Cognitive performance among carriers of pathogenic copy number variants: Analysis of 152,000 UK Biobank subjects

Short title: Cognitive performance of CNV carriers in the UK Biobank

Kimberley M Kendall[1] MRCPsych, Elliott Rees[1], PhD, Valentina Escott-Price[1], PhD, Mark Einon[1], Rhys Thomas[1] PhD, MRCP, Jonathan Hewitt[2] PhD, MRCP, Michael C O'Donovan[1] PhD, FRCPsych, Michael J Owen[1] PhD, FRCPsych, James Walters[1] PhD, MRCPsych, George Kirov[1]* PhD, MRCPsych.

[1] MRC Centre for Neuropsychiatric Genetics & Genomics, Cardiff University School of Medicine, Division of Psychological Medicine and Clinical Neuroscience, Hadyn Ellis Building, Maindy Road, Cardiff, CF24 4HQ, UK

[2] Department of Geriatric Medicine, Division of Population Medicine, 3rd Floor Academic Centre, Llandough Hospital, Penlan Road, Penarth, CF64 2XX

* Corresponding author - E-mail: kirov@cardiff.ac.uk; tel: +44 20 20688465

Key words: CNV; UK Biobank; schizophrenia; cognition; neurodevelopmental; Affymetrix

Abstract

Background: The UK Biobank is a unique resource for biomedical research, with extensive phenotypic and genetic data on half a million adults from the general population. We aimed to examine the effect of neurodevelopmental copy number variants (CNVs) on the cognitive performance of participants.

Methods: We used Affymetrix Power Tools and PennCNV-Affy software to analyse Affymetrix microarrays of the first 152,728 genotyped individuals. We annotated a list of 93 CNVs and compared their frequencies with control datasets. We analysed the performance on seven cognitive tests of carriers of 12 CNVs associated with schizophrenia (N = 1,087) and of carriers of another 41 neurodevelopmental CNVs (N = 484).

Results: The frequencies of the 93 CNVs in the Biobank subjects were remarkably similar to those among 26,628 controls from other datasets. Carriers of schizophrenia-associated CNVs and of the group of 41 other neurodevelopmental CNVs had impaired performance on the cognitive tests, with nine of 14 comparisons remaining statistically significant after correction for multiple testing. They also had lower educational and occupational attainment (p-values between 10^{-7} and 10^{-18}). The deficits in cognitive performance were modest (z score reductions between 0.01 and 0.51), compared to individuals with schizophrenia in the Biobank (z score reductions between 0.35 and 0.90).

Conclusions: This is the largest study on the cognitive phenotypes of CNVs to date. Adult carriers of neurodevelopmental CNVs from the general population have

significant cognitive deficits. The UK Biobank will allow unprecedented opportunities for analysis of further phenotypic consequences of CNVs.

Introduction

Copy number variants (CNVs) are >1,000 base pair DNA segments that are present at a variable copy number in comparison with a reference genome (1, 2). CNVs may be recurrent and have similar breakpoints when they are formed by non-allelic homologous recombination between sites of low copy repeats, or non-recurrent, with variable breakpoints, when formed as a result of defects in DNA replication or repair (3). There is an increased rate of CNVs in neurodevelopmental spectrum disorders including intellectual disability (ID), autism spectrum disorder (ASD) (4, 5), epilepsy (6) and schizophrenia (7). To date, 12 CNVs have been robustly associated with risk of schizophrenia and they also increase risk of ASD and ID (7, 8). Many more CNVs are implicated in ID, ASD and cases with congenital anomalies (4, 5) but have not been implicated in schizophrenia, although this could be due to insufficient statistical power (8).

The phenotypes of many highly penetrant CNVs, such as those implicated in Prader-Willi/Angelman or DiGeorge Syndrome, are well established. Many others have incomplete penetrance (9) and their phenotypic spectrum is not fully established (e.g. 1q21.2 duplication, 15q11.2 deletion). There are many adult carriers of incompletely penetrant CNVs in the general population, who have escaped the development of early-onset developmental disorders and are apparently healthy. However, they might still have an increased burden of cognitive or physical impairments.

Limited data, based on relatively small sample sizes, are available on the cognitive phenotypes of CNVs in adults. An Icelandic study of 144 carriers of 11 pathogenic CNVs found that healthy CNV carriers had impaired cognition and performed intermediately between non-CNV carriers and individuals with schizophrenia (10). An Estonian study of 56 CNV carriers found an association between rare CNVs and lower educational attainment (11).

The UK Biobank provides a great opportunity to study the effects of pathogenic CNVs on physical and mental health characteristics, especially the incompletely penetrant ones. The half a million individuals recruited by the Biobank have provided extensive demographic, health and cognitive data, will be followed up prospectively and are being genotyped on Affymetrix microarrays. We aimed to identify pathogenic CNVs in the first 152,728 individuals in the UK Biobank for whom genotype data have been released so far and to analyse the cognitive consequences of neurodevelopmental CNVs.

Methods and Materials

Participants: The UK Biobank (www.ukbiobank.ac.uk) recruited half a million participants in the UK between 2006 and 2010, aged 40-69 years, 53% female. All subjects provided informed consent to participate in UK Biobank projects and agreed to have their health followed over many years. Ethical approval for the study was granted by the North West multi-centre ethics committee. The Biobank used 22 assessment centres in England, Scotland and Wales. Participants were recruited from National Health Service (NHS) patient registers, with no exclusion criteria, provided they lived within a reasonable proximity to an assessment centre. Participants spent approximately three hours at the assessment centres, providing detailed demographic, socioeconomic and health related data via a touch screen questionnaire. Following this, a trained nurse performed an interview to clarify any questions arising from the touch screen questionnaire. Participants underwent several physical assessment measures and provided blood, urine and saliva samples. Data were released to Cardiff University following application to the UK Biobank.

Genotyping: Samples were genotyped at Affymetrix Research Services Laboratory, Santa Clara, CA. Around 100,000 samples were genotyped on the UK Biobank Axiom Array (820,967 probes) and ~50,000 samples were genotyped on the UK BiLEVE Array (807,411 probes) (12). There is 95% common content between the two arrays (https://biobank.ctsu.ox.ac.uk/crystal/docs/genotyping_sample_workflow.pdf). Sample processing at UK Biobank is described in the Supplementary Material. Here we present data from 152,728 individuals genotyped in the first phase of genotyping, of which 151,659 passed our quality control (QC) filters (Table S1).

CNV Calling: In our previous work on CNVs, we analysed datasets of up to 20,000 samples at a time (13-16). In order to call CNVs in the UK Biobank, we had to substantially increase the speed of processing, by analysing batches in parallel, standardising QC processes across all batches, and omitting z score analysis of CNVs, re-clustering after removal of poorly performing samples, and manual inspection of CNV traces. Anonymised genotype data were downloaded as raw (CEL) files from the UK Biobank website, stored on a secure Linux server and analysed with UNIX-based commands (detailed in the Supplementary Material, Section 2). Briefly, we used the *apt-probeset-genotype* command with Affymetrix Power Tools (APT) software (www.affymetrix.com/estore/partners_programs/programs/developer/tools/powertools.affx) to generate normalised signal intensity data, genotype calls and confidences. We analysed only the ~750,000 biallelic markers as these can be used in PennCNV-Affy to generate cluster plots. Each pre-defined batch of ~4,600 CEL files was analysed separately, to reduce potential batch effects. The genotype calls, confidences and summary files were processed with PennCNV-Affy software (17). We generated canonical genotype clusters, Log R Ratios (LRR) and B Allele Frequencies (BAF) and completed the subsequent steps recommended in PennCNV for the generation of CNV calls (17). Following CNV detection, adjacent CNVs were joined if separated by <25% of their combined length. Individual samples were excluded if they had 30 or more CNVs, had a waviness factor (WF) >0.03 or <-0.03 or a call rate <96%. Individual CNVs were excluded if they were covered by <10 probes or had a density coverage of <1 probe per 20,000 bp. A total of 1,069 samples (0.7%) were excluded during QC (Table S1).

CNV Annotation: We compiled a list of 93 CNVs proposed to be pathogenic in two widely accepted sources (4, 5). The full list with the corresponding critical region coordinates is presented in Table S2. The breakpoints of the initially called CNVs were inspected to confirm that they met our CNV calling criteria (Table S4). Briefly, we required a CNV to cover >50% of the critical interval and to include the key genes in the region (if known), or in the case of single gene CNVs (e.g. *NRXN1*) we required deletions to intersect at least one exon and duplications to cover the whole gene. For comparison of CNV frequencies with previous results, we used data from 26,628 control individuals from other large datasets, for which we have access to the raw CNV data and used the same CNV calling methods (Table S3) (7, 18, 19). As not all 93 of these CNVs are known to affect cognition (or such link has not been statistically confirmed), for our analyses we selected 53 of these CNVs that have been statistically associated with neurodevelopmental phenotypes (4), (after removing the common duplications at 15q11.2, as they would account for over half of the "other neurodevelopmental CNVs", thus skewing all analyses). The 53 CNVs were further subdivided into 12 CNVs associated with schizophrenia (7, 8) and the remaining 41, denoted as "other neurodevelopmental CNVs" (Table S2).

Cognitive Tests: Participants completed a battery of cognitive tests organised by the UK Biobank. We chose to analyse tests done by at least 10% of participants and restricted the analysis to individuals who self-reported as being of white British or Irish descent. Different numbers of participants were asked by the Biobank to complete the various tests, ranging from nearly the complete sample (Pairs Matching, Reaction Time), to only parts of the sample (the remaining tests). The scores were first normalised, if not normally distributed, and then converted to z scores (Supplementary material, Section 7). We analysed data on the following tests:

Pairs Matching Test (testing episodic memory, completed at the assessment centres at the first visit by 136,292 individuals with available genotypes). Participants were shown 6 pairs of cards for 3 seconds, which were then turned over. Participants were asked to identify the matching pairs. We used the total number of errors made during this task and restricted our analyses to individuals who finished the test. We applied a log +1 transformation to these data.

Reaction Time Test (simple processing speed, completed at the assessment centres at the first visit by 138,603 individuals with available genotypes). Participants were asked to play 12 rounds of a computerised 'Snap' game where they had to click a button as quickly as possible when shown two matching cards. We used the mean reaction time from their attempts. The data were log transformed.

Fluid Intelligence Test (reasoning and problem solving, completed at the assessment centres at the first visit by 44,575 individuals with available genotypes, i.e. this test was not given to everybody). Participants were presented with 13 verbal and numerical reasoning questions and had to answer as many as they could, within two minutes. We used the total number of correct answers for our analyses. Data were normally distributed and did not require transformation.

Digit Span (numeric working memory, completed at the assessment centres at the first visit by 14,495 individuals with available genotypes). Participants were presented with progressively longer numbers (maximum 12) and asked to enter them back once the number had disappeared. We used the maximum number of digits remembered for our analyses. Data were normally distributed and did not require transformation.

Symbol Digit Substitution Test (complex processing speed, completed at follow-up on home computers by 33,057 individuals with available genotypes). This test involves matching numbers to a set of symbols. We used the number of correct substitutions for our analyses. There were small numbers of outliers, that suggested technical issues, so we excluded results of <3 and >36 substitutions. The remaining scores were normally distributed and did not require transformation.

Trail Making Test A and B (visual attention, completed at follow-up on home computers by 29,251 individuals with available genotypes). Participants were asked to connect scattered circles according to numbers (Trail A) and to alternating numbers and letters (Trail B). We used the time taken to complete these tests for our analyses and these data were log transformed prior to the generation of z scores.

For comparison, we also analysed data obtained from 507 individuals from the Biobank, who self-reported a diagnosis of schizophrenia at the interviews at the assessment centres, because schizophrenia is associated with impaired cognitive performance (10). Only 169 of the 507 have been genotyped so far and just three of them carried a CNV from our list of 53 loci.

We calculated the effect size reductions on cognitive tests of the carriers of the two CNV groups, and of individuals with schizophrenia, in linear regression analyses, corrected for age and sex (Table 1). We also compared the carriers of “schizophrenia CNVs” versus those with “other neurodevelopmental CNVs” using linear regression analyses, corrected for age and sex (Table 2).

Educational attainment and occupational level are highly correlated with cognitive performance (20, 21). We analysed these variables using ordinal regression with CNV carrier status and sex as factors and age as a covariate.

Results

151,659 samples passed QC (99.3%). 790,761 CNVs were retained after QC (without filtering for frequency or size), an average of 5.2 per person (range 0 to 29, Figure S1). The frequencies of individual CNVs were consistent between batches, indicating a lack of significant batch effects (Table S12). Overall, the CNVs from the UK Biobank sample occurred at rates strikingly similar to those among the control datasets genotyped on different arrays and called by us with the same methods (Table S2). Only two CNVs reached nominally significant difference between the two datasets, but neither survives correction for multiple testing of 93 loci. 3.8% of people in the Biobank carry a CNV from the list of 93 that were annotated. (The list of CNVs will be made available for download from the UK Biobank.) Of those CNVs, 54 have been shown to be significantly associated with neurodevelopmental disorders (4), including all 12 schizophrenia CNVs. In the Biobank, 1.12% of participants carried one of these neurodevelopmental CNVs (after excluding the common 15q11.2 duplication, found in 0.5% of subjects).

Cognitive Test Results: Carriers of both the “schizophrenia CNVs” and the “other neurodevelopmental CNVs” had impaired performance on the seven cognitive tests, compared to CNV non-carriers, with 9 of the 14 comparisons reaching statistically significant differences that survive a conservative Bonferroni correction for multiple testing for 14 tests. Table 1 presents these differences, expressed as un-standardised B coefficients (z scores), corrected for age and sex in linear regression analyses. Most differences were modest in magnitude (z scores between 0.01 and 0.51 below the non-CNV carriers). Individuals with schizophrenia performed worse than either group of CNV carriers (z scores between 0.35 and 0.90 below the non-CNV carriers) and all differences were highly significant.

<Table 1>

We then examined the differences in cognitive performance between carriers of the 12 schizophrenia CNVs (N = 1,087) and carriers of the remaining 41 neurodevelopmental CNVs (N = 484), generated again from linear regression analysis, corrected for age and sex. Their performance tended to be similar (Figure 1 and Table 2). Although two of the tests reached nominal levels of statistical significance, these do not survive correction for multiple testing for seven tests and were in opposite directions (Table 2).

<Figure 1 and Table 2>

Educational and Occupational Attainment: We compared the educational and occupational attainment of neurodevelopmental CNV carriers against the non-CNV carriers. Both groups of CNVs carriers attained lower educational qualifications, e.g. a smaller proportion obtained a university/college degree, or achieved A/AS-levels at school (post-compulsory education qualifications taken at 16-18 years of age, Figure 2). We carried out ordinal regression analysis, with qualifications as the dependent variable, CNV status and sex as factors and age as a covariate. This indicated lower odds (0.61) for carriers of schizophrenia CNVs to finish in a higher qualifications group (95% CI 0.55-0.68, Wald 76.3, $p = 2.4\times10^{-18}$). Similar results were found for carriers of the other neurodevelopmental CNVs: lower odds (0.54), (95% CI 0.46-0.64, Wald 52.5, $p = 4.4\times10^{-13}$). CNV carriers also tended to have occupations that require less training or academic skills (Figure 3). Ordinal regression analysis, with major job group as the dependent variable, CNV status and sex as factors and age as a covariate indicated lower odds (0.64) for carriers of schizophrenia CNVs to have a job in an occupational group that requires higher skills and longer training, as defined by Office of National Statistics (22) (95% CI 0.56-0.73, Wald 43.7, $p = 3.7\times10^{-11}$). Similar results were found for the carriers of other neurodevelopmental CNVs: lower odds (0.58), (95% CI 0.47-0.71, Wald 28.4, $p = 1.0\times10^{-7}$).

<Figures 2 and 3>

Discussion

CNVs are a rare but important cause of serious neurodevelopmental disorders such as ID, ASD, schizophrenia and a variety of congenital malformations (4, 5, 23). The UK Biobank sample, with half a million participants, is a unique resource for establishing the effects of CNVs on phenotypic outcomes. The data quality was very high as indexed by the low fraction of samples that failed our QC (0.7%). Identical QC steps allowed us to call reliably the selected set of pathogenic CNVs, but researchers wishing to analyse smaller or common CNVs might have to use different filtering criteria.

Frequencies of pathogenic CNVs: We established the frequencies of a set of 93 CNVs that have been proposed to be pathogenic (4, 5) in ~150,000 participants in the UK Biobank genotyped so far (Table S2). We compared these with a large control dataset comprising 26,628 people, where we had access to raw CNV data or had ourselves called the CNVs from raw microarray files (Illumina or other versions of Affymetrix arrays). In the absence of an opportunity to perform technical replication on different arrays, we reasoned that finding similar CNV frequencies was the best validation of our CNV calling. These were indeed remarkably similar, with just two reaching nominally significant differences that would not survive correction for the multiple testing involved (Table S2).

Carriers of CNVs implicated in neurodevelopmental phenotypes have reduced cognitive performance: In order to assess the cognitive performance of CNV carriers, we first selected, from the 93 annotated CNVs, a list of 54 CNVs that have been statistically associated with neurodevelopmental phenotypes, such as ID and ASD (4). This was done in order to exclude CNVs that are not confirmed to be associated with cognitive impairment (although they could be pathogenic for other medical conditions). We excluded from this list the common 15q11.2 duplication, found in 0.5% of the sample, as it would disproportionately affect the results.

