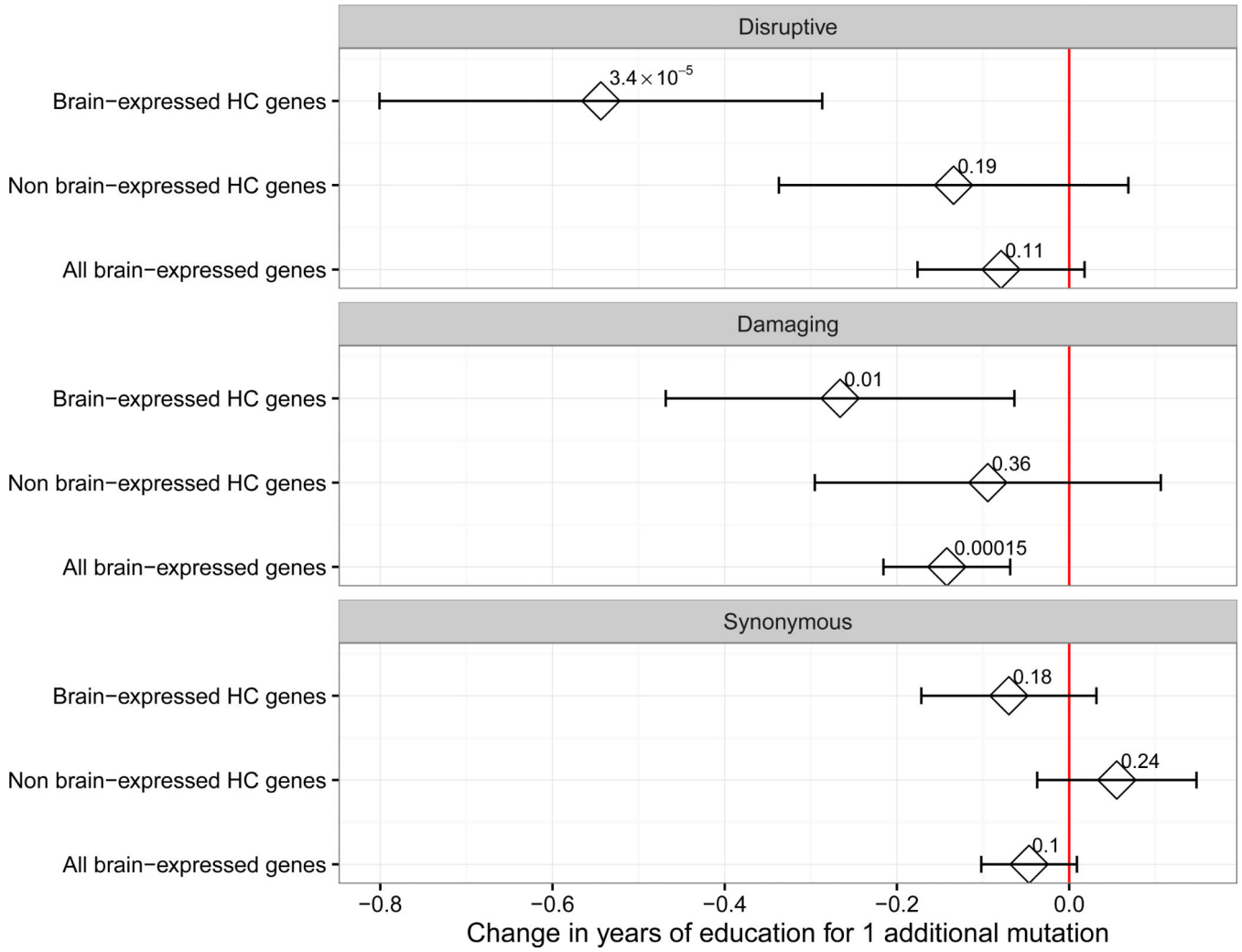

Figure 2.

Association between numbers of disruptive, damaging and synonymous URVs for different gene sets. The intersection between HC and brain-expressed genes yield the strongest reduction in YOE. We only report the meta-analysis results (N=14,133). The horizontal bars represent 95% confidence intervals. All the estimates are obtained from a linear regression model and combined using fixed-effect meta-analysis.

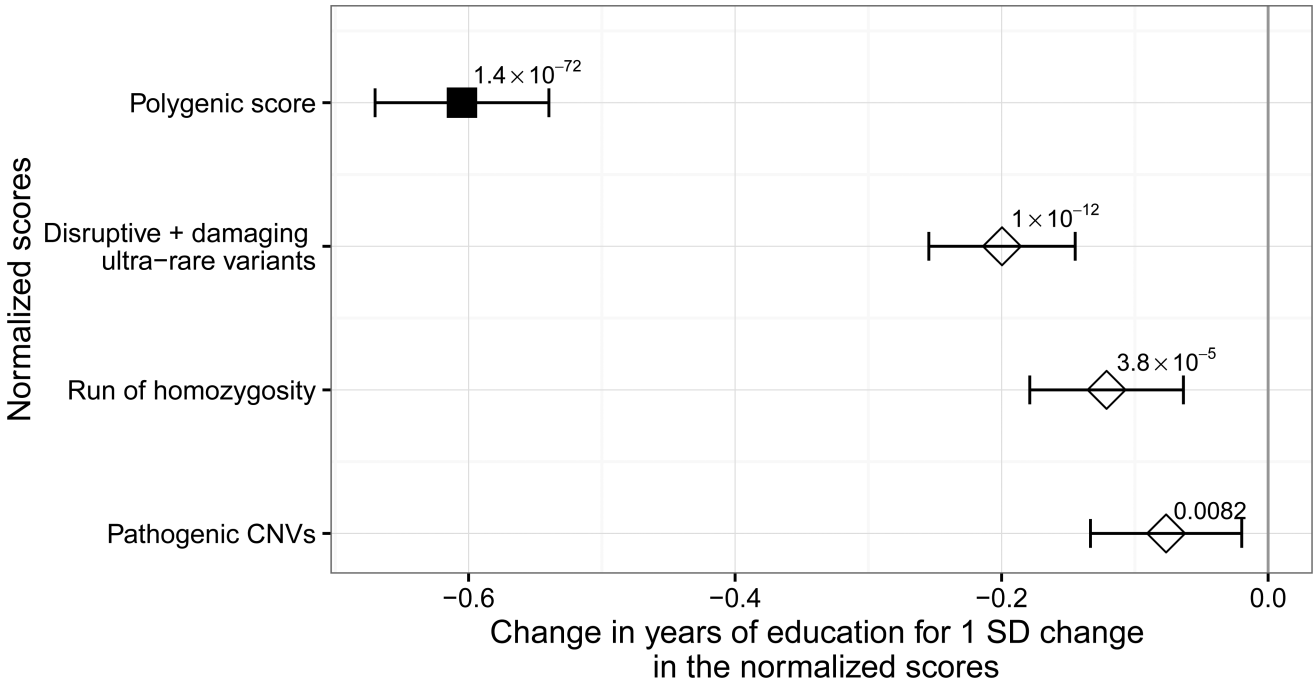

Figure 3.

Association between each of the normalized scores (polygenic, runs of homozygosity, URVs and pathogenic CNVs) and YOE. The results presented are from meta-analysis of Swedish WES, Estonian WGS and Finnish WGS studies (N=13,353), except for the polygenic score, which is calculated only in the Swedish WES study (N=10,651). Notice that we plot 1-polygenic score to obtain a negative association with YOE. The horizontal bars represent 95% confidence intervals. All the estimates are obtained from a linear regression model and combined using fixed-effect meta-analysis.