Each CNV is likely to have its own set of phenotypic characteristics and cognitive signature. It is premature to analyse each one separately, as the study will be better powered for such analysis after all participants are genotyped. In order to provide an initial sub-group analysis, we divided the CNVs into a set of 12 that have been confirmed as associated with schizophrenia (7, 8), and the 41 "other neurodevelopmental" CNVs. This division is unlikely to represent an actual dichotomy, as all 12 schizophrenia loci are also neurodevelopmental ones and we recently proposed (8) that many of the "other neurodevelopmental" CNVs increase risk for schizophrenia but this has escaped statistical confirmation, due to their rarity. Our finding of similar cognitive deficit among carriers of the two groups provides another argument that this distinction is somewhat arbitrary.

A study on the Icelandic population found reduced cognitive performance in 144 healthy carriers of 11 pathogenic CNVs: 1q21.2dup, *NRXN1* del, 13q31.3dup, 15q11.2del, 16p12.1del, 16p11.2del+dup, 16p13.11dup, 17p12del+dup and 22q11.21dup (10). Most of these (except the 13q31.3 duplication and 17p12del and dup) are on our list of 53 neurodevelopmental loci. We analysed the participants' performance on seven cognitive tests. In order to allow comparison with previous studies (10), we also show results for 507 individuals recruited in the Biobank, who self-reported to have schizophrenia, as they are expected to perform even worse on these tests. Carriers of neurodevelopmental CNVs (of both groups) performed intermediately between non-carriers and individuals with schizophrenia, with reductions between 0.01 and 0.51 standard deviations (z scores), compared with CNV non-carriers. The results reached significance on most tests, even if corrected conservatively for multiple testing for 14 tests (Figure 1 and Table 1). The two CNV groups were similar to each other, and no p-value would survive correction for multiple testing (Table 2). Individuals with schizophrenia performed worse than any other group on all tests, at 0.35 to 0.90 standard deviations (z score) reductions from non-CNV carriers.

It is well established that cognitive performance predicts achievement at school and employment (20, 21, 24). Here we show that adult carriers of neurodevelopmental CNVs also had lower educational attainment and tended to have occupations requiring less time in training, with all comparisons being highly significant. We wanted to address the question whether the effect on school/occupational attainment is entirely explained by a reduction in cognitive performance among CNV carriers. We tested in a logistic regression analysis the effect of CNV carrier status on educational and occupational attainment, with and without the fluid intelligence score as a covariate (this test showed the highest effect size). Most of the effect of the CNV status on education/occupation was explained via the effect of the fluid intelligence score (Table S7), indicating that, whilst cognitive impairment has a major effect on educational and occupational attainment, other phenotypic consequences of having a pathogenic CNV also play a role (Supplementary Material). We note that 30.9% of carriers of neurodevelopmental CNV hold managerial or professional occupations and that the distribution of their cognitive tests performance overlaps with that of non-CNV carriers, with only a modest shift (Figure S7c as an example for the Fluid Intelligence test scores). This suggests that significantly impaired performance in the presence of a pathogenic CNV is not inevitable, at least for many CNV loci. It is possible that these highly functioning individuals may have performed even better, had they not carried a CNV. It has been suggested (25) that pathogenic CNVs produce a consistent degree of cognitive impairment, in the context of the individual's genetic background (e.g. by 2 SD in the case of 22q11.2 deletions). Our results for the Fluid Intelligence Test are consistent with this observation (Figure S7c), albeit with a more modest difference.

The majority of previous studies on pathogenic CNVs have recruited individuals from health services for ID, ASD, congenital anomalies or schizophrenia. Much less is known about the effects of these CNVs in adults from the general population (9, 10). Carriers of neurodevelopmental CNVs had reduced cognitive performance, educational and occupational attainment, with highly significant differences compared to non-CNV carriers. The effect size however is modest, <0.5 standard deviations for all tests, with large overlap between the groups (Figure S7c). We suggest that this may partly be explained by many severely affected CNV carriers, such as adults with intellectual disability, having not taken part in UK Biobank. This would result in highly functioning individuals being overrepresented. This may be in part due to the recruitment strategy and in part because some people with complex disabilities may have died before the recruitment age of 40-69 years. For example, there were only 5 individuals with 22q11.2 deletions, while we would expect about 37 carriers in a population of this size (the rate of this deletion among newborns is ~1:4,000), (9). This is a severe disorder, with a decreased IQ of ~30 points, multiple congenital malformations and reduced life expectancy. Similarly, there were no cases consistent with Prader-Willi/Angelman syndromes or Down's syndrome and there was only a single case with deletion at the Smith-Magenis syndrome region. Individuals with schizophrenia are also underrepresented, at only 0.12%, despite the disorder's lifetime risk of 0.4-0.5% (26). An even smaller proportion of such individuals took part in the follow-up cognitive tests (e.g. only 7% of individuals with schizophrenia completed the Trail Making Tests at follow-up, compared to 21% of the remaining participants). Therefore, it is reasonable to assume that, although the UK Biobank attempted to recruit a sample that reflects the general population, it probably underrepresents seriously affected individuals. By analogy, carriers of pathogenic CNVs who have taken part in the UK Biobank might be among the higher functioning CNV carriers. The UK Biobank is recognised to be a generalisable sample, rather than one that is representative of the general population (27). However, it has been suggested that for data sources of this magnitude, data generated can still be applied to the population as a whole (28).

In order to further assess the potential of the Biobank to discover phenotypes caused by pathogenic CNVs, we checked whether we can detect the clearly defined phenotypes of certain common and incompletely penetrant CNVs. We examined CNVs at 16p11.2 (44 deletions and 42 duplications) and 17p12 (84 deletions and 45 duplications). Deletions at 16p11.2 have been associated with obesity, while duplications at the same locus have been associated with reduced weight (29). Carriers of deletions and duplications at this locus are expected to have reduced cognitive performance (30). In contrast, deletions and duplications at 17p12 cause peripheral neuropathies: hereditary neuropathy with liability to pressure palsies (HNPP) for deletions (31), or Charcot-Marie-Tooth disease type 1A (CMT1A), for duplications (32), but do not have an associated cognitive phenotype. These phenotypes were detected with very high levels of statistical significance (Supplementary material, Section 8). Thus, there were highly increased rates of peripheral neuropathy but normal cognitive performance in 17p12 CNV carriers, while 16p11.2 carriers had reduced cognitive performance and increased or reduced BMI for deletions and duplications respectively.

The full Biobank dataset will allow detailed analysis on the health consequences of many more individual CNV loci and we will provide our list of CNVs to the Biobank, to assist researchers.

Limitations: There are several limitations to this study. Whilst the UK Biobank made attempts to make their sample as representative as possible, the low proportion of individuals with severe disorders such as schizophrenia means that the sample cannot be considered perfectly representative of the general population (discussed above). There was also a considerable variation in the number of people who were approached to perform each cognitive test, and it is possible that some tests are more likely to have been performed by higher-functioning individuals. This results in large variability in power between the tests and limits inferences that may be made for the tests with a lower completion rate.

Acknowledgements:

This work used the UK Biobank health resource. The work at Cardiff University was funded by the Medical Research Council (MRC) Centre Grant (MR/L010305/1) and Program Grant (G0800509). We thank Geoff Scopes (Affymetrix) and Kai Wang (author of PennCNV) for their advice for the analysis of Axiom arrays and Louise Wain and Martin Tobin for allowing us access to the library files on BiLEVE arrays.

Financial Disclosures: The authors have no financial disclosures to declare.

Legends to Figures

Figure 1. z score differences in cognitive performance. The figure shows the z scores differences for seven cognitive tests in the different groups of individuals, after correction for age and gender in a linear regression analysis. The bars represent the z score means and s.e. of the means. In blue are the scores among individuals with schizophrenia (including those that have not been genotyped). In green are the scores of carriers of schizophrenia-associated CNVs, in red those of "other neurodevelopmental" CNVs. A minus sign on the x-axis indicates a worse score for all tests (e.g. a lower score or a longer time to complete a test).

Figure 2. **Distribution of the two groups of CNV carriers and non-CNV carriers in each educational qualification group**. The British qualifications are grouped as follows: College/university degree; A/AS levels or equivalent: qualifications taken at 16-18 years of age, post-compulsory education; O levels/GCSEs or equivalent: qualifications taken at 14-16 years of age, at the end of compulsory education; CSEs or equivalent: a predecessor to GCSEs including vocational subjects; NVQ or HND or HNC or equivalent: vocational qualifications. Black bars: schizophrenia-related CNV carriers; grey bars: "other neurodevelopmental" CNV carriers; white bars: CNV non-carriers.

Figure 3. Distribution of the two groups of CNV carriers and non-CNV carriers in each major job group as defined by the Office of National Statistics (22). X-axis coding: 1 – Managers and Senior Officials; 2 – Professional Occupations; 3 – Associate Professionals and Technical Occupations; 4 – Administrative and Secretarial Occupations; 5 – Skilled Trades Occupations; 6 – Personal Service Occupations; 7 – Sale and Customer Service Occupations; 8 – Process, Plant and Machine Operatives; 9 – Elementary Occupations. Black bars: schizophrenia-related CNV carriers; grey bars: "other neurodevelopmental" CNV carriers; white bars: CNV non-carriers.

References

1. Lee C, Scherer SW (2010): The clinical context of copy number variation in the human genome. *Expert Rev Mol Med* 12:e8.
2. Feuk L, Carson AR, Scherer SW (2006): Structural variation in the human genome. *Nat Rev Genet* 7:85-97.
3. Shaffer LG, Lupski JR (2000): Molecular mechanisms for constitutional chromosomal rearrangements in humans. *Annu Rev Genet* 34:297-329.
4. Coe BP, Witherspoon K, Rosenfeld JA, van Bon BW, Vulto-van Silfhout AT, Bosco P, et al. (2014): Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet* 46:1063-1071.
5. Dittwald P, Gambin T, Szafranski P, Li J, Amato S, Divon MY, et al. (2013): NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits. *Genome Res* 23:1395-1409.
6. Lal D, Ruppert AK, Trucks H, Schulz H, de Kovel CG, Kasteleijn-Nolst Trenité D, et al. (2015): Burden analysis of rare microdeletions suggests a strong impact of neurodevelopmental genes in genetic generalised epilepsies. *PLoS Genet* 11:e1005226.
7. Rees E, Walters JT, Georgieva L, Isles AR, Chambert KD, Richards AL, et al. (2014): Analysis of copy number variations at 15 schizophrenia-associated loci. *Br J Psychiatry* 204:108-114.
8. Rees E, Kendall K, Pardinas A, Legge S, Pocklington A, Escott-Price A, et al. (2016): Analysis of intellectual disability copy number variants for association with schizophrenia. *JAMA Psychiatry.* Published online: 17 August 2016.
9. Kirov G, Rees E, Walters JT, Escott-Price V, Georgieva L, Richards AL, et al. (2014): The penetrance of copy number variations for schizophrenia and developmental delay. *Biol Psychiatry* 75:378-385.
10. Stefansson H, Meyer-Lindenberg A, Steinberg S, Magnusdottir B, Morgen K, Arnarsdottir S, et al. (2014): CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* 505:361-366.
11. Männik K, Mägi R, Macé A, Cole B, Guyatt AL, Shihab HA, et al. (2015): Copy number variations and cognitive phenotypes in unselected populations. *JAMA* 313:2044-2054.
12. Wain LV, Shrine N, Miller S, Jackson VE, Ntalla I, Soler Artigas M, et al. (2015): Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir Med* 3:769-781.
13. Rees E, Walters JT, Chambert KD, O'Dushlaine C, Szatkiewicz J, Richards AL, et al. (2014): CNV analysis in a large schizophrenia sample implicates deletions at 16p12.1 and SLC1A1 and duplications at 1p36.33 and CGNL1. *Hum Mol Genet* 23:1669-1676.
14. Kirov G, Pocklington AJ, Holmans P, Ivanov D, Ikeda M, Ruderfer D, et al. (2012): De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol Psychiatry* 17:142-153.
15. Ikeda M, Aleksic B, Kirov G, Kinoshita Y, Yamanouchi Y, Kitajima T, et al. (2010): Copy number variation in schizophrenia in the Japanese population. *Biol Psychiatry* 67:283-286.

16. Grozeva D, Kirov G, Conrad DF, Barnes CP, Hurles M, Owen MJ, et al. (2013): Reduced burden of very large and rare CNVs in bipolar affective disorder. *Bipolar Disord* 15:893-898.
17. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. (2007): PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 17:1665-1674.
18. Levinson DF, Duan J, Oh S, Wang K, Sanders AR, Shi J, et al. (2011): Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. *Am J Psychiatry* 168:302-316.
19. International Schizophrenia Consortium (2008): Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 455:237-241.
20. Wagner R (1997): Intelligence, training and employment. *Am Psychologist* 52:1059-1069.
21. Schmidt F (2002): The role of general cognitive ability and job performance: why there cannot be a debate. *Hum Performance* 15:187-210.
22. Office for National Statistics (2000): *Standard occupational classification 2000*. London: The Stationery Office.
23. Kirov G (2015): CNVs in neuropsychiatric disorders. *Hum Mol Genet* 24:R45-49.
24. Strenze T (2007): Intelligence and socioeconomic success: a meta-analytic review of longitudinal research. *Intelligence* 35:401-426.
25. Moreno-De-Luca A, Myers SM, Challman TD, Moreno-De-Luca D, Evans DW, Ledbetter DH (2013): Developmental brain dysfunction: revival and expansion of old concepts based on new genetic evidence. *Lancet Neurol* 12:406-414.
26. McGrath J, Saha S, Chant D, Welham J (2008): Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiol Rev* 30:67-76.
27. Collins R (2012): What makes UK Biobank special? *Lancet* 379:1173-1174.
28. Manolio TA, Collins R (2010): Enhancing the feasibility of large cohort studies. *JAMA* 304:2290-2291.
29. McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S, et al. (2009): Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet* 41:1223-1227.
30. Jacquemont S, Reymond A, Zufferey F, Harewood L, Walters RG, Kutalik Z, et al. (2011): Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature* 478:97-102.
31. Chance PF, Abbas N, Lensch MW, Pentao L, Roa BB, Patel PI, et al. (1994): Two autosomal dominant neuropathies result from reciprocal DNA duplication/deletion of a region on chromosome 17. *Hum Mol Genet* 3:223-228.
32. Lupski JR, Wise CA, Kuwano A, Pentao L, Parke JT, Glaze DG, et al. (1992): Gene dosage is a mechanism for Charcot-Marie-Tooth disease type 1A. *Nat Genet* 1:29-33.

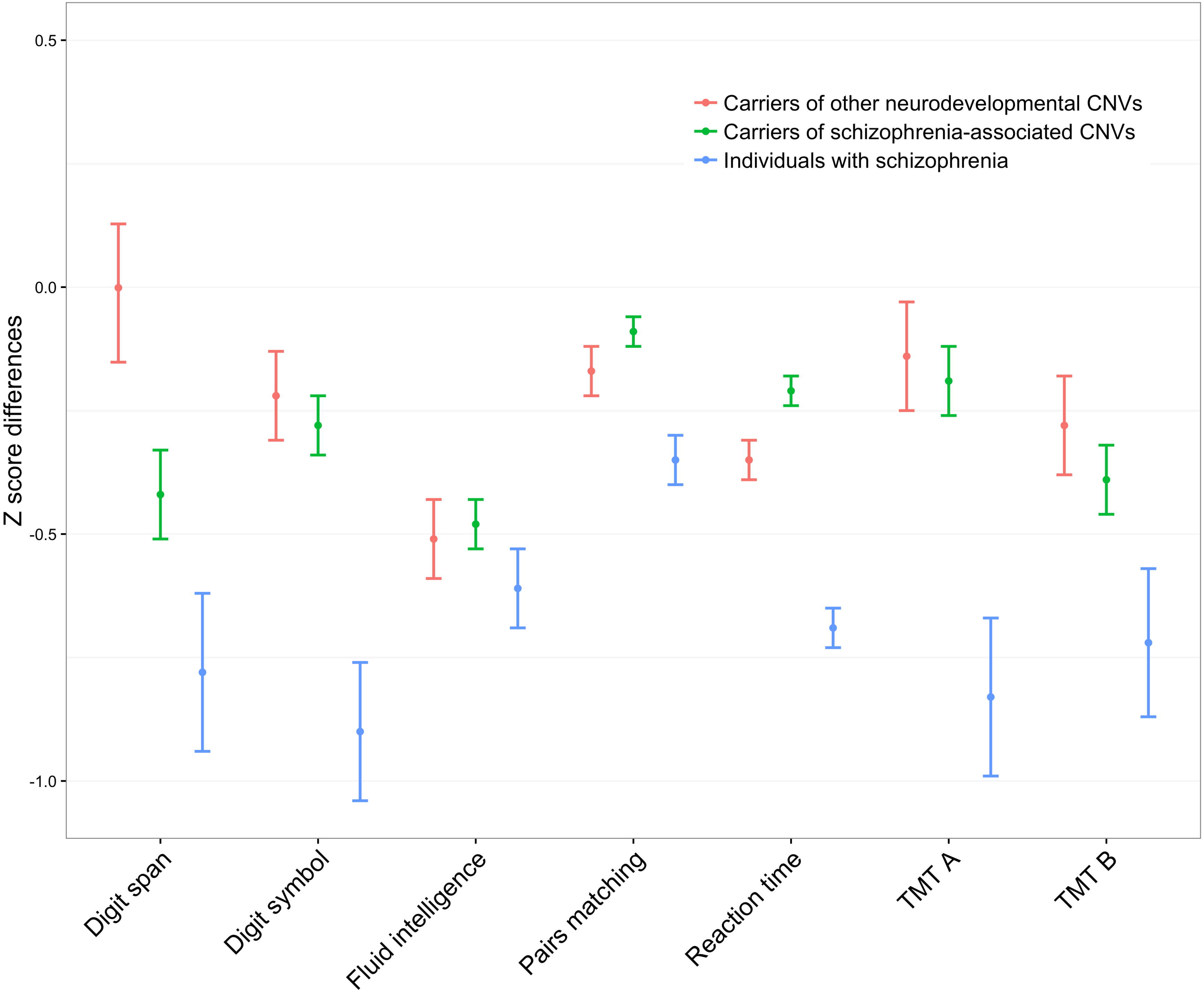
Cognitive Test	Non-CNV Carriers	Carriers of Schizophrenia CNVs			Carriers of Other Neurodevelopmental CNVs			Individuals with Schizophrenia		
	N	N	B (s.e.)	*p*	N	B (s.e.)	*p*	N	B (s.e.)	*p*
Pairs Matching – Number of Incorrect Matches	134781	1048	0.09 (0.03)	0.003	463	0.17 (0.05)	$3.2x10^{-4}$	436	0.35 (0.05)	$7.82x10^{-14}$
Reaction Time – Mean Time to Correctly Identify Matches	137053	1073	0.21 (0.03)	$1.52x10^{-13}$	477	0.35 (0.04)	$9.46x10^{-16}$	480	0.69 (0.04)	$1.05x10^{-57}$
Fluid Intelligence – Score	44107	321	-0.48 (0.05)	$5.29x10^{-19}$	147	-0.51 (0.08)	$3.97x10^{-10}$	168	-0.61 (0.08)	$2.51x10^{-15}$
Digit Span – Number of Digits Remembered	14343	102	-0.42 (0.09)	$1.8x10^{-5}$	50	-0.01 (0.14)	0.931	39	-0.78 (0.16)	$8.7x10^{-7}$
Symbol Digit Substitution – Number of Correct	32770	204	-0.29 (0.06)	$3.0x10^{-6}$	83	-0.22 (0.09)	0.024	42	-0.90 (0.14)	$8.33x10^{-11}$

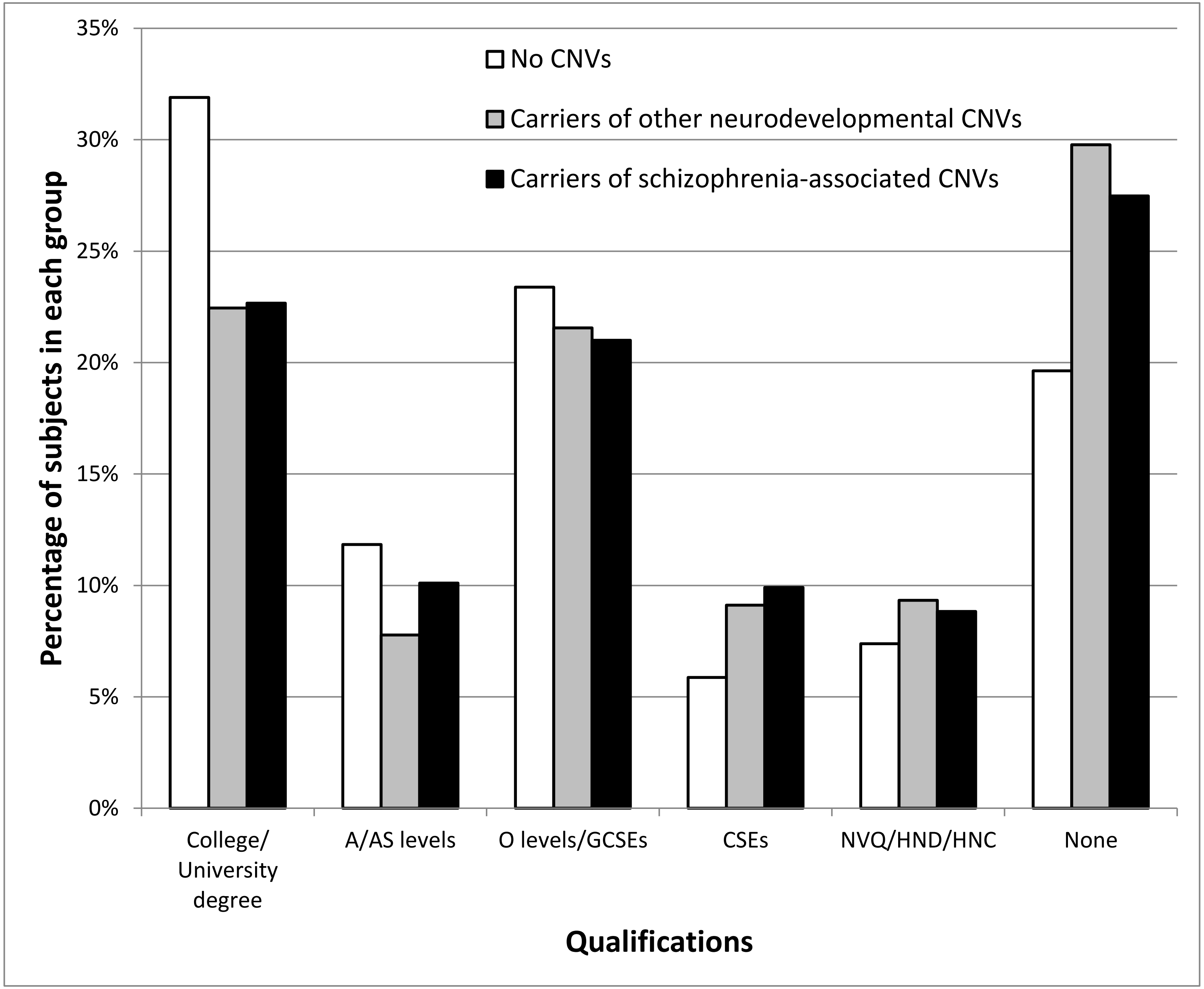
Matches										
Trail Making Test A – Time to Complete	28988	185	0.19 (0.07)	0.008	78	0.14 (0.11)	0.199	36	0.83 (0.16)	1.51×10^{-7}
Trail Making Test B – Time to Complete	28988	185	0.38 (0.07)	2.07×10^{-8}	78	0.28 (0.10)	0.008	36	0.72 (0.15)	3.00×10^{-6}

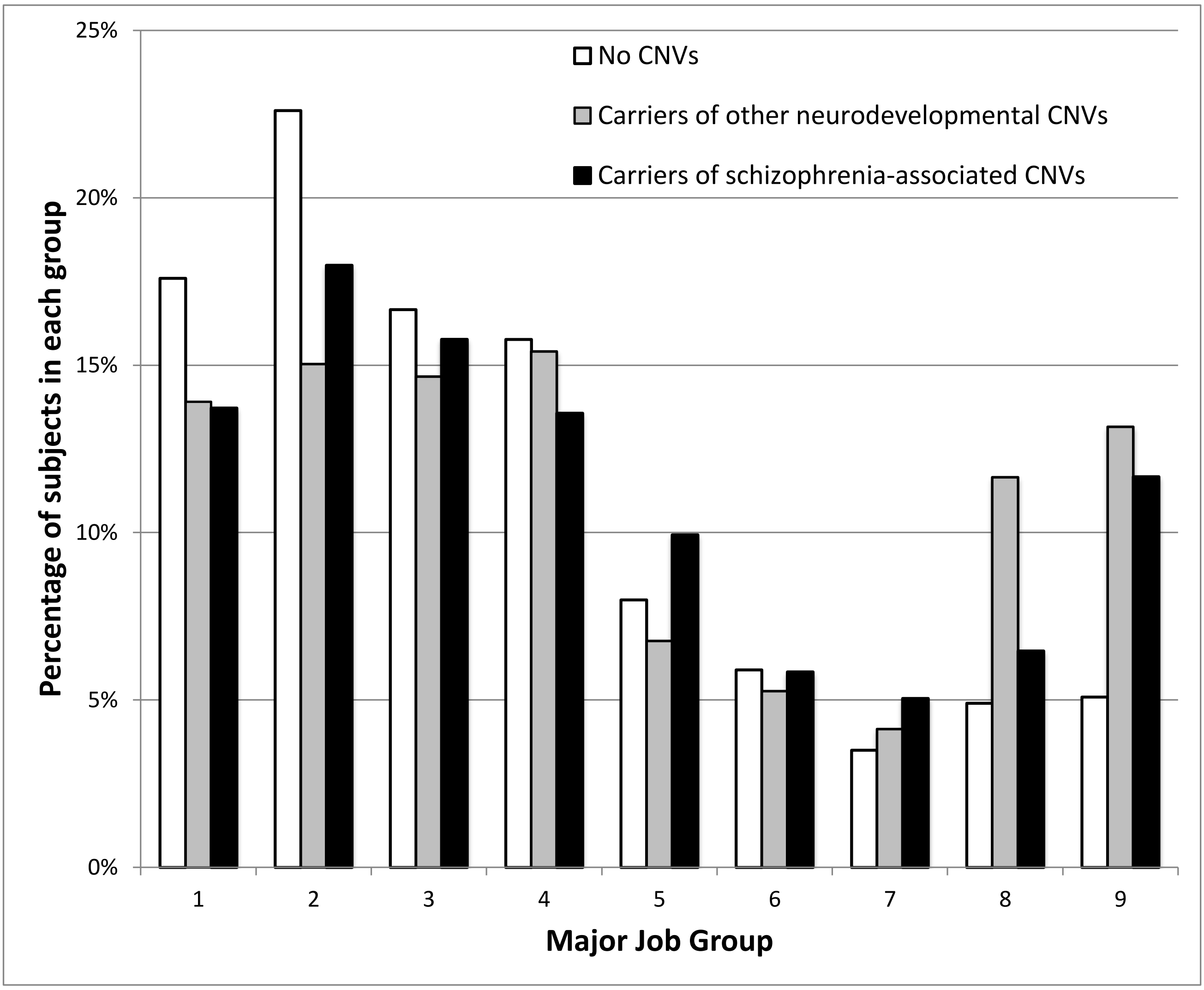
Table 1. Results on cognitive tests in the two groups of CNV carriers and individuals with schizophrenia, compared to non-CNV carriers. B indicates the differences from non-CNV-carriers, expressed as z scores (s.e. mean) in linear regression analysis, corrected for age and sex. Non-CNV carriers, by definition, have z scores of 0.

Cognitive Test	N Carriers of Schizophrenia CNVs	N Carriers of Other Neurodevelopmental CNVs	B (s.e.)	*p*
Pairs Matching	1048	463	-0.02 (0.05)	0.74
Reaction Time	1073	477	-0.13 (0.06)	0.026
Fluid Intelligence – Score	321	147	0.046 (0.09)	0.63
Digit Span	102	50	-0.39 (0.18)	0.031
Symbol Digit Substitution	204	83	-0.08 (0.12)	0.52
Trail Making Test A	185	78	0.06 (0.13)	0.68
Trail Making Test B	185	78	0.11 (0.13)	0.41

Table 2. Comparison of cognitive tests in carriers of 12 schizophrenia-associated CNVs and individuals with other neurodevelopmental CNVs. B indicates the differences between the two groups, expressed as z score (s.e. mean) as in Table 1. P-values are based on linear regression analysis, as above.


0.5
0.0
-0.5
-1.0
Z score differences
Carriers of other neurodevelopmental CNVs
Carriers of schizophrenia-associated CNVs
Individuals with schizophrenia
Digit span
Digit symbol
Fluid intelligence
Pairs matching
Reaction time
TMT A
TMT B


□ No CNVs
▨ Carriers of other neurodevelopmental CNVs
■ Carriers of schizophrenia-associated CNVs
35%
30%
25%
20%
15%
10%
5%
0%
Percentage of subjects in each group
College/ University degree
A/AS levels
O levels/GCSEs
CSEs
NVQ/HND/HNC
None
Qualifications


25%
20%
15%
10%
5%
0%
Percentage of subjects in each group
□ No CNVs
▨ Carriers of other neurodevelopmental CNVs
■ Carriers of schizophrenia-associated CNVs
1
2
3
4
5
6
7
8
9
Major Job Group

Cognitive Performance Among Carriers of Pathogenic Copy Number Variants: Analysis of 152,000 UK Biobank Subjects

Supplement 1

Contents

Sample Processing and Genotyping

The UK Biobank obtained two 10ml EDTA vacutainers of blood per participant. DNA was extracted and purified using a modified Maxwell 16 Blood DNA Purification Kit (Promega – AS1010X) (1). Details of the genotyping processes used are available on the UK Biobank website. Briefly, samples were genotyped at the Affymetrix Research Laboratory, Santa Clara, CA on 96-well plates. Samples were genotyped on two very similar arrays with ~95% probe overlap between them:

1. The UK Biobank Axiom Array includes over 820,000 variants, providing comprehensive coverage of the genome and including rare coding variants, pharmacogenomics markers, and extra coverage for known copy number regions, the HLA region, genes involved in inflammation and eQTL variants.
2. The UK BiLEVE array was designed for a University of Leicester project investigating genetic variation in Chronic Obstructive Pulmonary Disease (COPD). Approximately 50,000 participants in the UK Biobank were genotyped on this array prior to the introduction of the UK Biobank Axiom Array. The UK BiLEVE Array contains 807,411 probes.

CNV Calling at Cardiff University

Anonymized genotype data were downloaded as raw CEL files in batches of 1,000 files from the UK Biobank using an authentication key. They were stored and processed on a secure UNIX server.

For the initial step, we used the Affymetrix Power Tools software downloaded from the Affymetrix website (http://media.affymetrix.com/partners_programs/programs/developer/tools/powertools.affx). We used the *apt-probeset-genotype* command on batches of ~4,600 files using library files for the two arrays downloaded from the Affymetrix website and the following parameters:

```
/apt-probeset-genotype --analysis-files-path /Axiom_UKB_WCSG.r3/ --xml-file /Axiom_UKB_WCSG_96orMore_Step2_Bi-allelic.r3.apt-probeset-genotype.AxiomGT1.xml --out-dir /Batch1 --summaries --cel-files cel.list_batch1.txt
```

This command (example given arbitrarily for Batch1) creates four files that are required for subsequent steps:

AxiomGT1.calls.txt – genotype calls

AxiomGT1.confidences.txt – confidences for the genotype calls

AxiomGT1.report.txt – summaries for the samples analyzed including the computed gender, call rate and heterozygosity

AxiomGT1.summary.txt – normalized intensities of A and B alleles

We used only biallelic markers for our CNV calls as they are utilized by the PennCNV software for the creation of genotype clusters. Of the original ~835,000 markers, this left ~750,000 markers for analysis (Table S1 for details on each batch). Other researchers may use different .xml files and different steps, here we only present the methods we used.

For subsequent steps, we used the PennCNV-Affy program to complete CNV calling (http://penncnv.openbioinformatics.org/en/latest/user-guide/affy/). This free software has been developed and maintained by Dr Kai Wang and colleagues (2). First, we generated canonical genotype clustering files from the above files, e.g.:

```
/generate_affy_geno_cluster.pl AxiomGT1.calls.txt AxiomGT1.confidences.txt AxiomGT1.summary.txt --nopower2 -locfile mapfile.dat -sexfile sex_batch1.txt -out batch1.genocluster
```

As the signal intensity values had not been log2 normalized, the –nopower2 argument was used. We are grateful to Dr Wang for suggesting this solution.

Next, we calculated Log R Ratio (LRR) and B-Allele Frequency (BAF) values using the *normalize_affy_geno_cluster.pl* command, e.g.:

```
normalize_affy_geno_cluster.pl batch1.genocluster AxiomGT1.summary.txt -nopower2 -locfile mapfileAX.dat -out batch1_lrr_baf.txt
```

From here, our analyses followed the Affymetrix CNV Calling Overview described on the PennCNV website (http://penncnv.openbioinformatics.org/en/latest/user-guide/affy/).

Briefly, we split the signal files into individual files for CNV calling by PennCNV and then detected the raw CNV calls. We created pfb and gcmodel files from existing data and “trained” a .hmm file on 100 Axiom samples, using the available Affy 6.0 hmm file as a template.

Quality Control (QC) Filtering

We examined the distribution of the QC parameters, which we routinely used in our previous work to filter samples. Individuals were filtered out if they were outliers for the following parameters: if they had >30 CNVs (Figure S1), had a genotype call rate <96% (Figure S2), had a waviness factor (WF) >0.03 or <-0.03 (Figure S3). We decided not to filter on LRR SD (Figure S4). These are quite relaxed filtering criteria, compared to those we used in our previous studies (e.g. (3)). We wanted to review the results with these filtering steps first, and then filter more stringently if needed. As explained in the main text, the pathogenic CNVs were called confidently with these criteria, so we saw no reason to filter more stringently. Other researchers might want to set different filters, especially if analyzing small or common CNVs. 1,020 samples were excluded using these filters. This number varied between 0 and 146 samples for individual batches (see below, Table S1). We re-clustered Batch 14 after excluding 114 failed samples to see if this resulted in better quality CNVs in the remaining samples. The resulting set of CNVs in the re-clustered samples was practically identical, showing that the small number of bad samples did not introduce important biases in CNV calling. We elected not to repeat this process for other batches.

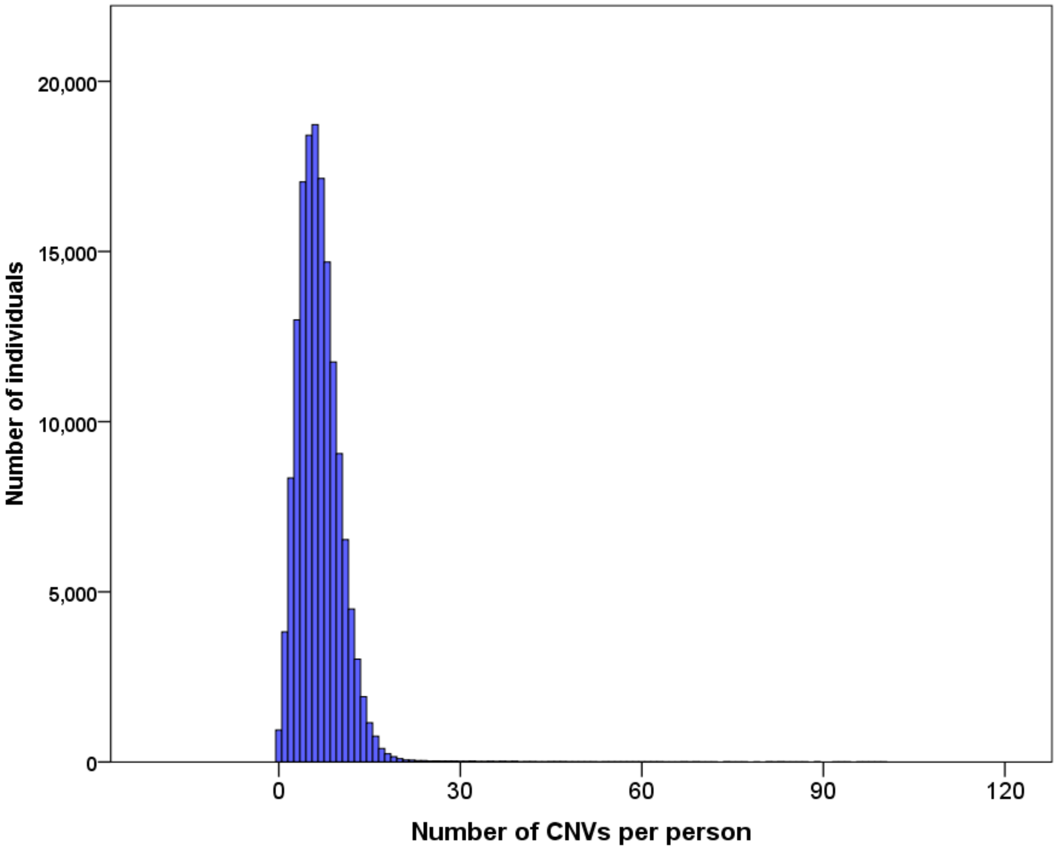


Figure S1. Distribution of the number of CNVs per person, before filtering (all CNVs included, starting from those called with 3 probes). We filtered out samples with 30 or more CNVs. Another 160 samples are not included in the graph, to improve the visualisation, as they had >120 CNVs each, ranging up to 1207 CNVs.

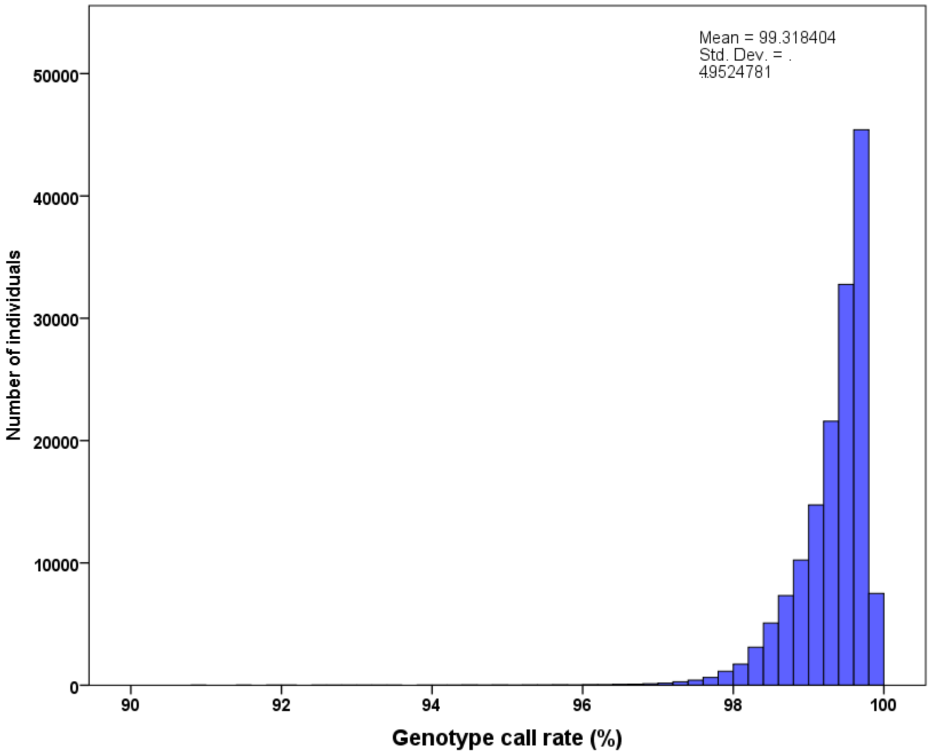


Figure S2. Distribution of genotype call rate scores before filtering. We excluded samples with a SNP call rate of <96%. The full distribution of samples is shown.

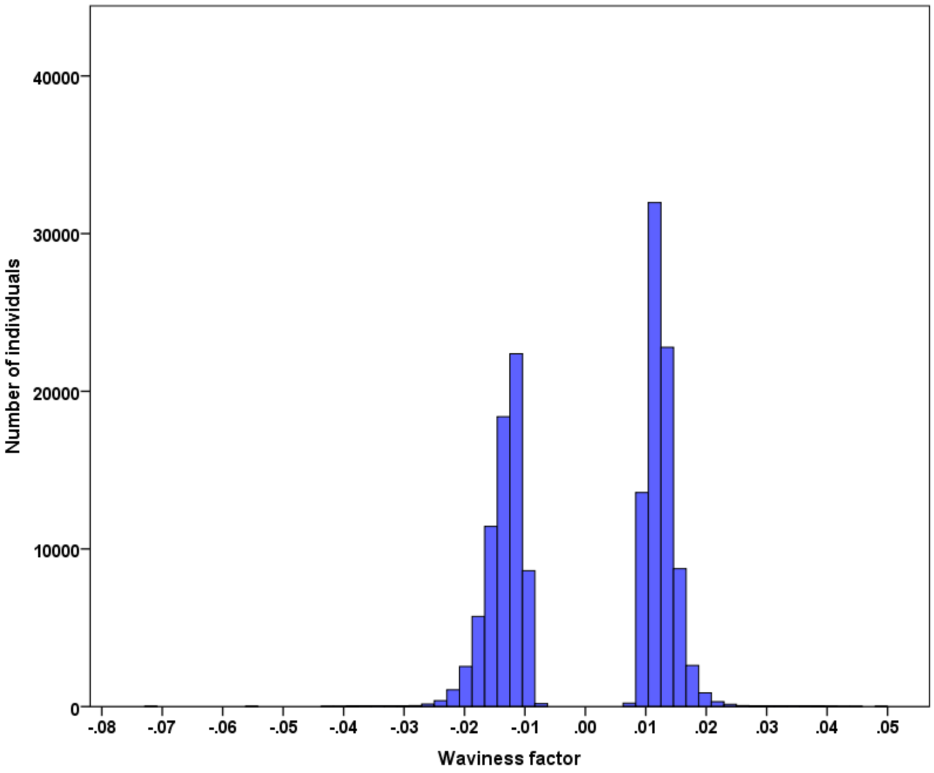


Figure S3. Distribution of the waviness factor (WF), before filtering. We excluded samples with WF >0.03 and <-0.03. The full distribution of samples is shown.

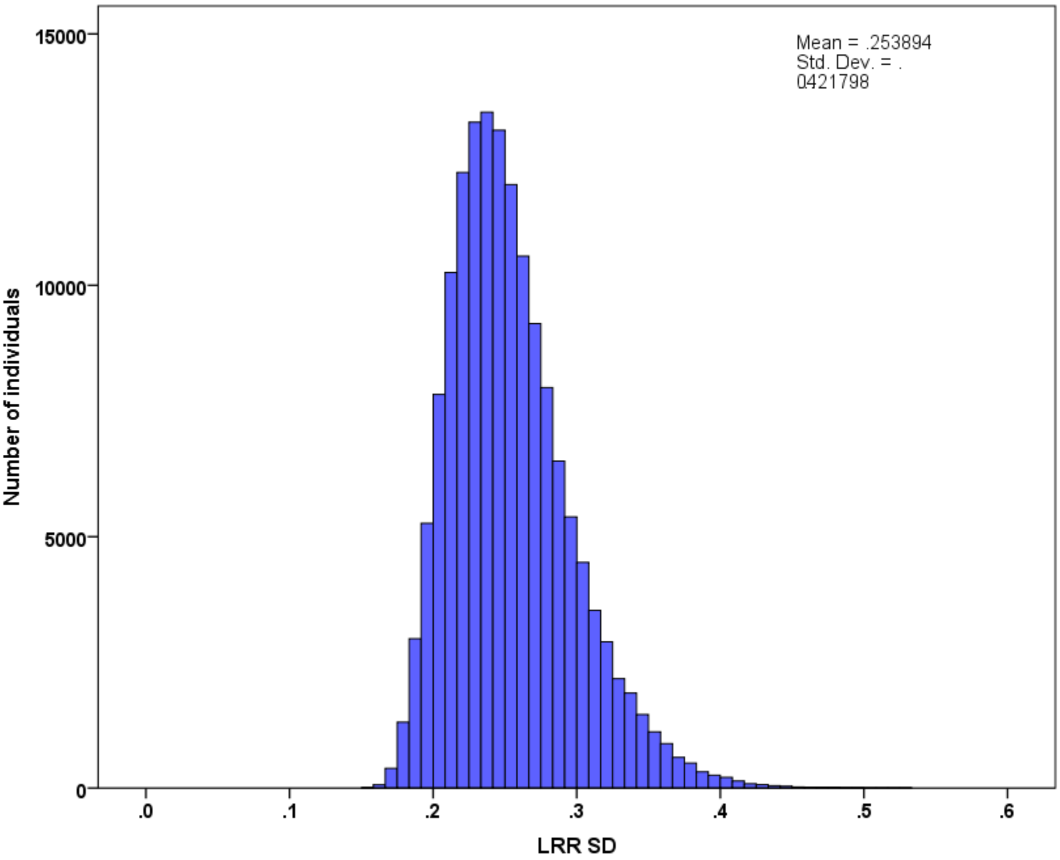


Figure S4. Distribution of the Log R Ratio standard deviation (LRR SD) scores. The full distribution of samples is shown. We decided not to filter on LRR SD, because poorly performing samples still picked up the pathogenic CNVs at similar rates.

Annotation of CNVs

For our annotations, we selected 93 CNVs proposed as possible intellectual disability CNVs in two of the largest and best-known studies (4, 5). These CNVs and their critical regions are shown in Table S2. The original papers divided some of the regions into sub-regions. To simplify the presentation and interpretation we grouped such regions together (e.g. “large” and “small” 22q11.2 CNVs are presented as the same). We used a script to annotate these CNVs in each batch and then manually inspected the positions of the annotated CNVs, filtering out those that did not cover the correct interval. Whilst it is relatively straightforward to call recurrent CNVs when their breakpoints are close to the low copy repeats (LCR) flanking such CNVs, in some cases this is not so obvious. For example, we had to decide how to call CNVs that do not cover the full critical region. We compiled a set of rules to facilitate this process (Table S4). As a general rule, a CNV was called if it covered >50% of the region, including any key genes in that region (if known). In the case of single gene CNVs, deletions were called if they intersected an exon. For duplications of single genes, we required the whole gene to be duplicated. If a CNV covered two known, adjacent loci, we called it according to the more pathogenic of these loci. If a CNV was >20Mb in size, we did not give it the name of a specific CNV locus, but instead called it “large” even if it included a known CNV. Some loci had specific rules (Table S4). Other researchers may elect to use other criteria in the annotation of CNVs.

Samples Processed by Stage and Batch

Table S1. Samples processed at each stage per batch.

Batch	CEL	Genotype	Cluster	LRR/BAF	Initial CNVs	QC CEL	Excluded samples	Final CNVs
1	4708	834897	757018	756750	30266	4705	3	23977
2	4656	834845	759759	759490	31374	4652	4	24476
3	4648	834837	757910	757636	29291	4648	0	23197
4	4651	834840	759630	759364	29809	4650	1	23732
5	4661	834850	758994	758727	30253	4658	3	24035
6	4688	834877	759928	759660	30978	4686	2	24434
7	4677	834866	760538	760259	34025	4672	5	26013
8	4755	834944	760228	759953	30351	4754	1	24055
9	4692	834881	761596	761319	31252	4689	3	24373
10	4713	834902	760493	760494	31809	4709	4	24616
11	4700	834889	759773	759506	28644	4699	1	22658
12	4705	834894	760790	760510	28374	4701	4	22506
13	4691	834880	759337	759071	28913	4691	0	22792
14	4708	834897	757404	757140	25011*	4594	114	20509
15	4713	834902	757852	757594	27731	4699	14	21094
16	4604	834793	759225	758962	39377	4491	113	21869
17	4598	834787	758505	758238	29262	4552	46	23132
18	4621	834810	757875	757617	36243	4525	96	23188
19	4626	834815	758296	758036	35166	4490	136	22433
20	4635	834824	758915	758653	30530	4531	104	22327
21	4581	834770	759645	759383	27507	4569	12	22048
22	4719	834908	759932	759663	31807	4709	10	24695
-1	4540	837704	744454	744161	35746	4498	42	25817
-2	4551	837715	745106	744816	68699	4465	86	24788
-3	4530	837695	743842	743559	30807	4527	3	23669
-4	4548	837712	744820	744536	33428	4548	0	25424
-5	4531	837695	744490	744199	32012	4528	3	24543
-6	4526	837690	744022	743728	34067	4437	89	24121
-7	4532	837696	744520	744244	35741	4515	17	26300
-8	4557	837721	745555	745270	34516	4554	3	25898
-9	4537	837701	746022	745729	41317	4391	146	25554
-10	4563	837727	744593	744303	33977	4560	3	25523
-11	4563	837727	744880	744584	35378	4562	1	26965

Batches 1 to 22 were genotyped on the Axiom array, batches -1 to -11 on the BiLEVE array.

CEL – number of CEL files downloaded; Genotype – number of genotypes generated; Cluster – number of genotypes following clustering; LRR/BAF – number of Log R Ratios and B Allele Frequencies generated; Initial CNVs – initial number of CNVs detected; QC CEL – number of samples after QC filtering; Excluded samples – number of individuals excluded; Final CNVs – number of CNVs following completion of all QC stages.

*Batch 14 was re-clustered after removing the 114 failed samples. We show the number of CNVs following re-clustering.

CNV Breakpoints and Frequencies in the UK Biobank and in the Other Control Samples

Table S2. CNV loci, their genomic coordinates (according to Feb. 2009 UCSC GRCh37/hg19 Assembly) and their frequencies in the UK Biobank and in the other control datasets. This list was compiled from studies establishing association of these CNVs with neurodevelopmental disorders including developmental delay and autism spectrum disorder (4, 5). The "other control" datasets are described elsewhere (6-9) and are summarized in the next section. We also show the results of a comparison of the frequencies between the two datasets, using Fisher Exact test. The CNVs included in our list of "schizophrenia CNVs" and "other neurodevelopmental CNVs" are indicated. The last column lists genes from the PSD complex that are within the CNV loci (following on our previous work, (10)).

See Supplement 2 (Excel file) for Table S2.

Other Control Datasets

We established the frequencies of these CNVs in other control datasets, using the same criteria as for the UK Biobank. We used a total of 26,628 controls from various datasets where we had access to the raw data (or had completed all the analysis ourselves), so were able to apply the same criteria. Here we provide a brief list of these datasets. Some of these are not strict 'controls' but have phenotypes that we considered unlikely to be related to our list of pathogenic CNVs, e.g. smoking, melanoma.

Table S3. Other control datasets used for comparison of frequencies.

Dataset	Reference	Array	N after QC
International Schizophrenia Consortium	(11)	Affy 5.0, Affy 6.0	3,181
Molecular Genetics of Schizophrenia	(7)	Affy 6.0	3,437
The Genetic Architecture of Smoking and Smoking Cessation	dbGaP (phs000404.v1.p1) (8)	Illumina HumanOmni2.5 520,766 overlapping probes used	1,488
High Density SNP Association Analysis of Melanoma: Case-Control and Outcomes Investigation	dbGaP (phs000187.v1.p1) (8)	Illumina HumanOmni1_Quad_v1-0-B 520,766 overlapping probes used	2,971
Genetic Epidemiology of Refractive Error in the KORA Study	dbGaP (phs000303.v1.p1) (8)	Illumina HumanOmni2.5 520,766 overlapping probes used	1,857
National Blood Donors (NBS) Cohort (WTCCC2)	EGA (EGAD00000000024) (8)	Illumina 1.2M 520,766 overlapping probes used	2,375
1958 British Birth Cohort (WTCCC2)	EGA (EGAD00000000022) (8)	Illumina 1.2M 520,766 overlapping probes used	2,564

Dataset	Reference	Array	N after QC
Chronic Obstructive Pulmonary Disease	dbGAP (phs000179.v3.p2) (9)	HumanOmni1-Quad_v1-0-Multi_H 666,868 overlapping probes used	992
Corneal dystrophy	dbGAP (phs000421.v1.p1) (9)	HumanOmni2.5-4v1_H 666,868 overlapping probes used	3,529
Mammography	dbGAP (phs000395.v1.p1) (9)	HumanOmni1_Quad_v1-0_B 666,868 overlapping probes used	954
Melanoma 2	dbGAP (phs000519.v1.p1) (9)	HumanOmniExpressExome-8v1_A 666,868 overlapping probes used	2,416
Cardiff Controls (blood donors)	Cardiff (9)	HumanOmniExpress-12v1-1_H 666,868 overlapping probes used	860

We would like to acknowledge the use of these freely available datasets:

International Schizophrenia Consortium (ISC)

Details of the ISC dataset have been previously published (6). The sample consists of six European populations genotyped at the Broad Institute, Cambridge, Massachusetts using Affymetrix 6.0 or 5.0 genotyping arrays. We analyzed CNVs in 3,185 controls.

Molecular Genetics of Schizophrenia (MGS)

Details of the MGS dataset have been described elsewhere (7). The sample consists of individuals of European American ancestry and African American ancestry genotyped at the Broad Institute, Cambridge, Massachusetts using Affymetrix 6.0 genotyping arrays. We analyzed CNVs in 2,556 controls of European American ancestry and 881 controls of African American ancestry (passing QC). CNVs were called using the Birdsuite algorithm (12). All individuals with schizophrenia met DSM-IV criteria for schizophrenia or schizoaffective disorder (13).

We obtained access to the MGS dataset through dbGaP. Funding support for the Genome-Wide Association of Schizophrenia Study was provided by the National Institute of Mental Health (R01 MH67257, R01 MH59588, R01 MH59571, R01 MH59565, R01 MH59587, R01 MH60870, R01 MH59566, R01 MH59586, R01 MH61675, R01 MH60879, R01 MH81800, U01 MH46276, U01 MH46289 U01 MH46318, U01 MH79469, and U01 MH79470) and the genotyping of samples was provided through the Genetic Association Information Network (GAIN). The datasets used for the analyses described in this manuscript were obtained from the database of Genotypes and Phenotypes (dbGaP) found at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000021.v3.p2. Samples and associated phenotype data for the Genome-Wide Association of Schizophrenia Study were provided by the Molecular Genetics of Schizophrenia Collaboration (PI: Pablo V. Gejman, Evanston Northwestern Healthcare (ENH) and Northwestern University, Evanston, IL, USA). Funding support for the companion studies, Genome-Wide Association Study of Schizophrenia (GAIN) and Molecular Genetics of Schizophrenia - nonGAIN Sample (MGS_nonGAIN), was provided by Genomics Research Branch at NIMH and the genotyping and analysis of samples was provided through the Genetic Association Information Network (GAIN) and under the MGS U01s: MH79469 and MH79470. Assistance with data cleaning was provided by the National Center for Biotechnology Information. The MGS dataset(s) used for the analyses described in this manuscript were obtained from the database of Genotype and Phenotype (dbGaP) found at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession numbers phs000021.v2.p1 (GAIN) and phs000167.v1.p1 (nonGAIN). Samples and associated phenotype data for the MGS GWAS study were collected under the following grants: NIMH Schizophrenia Genetics Initiative U01s: MH46276 (CR Cloninger), MH46289 (C Kaufmann), and MH46318 (MT Tsuang); and MGS Part 1 (MGS1) and Part 2 (MGS2) R01s: MH67257 (NG Buccola), MH59588 (BJ Mowry), MH59571 (PV Gejman), MH59565 (Robert Freedman), MH59587 (F Amin), MH60870 (WF Byerley), MH59566 (DW Black), MH59586 (JM Silverman), MH61675 (DF Levinson), and MH60879 (CR Cloninger). Further details of collection sites, individuals, and institutions may be found in data supplement Table 1 of Sanders et al (15) and at the study dbGaP pages.

The Genetic Architecture of Smoking and Smoking Cessation

These data were accessed through dbGAP: study accession phs000404.v1.p1. Funding support for genotyping, which was performed at the Center for Inherited Disease Research (CIDR), was provided by 1 X01 HG005274-01. CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268200782096C. Assistance with genotype cleaning, as well as with general study coordination, was provided by the Gene Environment Association Studies (GENEVA) Coordinating Center (U01 HG004446). Funding support for collection of datasets and samples was provided by the Collaborative Genetic Study of Nicotine Dependence (COGEND; P01 CA089392) and the University of Wisconsin Transdisciplinary Tobacco Use Research Center (P50 DA019706, P50 CA084724).

High Density SNP Association Analysis of Melanoma: Case-Control and Outcomes Investigation

These data were accessed through dbGaP: study accession phs000187.v1.p1: Research support to collect data and develop an application to support this project was provided by 3P50CA093459, 5P50CA097007, 5R01ES011740, and 5R01CA133996.

Genetic Epidemiology of Refractive Error in the KORA Study

These data were accessed through dbGaP: study accession phs000303.v1.p1. Principal Investigators: Dwight Stambolian, University of Pennsylvania, Philadelphia, PA, USA; H. Erich Wichmann, Institut für Humangenetik, Helmholtz-Zentrum München, Germany, National Eye Institute, National Institutes of Health, Bethesda, MD, USA. Funded by R01 EY020483, National Institutes of Health, Bethesda, MD, USA.

National Blood Donors (NBS) Cohort and 1958 British Birth Cohort (Wellcome Trust Case Control Consortium 2 - WTCCC2)

Samples were downloaded from https://www.ebi.ac.uk/ega/ and included samples from the National Blood Donors Cohort, EGAD00000000024 and samples from the 1958 British Birth Cohort, EGAD00000000022. Funding for these projects was provided by the Wellcome Trust Case Control Consortium 2 project (085475/B/08/Z and 085475/Z/08/Z), the Wellcome Trust (072894/Z/03/Z, 090532/Z/09/Z and 075491/Z/04/B) and NIMH grants (MH 41953 and MH083094).

Chronic Obstructive Pulmonary Disease (COPD)

These data from the Genetic Epidemiology of COPD study were accessed through dbGaP – study accession phs000179.v3.p2. SNP genotyping data was obtained from 998 individuals recruited for a genome-wide association study of chronic obstructive pulmonary disease (COPD). These participants had COPD or were control individuals who smoked, and were recruited in the USA. They were aged 45-80 years, white or African American and approximately 50% were male. The principal investigators were James D Crapo (National Jewish Health, Denver, CO, USA) and Edwin K Silverman (Brigham and Women's Hospital, Boston, MA, USA). The study was run at the National Heart, Lung and Blood Institute, Bethesda, MD, USA and funded by the National Institutes of Health, Bethesda, MD USA (U01HL089897, U01HL089856).

Corneal Dystrophy

These data were accessed through dbGaP – study accession dbGAP phs000421.v1.p1. SNP genotyping data was obtained from 3,640 individuals recruited for a genome-wide association study of Fuch’s Endothelial Corneal Dystrophy (FECD). Participants had FECD or were control individuals, and were recruited in the USA. Participants were 60 years of age or older and approximately 33% were male. This sample comes from the Genome-Wide Association Study of Fuch’s Endothelial Corneal Dystrophy (FECD) held in dbGAP. The principal investigators were Natalie Afshari (Duke University, Durham, NC, USA), John Gottsch (Johns Hopkins University, Baltimore, MD, USA), Sudha K Iyengar (Case Western Reserve University, Cleveland, OH, USA), Nicholas Katsanis (Johns Hopkins University, Baltimore, MD, USA), Gordon Klintworth (Duke University, Durham, NC, USA) and Jonathan Lass (Case Western Reserve University, Cleveland, OH, USA). Co-investigators were Simon Gregory (Duke University, Durham, NC, USA) and Yi-Ju Li (Duke University, Durham, NC, USA). The study was funded by the National Eye Institute, National Institutes of Health, Bethesda, MD, USA (R01EY016482, CWRU, PI: Sudha Iyengar; R01EY016514, DUEC, PI: Gordon Klintworth; R01EY016835, JHU, PI: John Gottsch). Genotyping was carried out at the Center for Inherited Disease Research (CIDR), Johns Hopkins University, Baltimore, MD, USA and was funded by the National Institutes of Health, Bethesda, MD, USA. HHSN268200782096C, NIH contract "High throughput genotyping for studying the genetic contributions to human disease"; HHSN268201100011I, NIH contract "High throughput genotyping for studying the genetic contributions to human disease".

Mammography

These data were accessed through dbGaP – study accession dbGAP phs000395.v1.p1. SNP genotyping data was obtained from 974 women undergoing mammography through participation in the San Francisco Mammography Registry as part of the Breast Cancer Screening Consortium. This sample comes from the California Pacific Medical Center Research Breast Health Cohort held in dbGAP. The principal investigator was Elad Ziv (University of California, San Francisco, CA, USA). Co-investigators were Steven Cummings (California Pacific Medical Center Research Institute and University of California, San Francisco, CA, USA), Karla Kerlikowske (University of California, San Francisco, CA, USA) and John Shepherd (University of California, San Francisco, CA, USA). The study was run at the National Cancer Institute, National Institutes of Health, Bethesda, MD, USA and was funded by the National Institutes of Health, Bethesda, MD, USA (P01 CA107584; R01 CA120120). Genotyping was carried out at Johns Hopkins University Center for Inherited Disease Research (CIDR), Baltimore, MD, USA and was funded by the National Institutes of Health, Bethesda, MD, USA (HHSN268200782096C, "NIH contract High throughput genotyping for studying the genetic contributions to human disease"; HHSN268201100011I, "NIH contract High throughput genotyping for studying the genetic contributions to human disease"). Genotyping quality control was carried out at the Genetics Coordinating Center, Dept. of Biostatistics, University of Washington, WA, USA.

Melanoma 2

These data were accessed through dbGaP – study accession dbGAP phs000519.v1.p1. SNP genotyping data was obtained from 2,598 individuals recruited for a genome-wide association study of melanoma. Participants had a diagnosis of cutaneous melanoma or were population-based controls. 52% were male. This sample comes from the Study of Melanoma Risk in Australia and the United Kingdom study held in dbGAP. The principal investigator was Nicholas Hayward (Queensland Institute of Medical Research, Brisbane, QLD, Australia). The study was funded by the National Cancer Institute of the National Institutes of Health, Bethesda, MD, USA (R01CA088363). Genotyping was carried out at the Center for Inherited Disease Research (CIDR), Johns Hopkins University, Baltimore, MD, USA and funded by the National Institutes of Health, Bethesda, MD, USA (HHSN268201100011I). We refer to this study as Melanoma 2, to point out that it is different from the Melanoma study we used in our previous work (8).

Cardiff Controls (Blood Donors)

878 blood donors in the UK were recruited by Cardiff University and blood donation centers. These individuals were not screened for psychiatric disorders. However, in the UK, individuals cannot donate blood if they are taking medications. Therefore, individuals in this dataset are likely to be healthy. 57% of individuals were male.

CNV Calling Criteria

Table S4. Criteria used for calling CNVs. CNVs at *EHMT1* and *SHANK3* were required to intersect at least 1Mbp distance, as small deletions and duplications were found to be common in samples with poor QC criteria, indicating that small CNVs in these telomeric regions were likely to be false-positives.

CNV	**Criteria**
1p36 del/dup	Size >50% of critical region, affecting *GABRD*
TAR del/dup	Size >50% of critical region
1q21.1 del/dup	Size >50% of critical region
NRXN1 del	Exonic deletions
2q11.2 del/dup	Size >50% of critical region, affecting both *LMAN2L* and *ARID5A*
2q13 del/dup	Size >50% of critical region
2q13 del/dup (*NPHP1*)	Size >50% of critical region, affecting *NPHP1*
2q21.1 del/dup	Size >50% of critical region
2q37 del/dup (*HDAC4*)	Size >50% of critical region, affecting *HDAC4*
3q29 del/dup	Size >50% of critical region
Wolf-Hirschhorn del/dup	Size >50% of critical region
Sotos Syn/5q35 dup	Size >50% of critical region
6q16 del/dup (*SIM1*)	Exonic deletions; whole gene duplications
Williams Beuren Syn del/dup	Size >50% of critical region
7q11.23 distal del/distal dup	Size >50% of critical region
8p23.1 del/dup	At least 1Mbp of critical region
9q34 del/dup (*EHMT1*)	At least 1Mbp CNVs, including *EHMT1*
10q11.21q11.23 del/dup	Size >50% of critical region
10q23 del/dup	At least 1Mbp, including *NRG3* and *GRID1*
Potocki-Shaffer Syn del/11p11.2 dup (*EXT2*)	Size >50% of critical region, including *EXT2*
13q12 del/dup (*CRYL1*)	Exonic deletions; whole gene duplications
13q12.12 del/dup	Size >50% of critical region
15q11.2 del/dup	Size >50% of critical region
PWS del/dup	Full critical region, ~4Mbp

CNV	**Criteria**
15q11q13 del/dup BP3-BP4	Size >50% of critical region
15q11q13 del/dup BP3-BP5	Size >50% of critical region
15q13.3 del/dup	Size >50% of critical region
15q13.3 del/dup (*CHRNA7*)	Size >50% of critical region, affecting *CHRNA7*
15q24 del/dup	At least 1Mbp between the A-E intervals
15q25 del/dup	At least 1Mbp between the A-D intervals
Rubinstein-Taybi del/dup (*CREBBP*)	Exonic deletions; whole gene duplications
16p13.11 del/dup	Size >50% of critical region
16p12.1 del/dup	Size >50% of critical region
16p12.2-p11.2 del/dup (7.1-8.7 Mb)	Size >50% of critical region
16p11.2 distal del/distal dup	Size >50% of critical region
16p11.2 del/dup	Size >50% of critical region
17p13.3 del/dup (*YWHAE*)	Exonic deletions; whole gene duplications
17p13.3 del/dup (*PAFAH1B1*)	Exonic deletions; whole gene duplications
17p12 del (HNPP)/dup (CMT1A)	Size >50% of critical region, affecting *PMP22*
Smith-Magenis/Potocki-Lupski Syn	Size >50% of critical region
17q11.2 del/dup (*NF1*)	Size >50% of critical region, affecting *NF1*
17q12 del/dup	Size >50% of critical region
17q21.31 del/dup	Size >50% of critical region
17q23.1q23.2 del/dup	Size >50% of critical region
22q11.2 del/dup	Size >50% of critical region
22q11.2 distal del/dup	Size >50% of critical region
SHANK3 del/dup	At least 1Mbp CNVs, including *SHANK3*
"Large" CNVs	Size > 20Mbp + >50 genes

Analysis of Cognitive Tests

Analyses were carried out on data from cognitive tests performed by UK Biobank participants at their first assessment at the assessment centers, or at follow-up on home computers. Details of these tests are available on the UK Biobank website (http://biobank.ctsu.ox.ac.uk/crystal/). All participants included in these analyses were of white British and Irish ancestry.

Pairs Matching Test

The Pairs Matching Test examines episodic memory. In this test, participants were shown cards displaying symbols for 3 seconds (6 cards in the training round and 12 cards in round 2, the testing round). The cards were then turned over and the participant's task was to identify all correct pairs in as few tries as possible. Data collected included the number of correct and incorrect matches per round and the time taken to complete the round. We used the number of incorrect matches in round 2 (field 399) of the Pairs Matching Test (field 100030) as our outcome. We examined only results from individuals who achieved 6 correct matches in order to exclude results on people who did not complete the test. These results were not normally distributed (Figure S5A). We applied a log+1 transformation (Figure S5B) before converting them to z scores (Figure S5C).

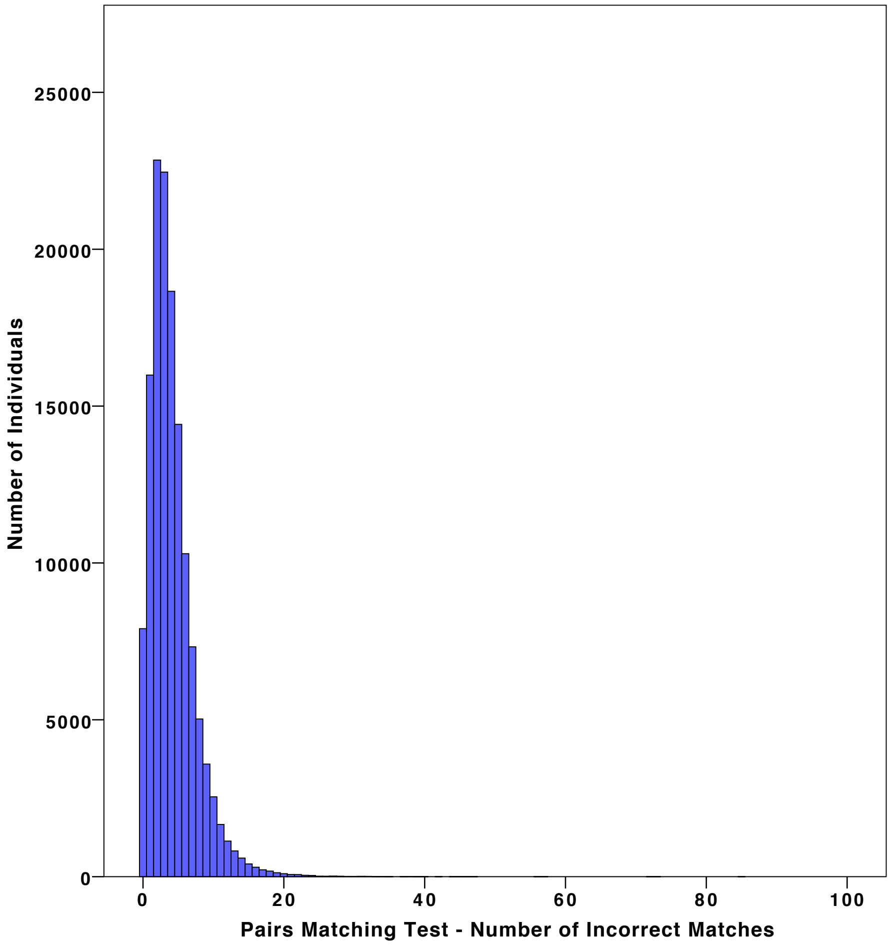


Figure S5A. Distribution of results for the number of incorrect matches in the Pairs Matching Test.

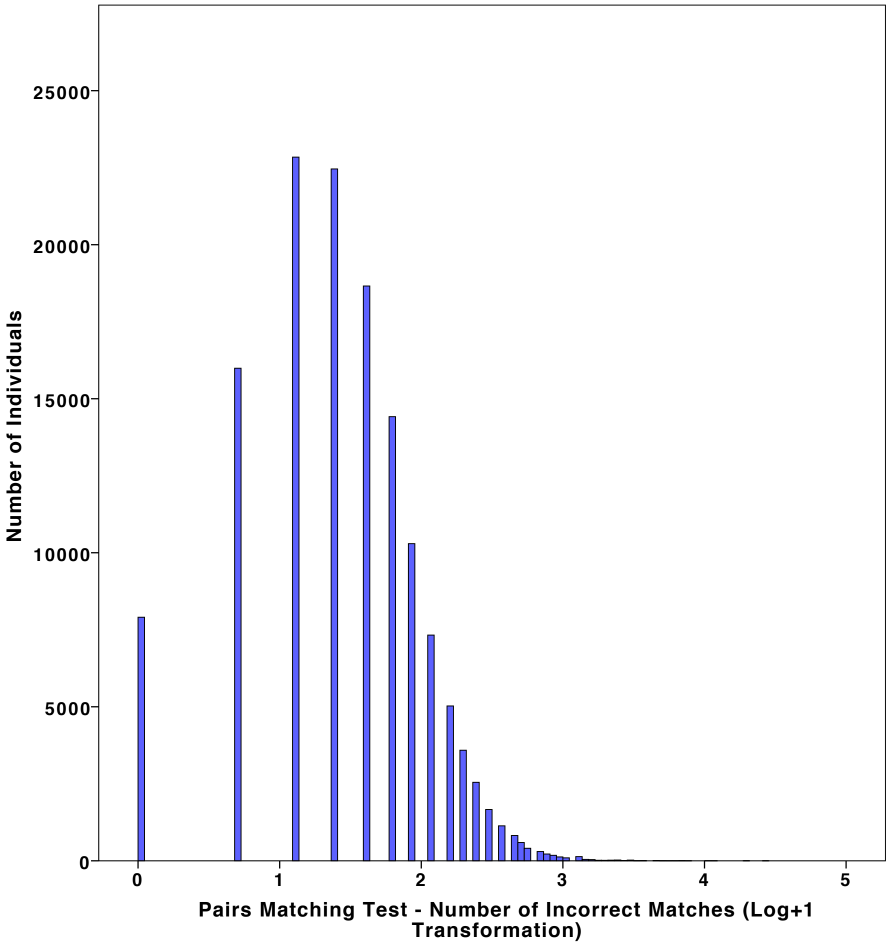


Figure S5B. Distribution of results for the number of incorrect matches in the Pairs Matching Test after log+1 transformation. The resulting values for skewness = -0.334 and kurtosis = 0.143 are considered acceptable in order to prove normal univariate distribution (15).

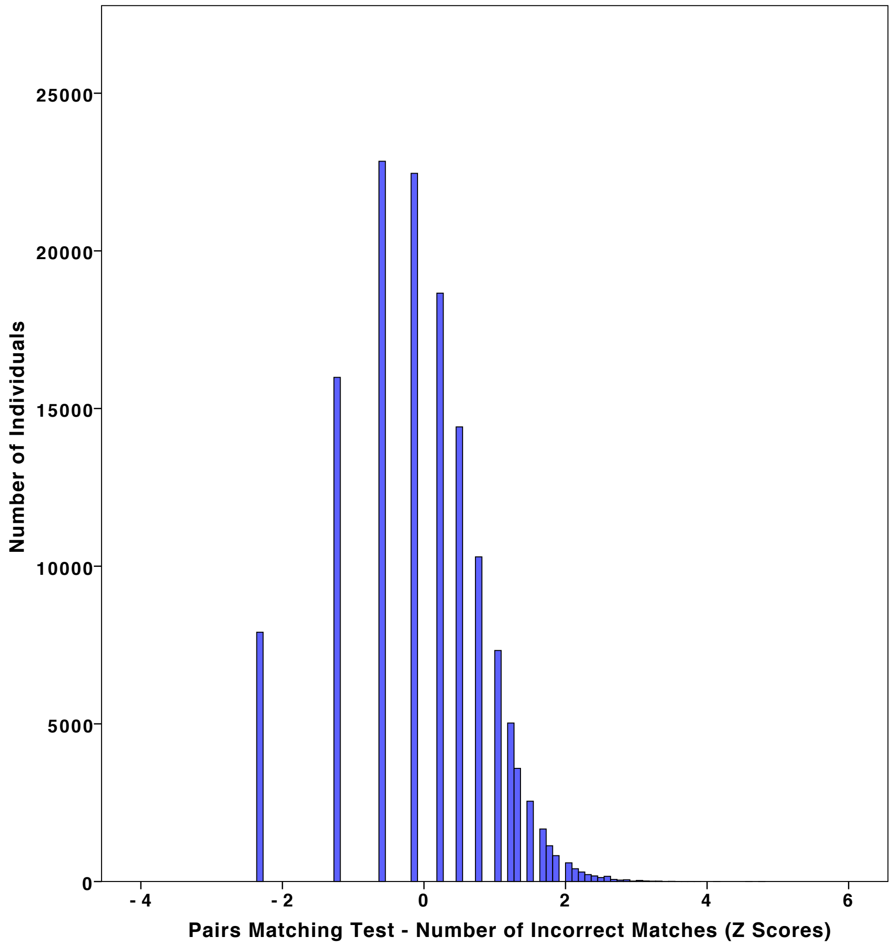


Figure S5C. Distribution of the log+1 transformed results for the number of incorrect matches in the Pairs Matching Test after conversion to z scores.

We compared the results (z scores) between both carriers of schizophrenia CNVs (n = 1,048) and other neurodevelopmental CNVs (n = 463) with non-CNV carriers (134,781) in linear regression analyses corrected for age and sex (Table 1 main text). We also compared the results between individuals with schizophrenia and those without the disorder in a linear regression analysis, corrected for age and sex (Table 1 main text).

Reaction Time Test

The Reaction Time Test examines simple processing speed. In this test, participants were required to play 12 rounds of the card game 'Snap'. Participants were shown two cards at a time and were required to press a button as quickly as possible if the cards were the same. Data collected included the cards used, the number of times the button was pressed, the duration to the first press of the button in each round and the mean time to correctly identify matches. We used the mean time to correctly identify matches as our outcome (field 20023). These results were not normally distributed (Figure S6A). We excluded outlying scores (<100ms and >1500ms) and applied a log transformation to the results (Figure S6B) before converting them to z scores (Figure S6C).

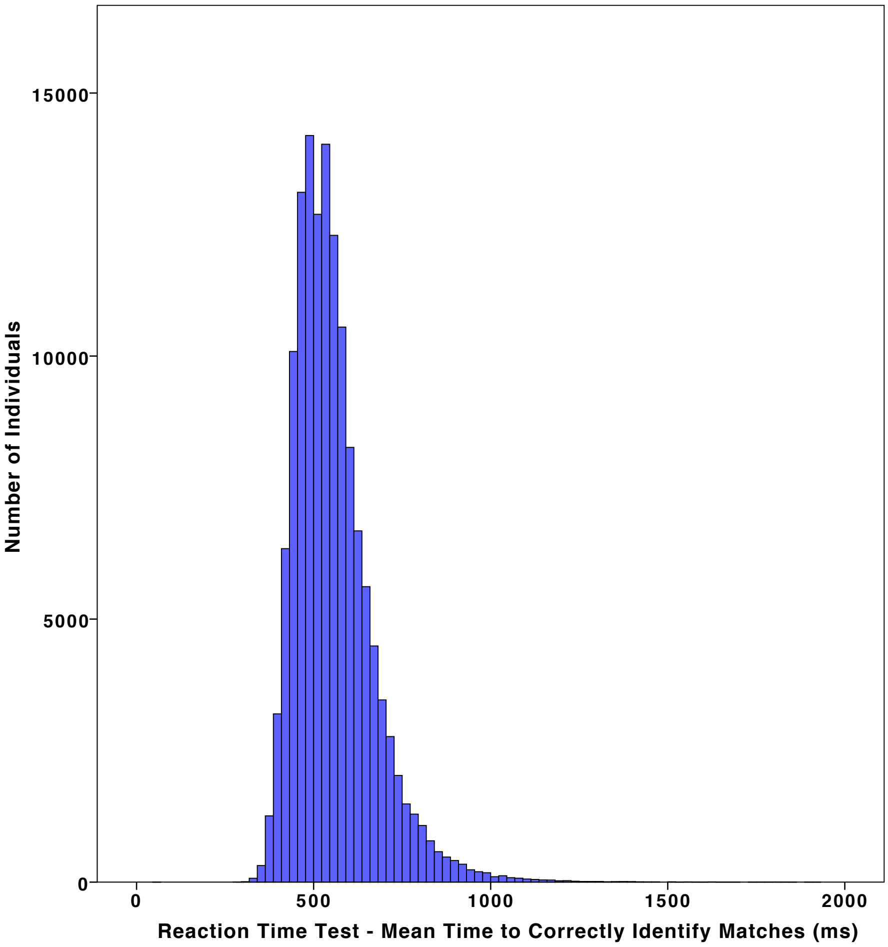


Figure S6A. Distribution of results for the mean time to correctly identify matches in the Reaction Time Test in milliseconds. All results are shown.

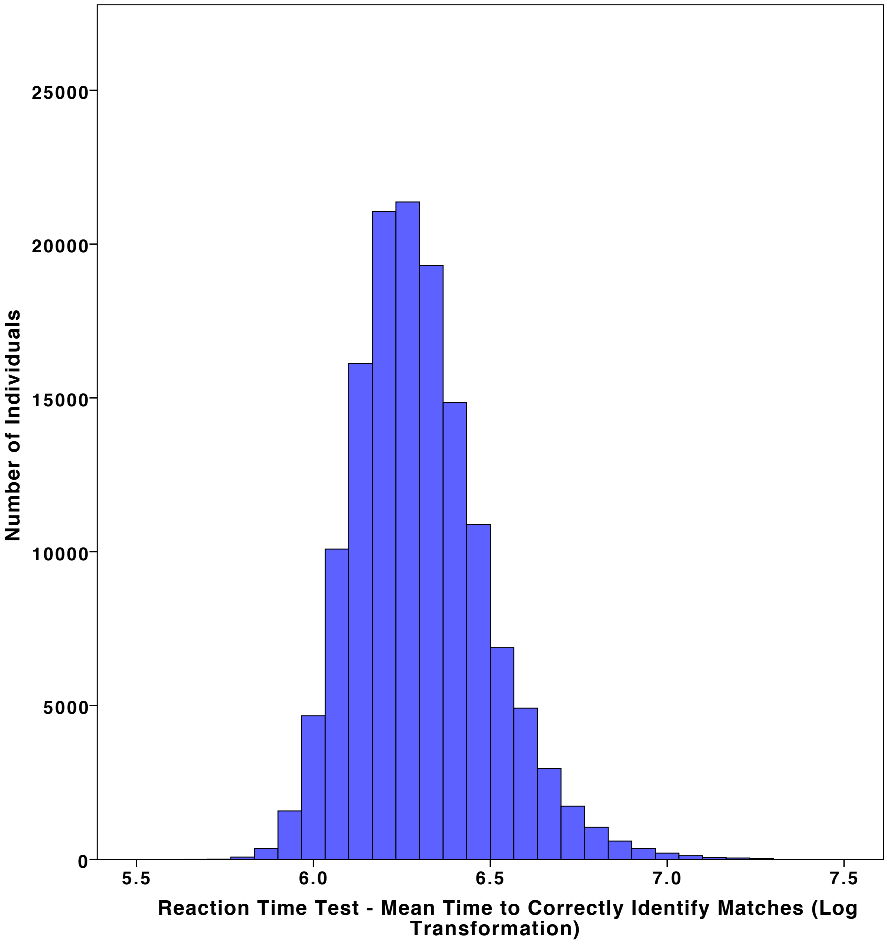


Figure S6B. Distribution for the mean time to correctly identify matches in the Reaction Time Test after log transformation. Skewness = 0.692 and kurtosis = 0.985.

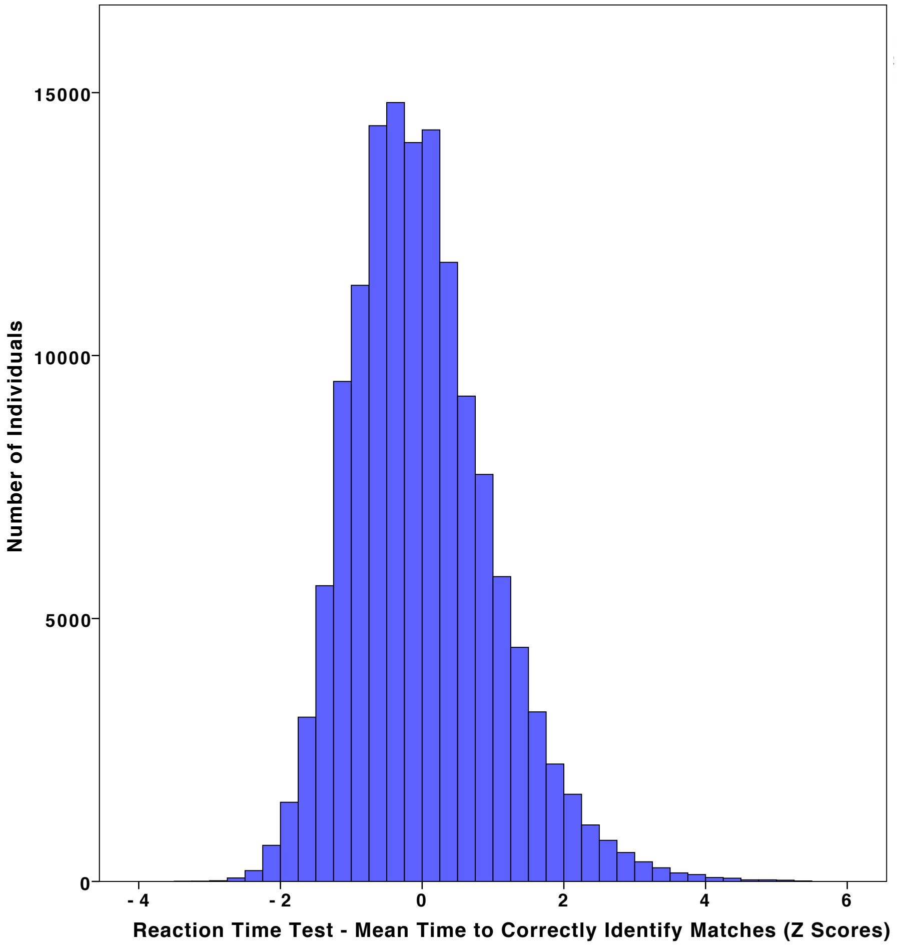


Figure S6C. Distribution of the log transformed results for the mean time to correctly identify matches in the Reaction Time Test after conversion to z scores.

We compared the results (z scores) between both carriers of schizophrenia CNVs (n = 1,073) and other neurodevelopmental CNVs (n = 477) with non-CNV carriers (n = 137,053) in linear regression analyses corrected for age and sex (Table 1 main text). We also compared the results between individuals with schizophrenia and those without the disorder in a linear regression analysis, corrected for age and sex (Table 1 main text).

Fluid Intelligence Test

The Fluid Intelligence Test examines reasoning and problem solving ability of participants. In this test, participants were asked to complete as many questions as possible within 2 minutes (maximum score = 13). Data collected included participants' results for each question, the number of questions attempted within the time limit and participants' total scores. We used the number of correct answers (field 20016) as our outcome. These results were normally distributed (Figure S7A). We converted them to z scores (Figure S7B).

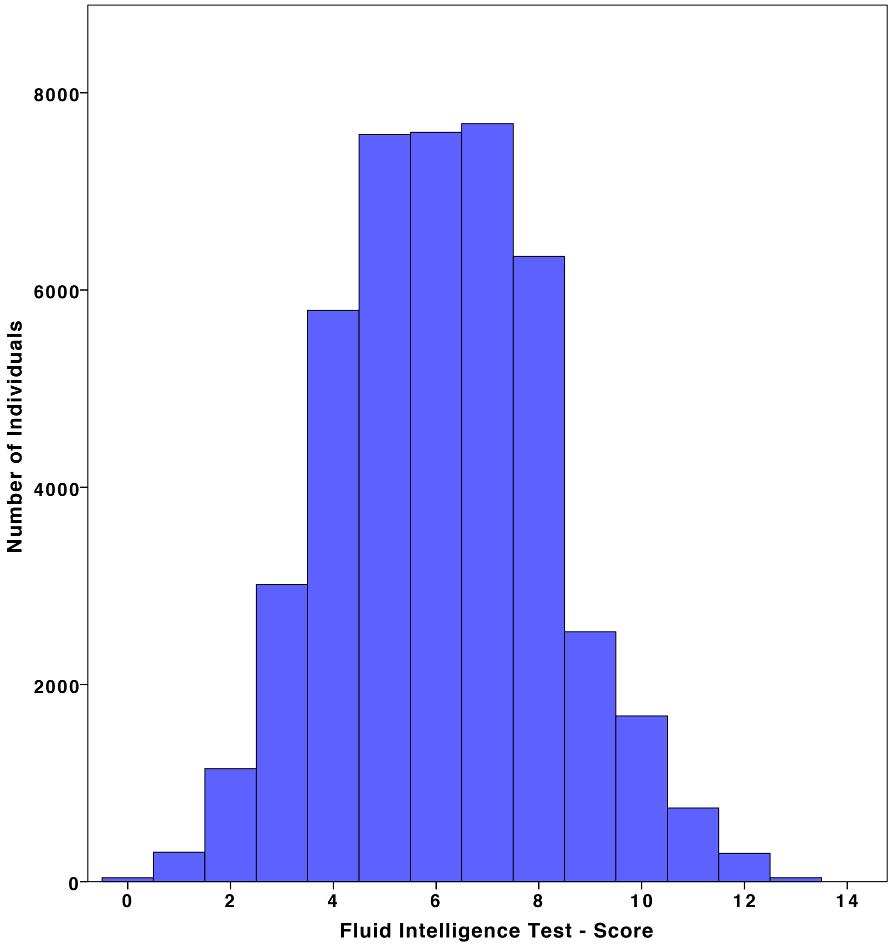


Figure S7A. Distribution of scores in the Fluid Intelligence Test. Skewness = 0.169 and kurtosis = -0.147.

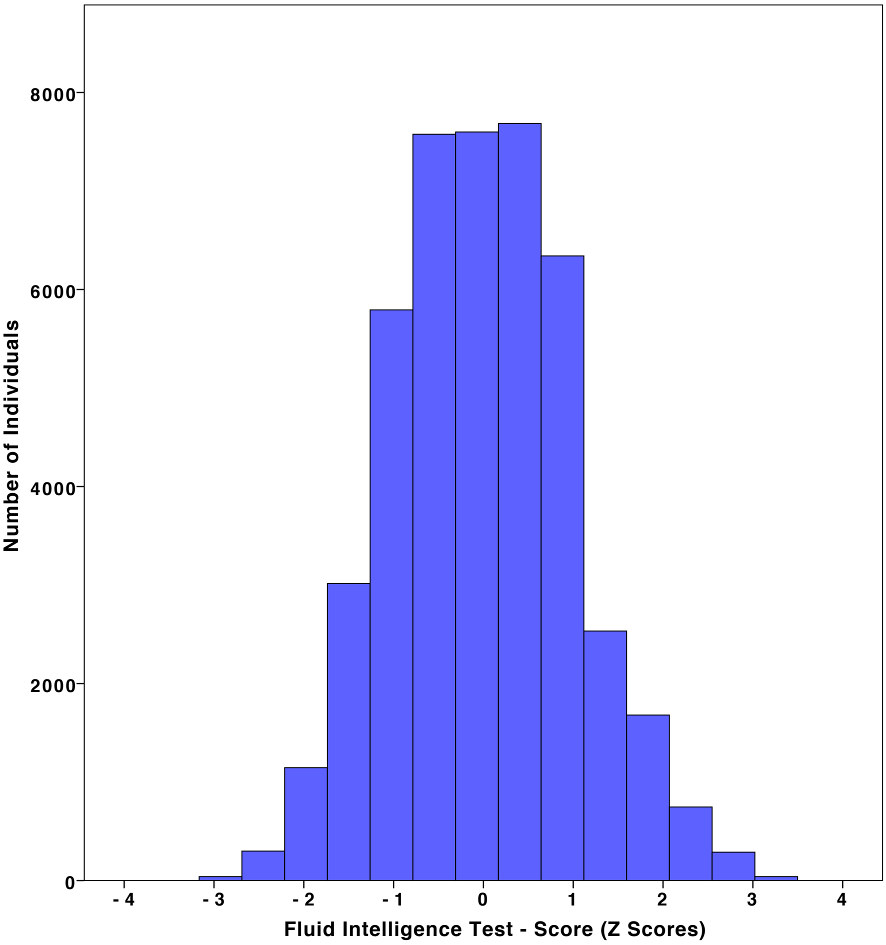


Figure S7B. Distribution of scores in the Fluid Intelligence Test after conversion to z scores.

We compared the results (z scores) between both carriers of schizophrenia CNVs (n = 321) and other neurodevelopmental CNVs (n = 147) with non-CNV carriers (n = 44,107) in linear regression analyses corrected for age and sex (Table 1 main text). We also compared the results between individuals with schizophrenia and those without the disorder in a linear regression analysis, corrected for age and sex (Table 1 main text).

Figure S7c shows the distribution of the raw scores on the 13 questions, separately for carriers of schizophrenia CNVs, other neurodevelopmental CNVs and non-CNV carriers. The distributions are similar, with a shift towards lower scores amongst carriers of schizophrenia and other neurodevelopmental CNVs.

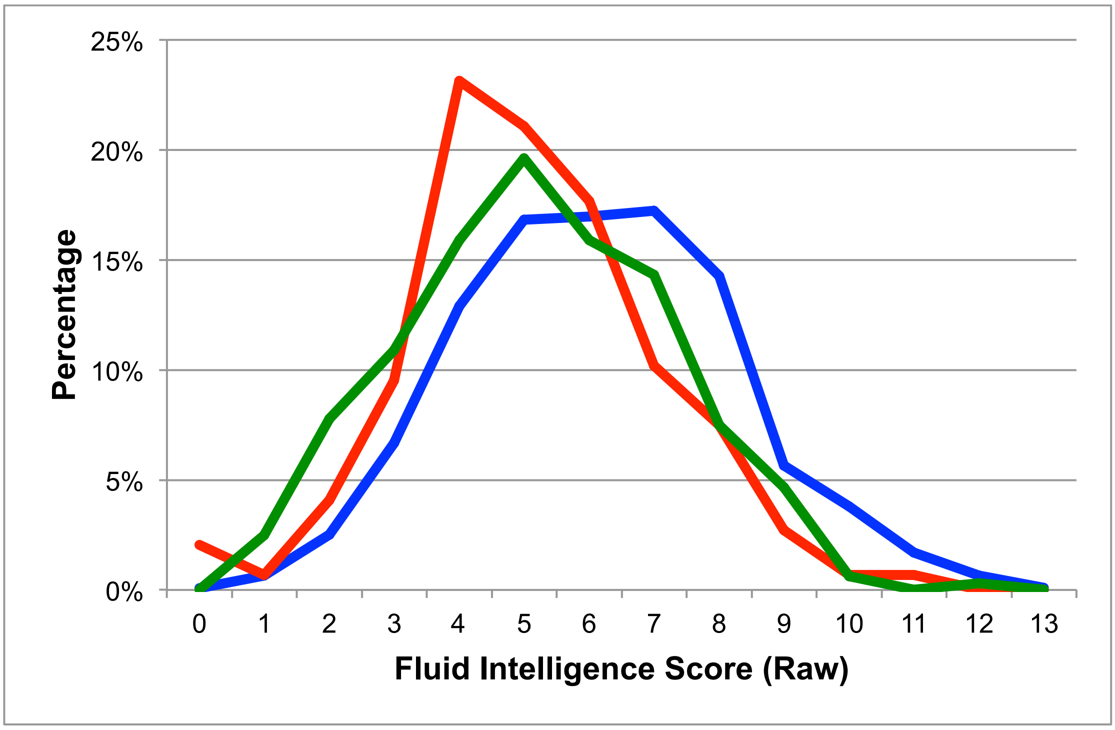


Figure S7C. Distribution of scores in the Fluid Intelligence Test in carriers of schizophrenia CNVs (green), carriers of other neurodevelopmental CNVs (red) and non-CNV carriers (blue).

Digit Span (Numeric Memory Test)

The Digit Span Test examines working memory. In this test, participants were shown a two-digit number, which then disappeared from the screen. Participants were asked to enter the number they had seen. The number became one digit longer in each round (maximum 12). Data collected included target numbers to be memorized, the number of correct entries, time to complete the test and the maximum number of digits remembered correctly. We used the maximum number of digits remembered correctly (field 4282) as our outcome. These results were normally distributed (Figure S8A). We converted the results to z scores (Figure S8B).

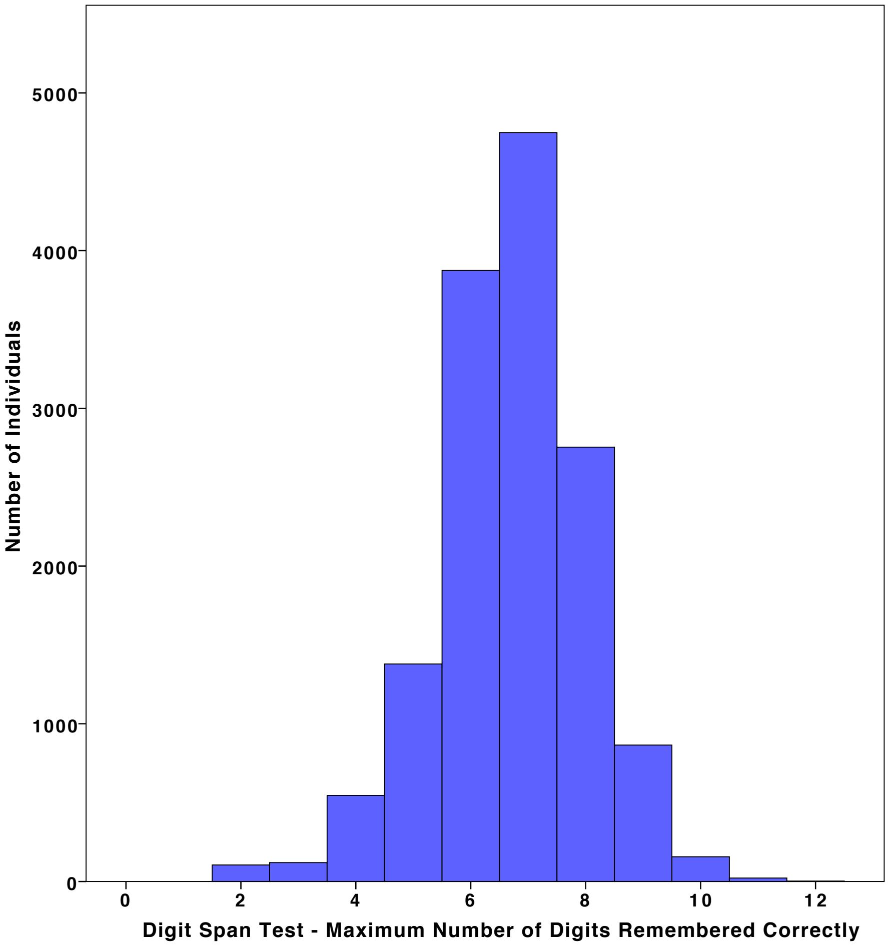


Figure S8A. Distribution of results for the maximum number of digits remembered correctly in the Digit Span Test (Numeric Memory Test). Skewness = -0.355 and kurtosis = 0.896.

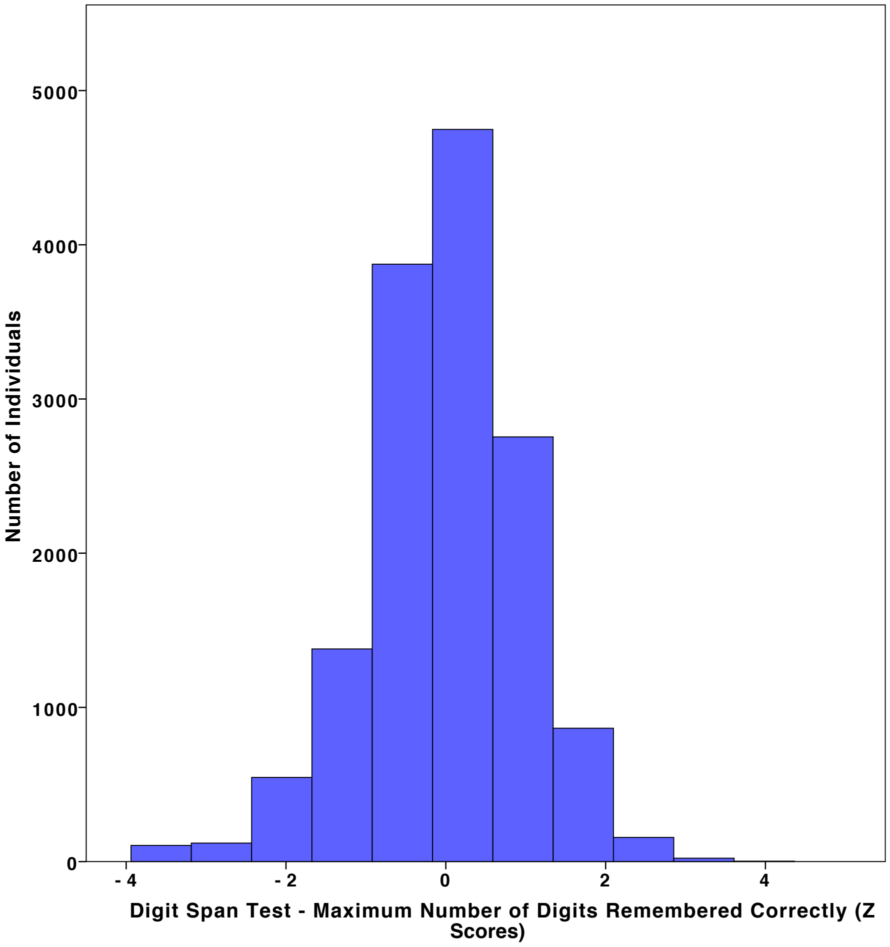


Figure S8B. Distribution of results for the maximum number of digits remembered correctly in the Digit Span Test (Numeric Memory Test) after conversion to z scores.

We compared the results (z scores) between both carriers of schizophrenia CNVs (n = 102) and other neurodevelopmental CNVs (n = 50) with non-CNV carriers (n = 14,343) in linear regression analyses corrected for age and sex (Table 1 main text). We also compared the results between individuals with schizophrenia and those without the disorder in a linear regression analysis, corrected for age and sex (Table 1 main text).

Symbol Digit Substitution Test

The Symbol Digit Substitution Test examines complex processing speed. In this test, participants were asked to play a code-breaking game. Participants were required to replace symbols using a number pad within 2 minutes. Data collected included the number of matches attempted and the number of correct matches. We used the number of symbol digit matches made correctly (field 20159) as our outcome. These results were not normally distributed (Figure S9A). We excluded outlying scores (0-1 and >36) and converted the remaining scores to z scores (Figure S9B).

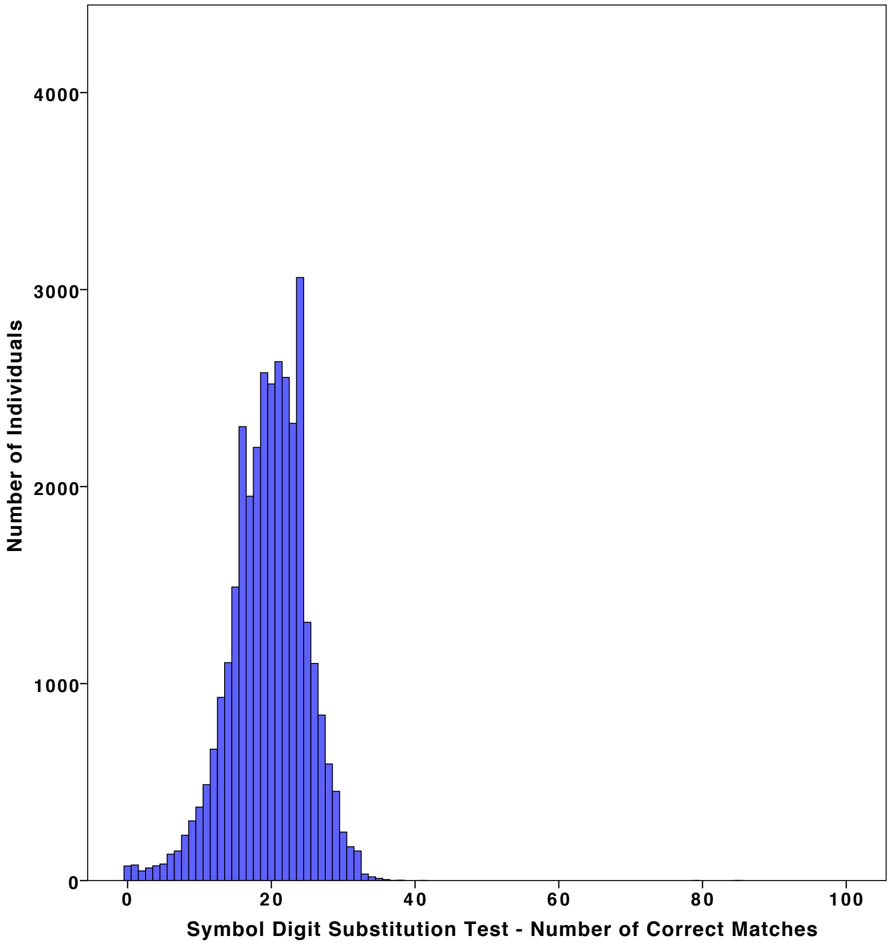


Figure S9A. Distribution of results for the number of correct matches in the Symbol Digit Substitution Test.

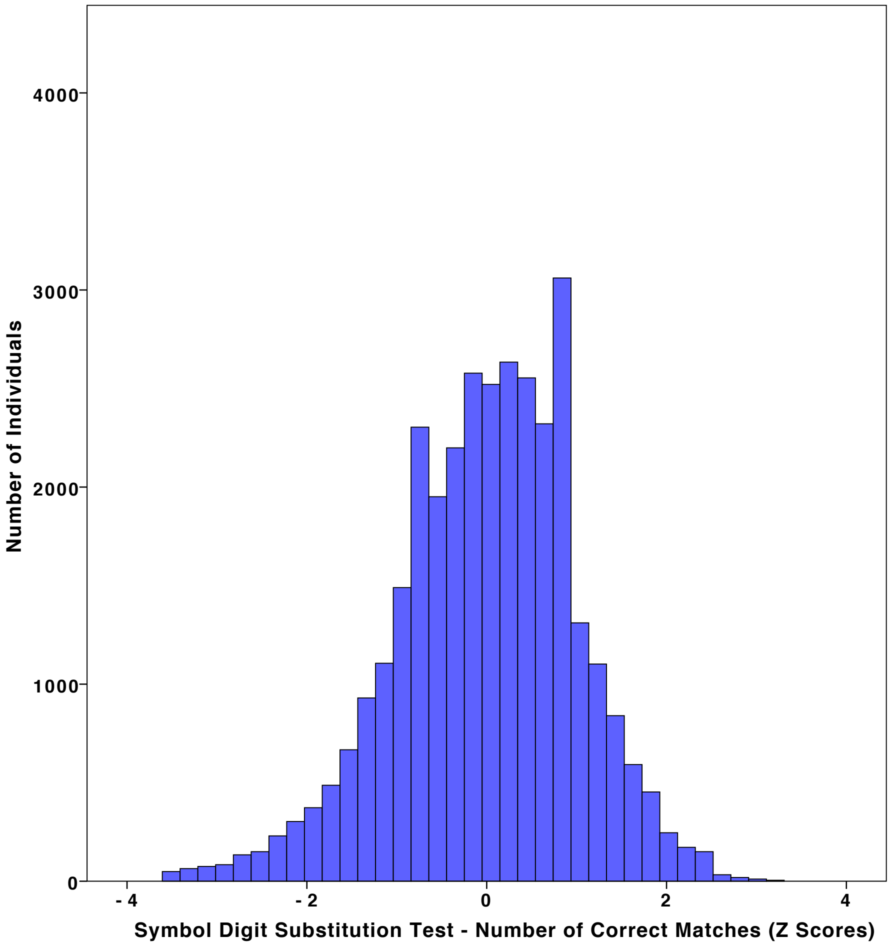


Figure S9B. Distribution of results for the number of correct matches in the Symbol Digit Substitution Test after conversion to z scores. Skewness = -0.359 and kurtosis = 0.338.

We compared the results (z scores) between both carriers of schizophrenia CNVs (n = 204) and other neurodevelopmental CNVs (n = 83) with non-CNV carriers (n = 32,770) in linear regression analyses corrected for age and sex (Table 1 main text). We also compared the results between individuals with schizophrenia and those without the disorder in a linear regression analysis, corrected for age and sex (Table 1 main text).

Trail Making Test

The Trail Making Test examines visual attention and consists of two parts – numeric (A) and alphanumeric (B). In Trail Making Test A, participants must link circles with digits in the correct numeric order. In Trail Making Test B, participants must link circles by alternating between numbers and letters of the alphabet. Data collected included the total errors made per test, intervals between points and the total time taken to complete each test. We used the duration to complete the test as our outcome on both tests (TMT A – field 20156, TMT B – field 20157). These results were not normally distributed (Figures S10A and S11A). We applied a log transformation to the results (Figures S10B and S11B) before converting them to z scores (Figures S10C and S11C).

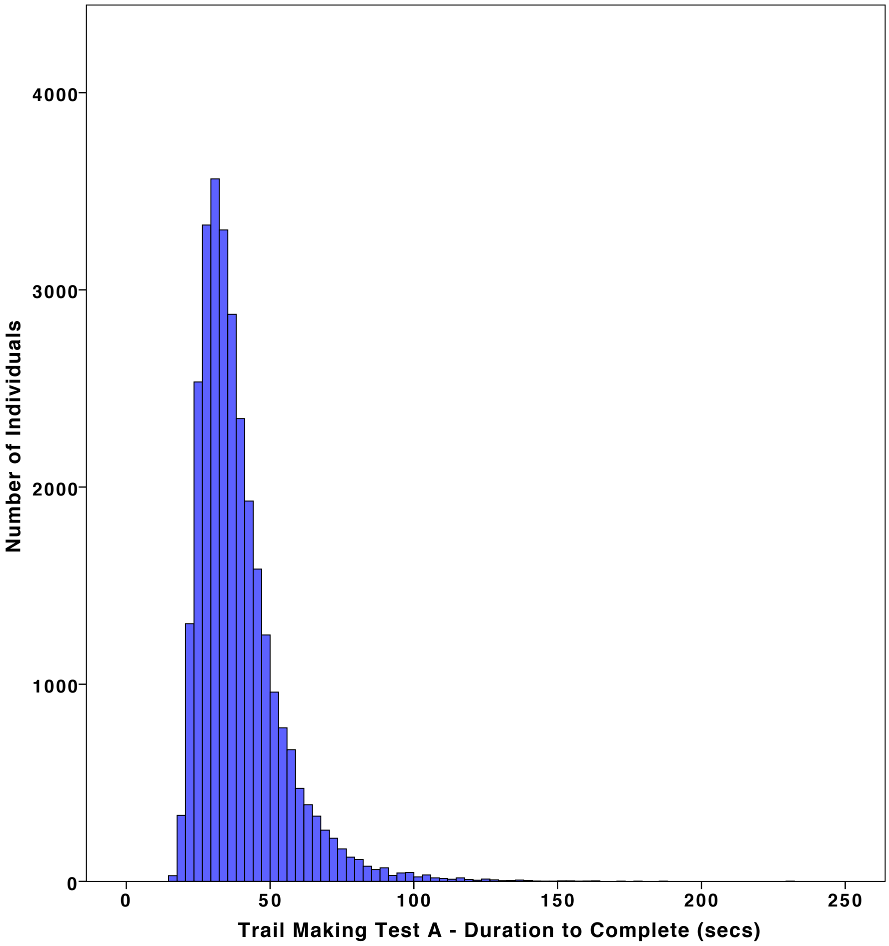


Figure S10A. Distribution of results for the duration to complete Trail Making Test A.

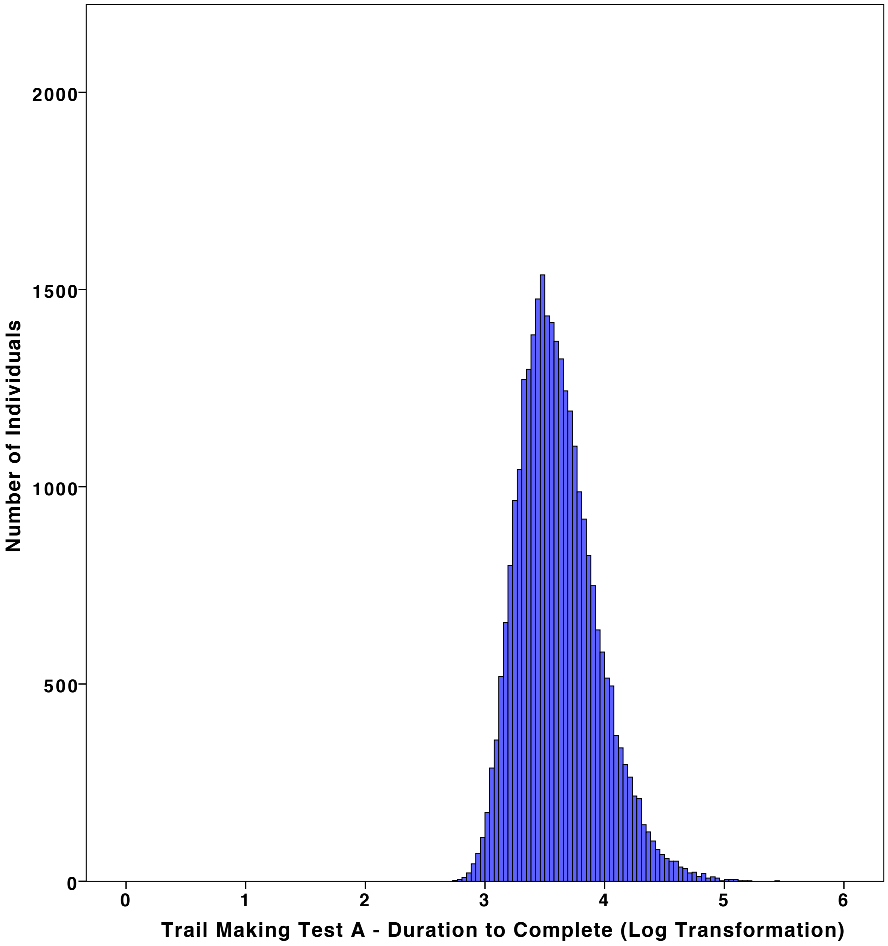


Figure S10B. Distribution of results for the duration to complete Trail Making Test A after log transformation. Skewness = 0.665 and kurtosis = 0.624).

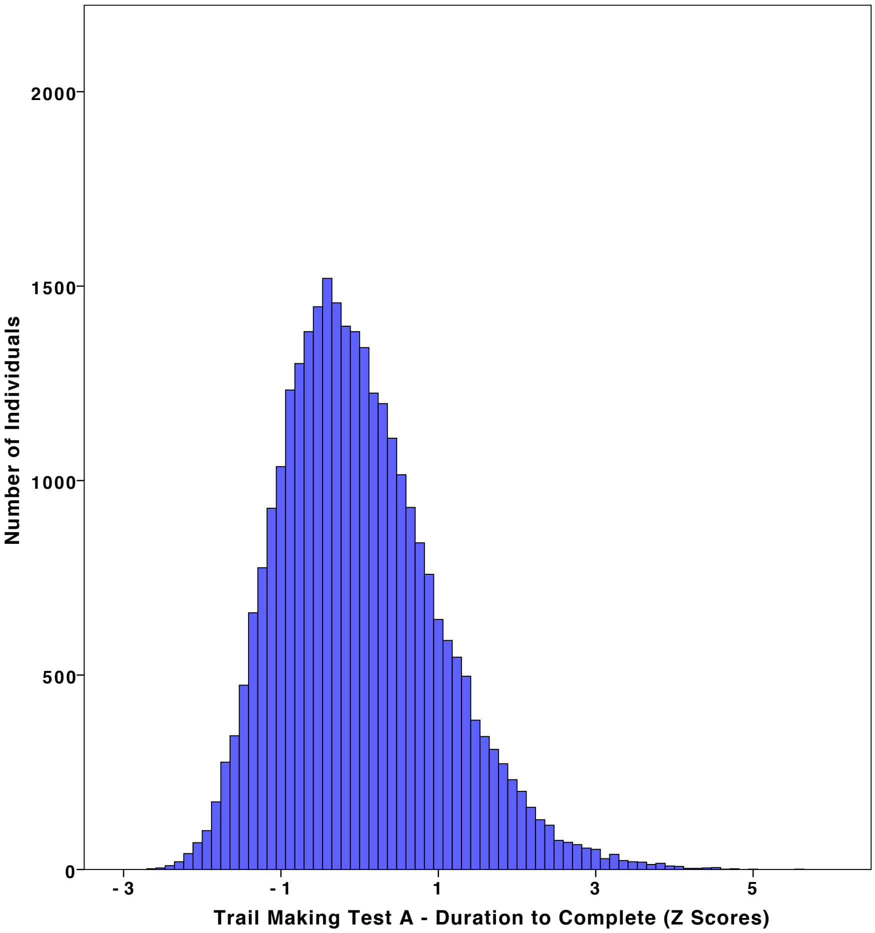


Figure S10C. Distribution of results for the duration to complete Trail Making Test A after conversion to z scores.

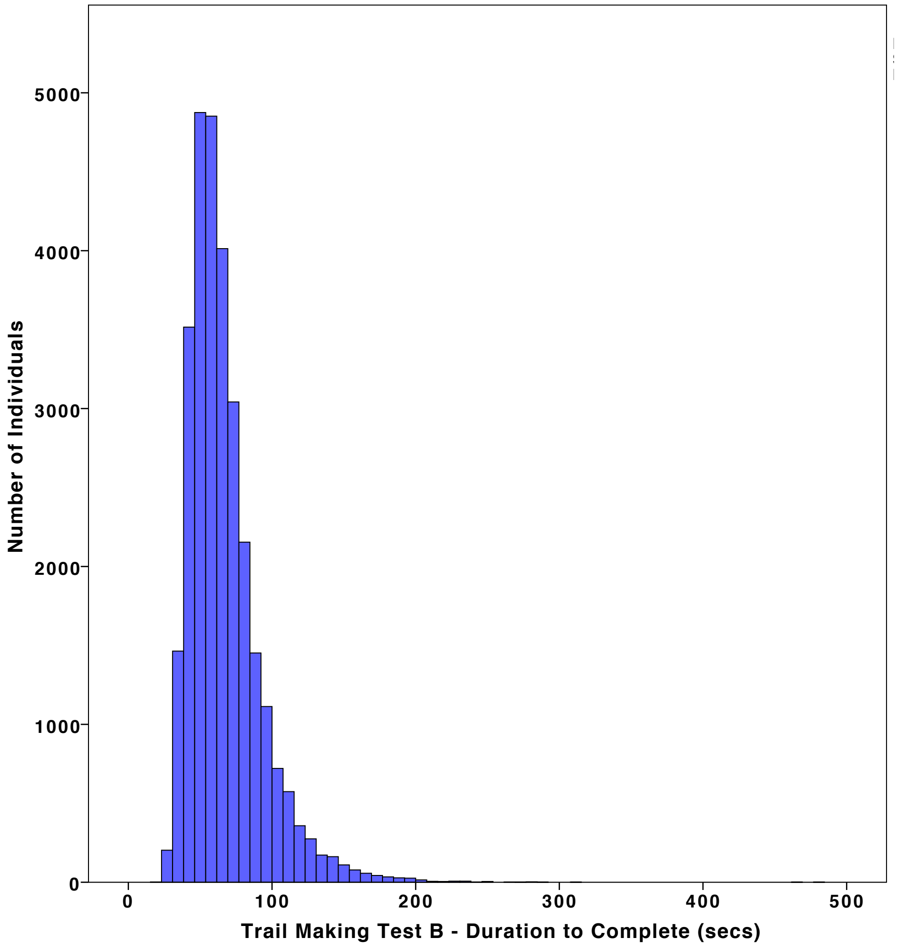


Figure S11A. Distribution of results for the duration to complete Trail Making Test B.

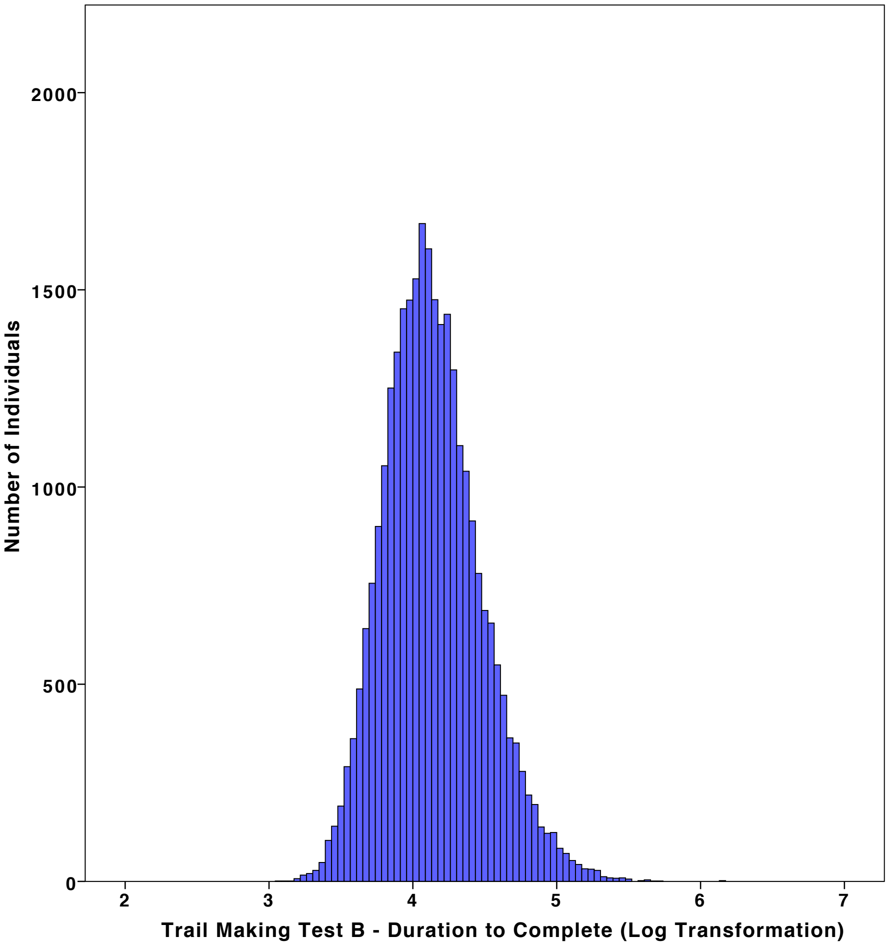


Figure S11B. Distribution of results for the duration to complete Trail Making Test B after log transformation. Skewness = 0.482 and kurtosis = 0.449.

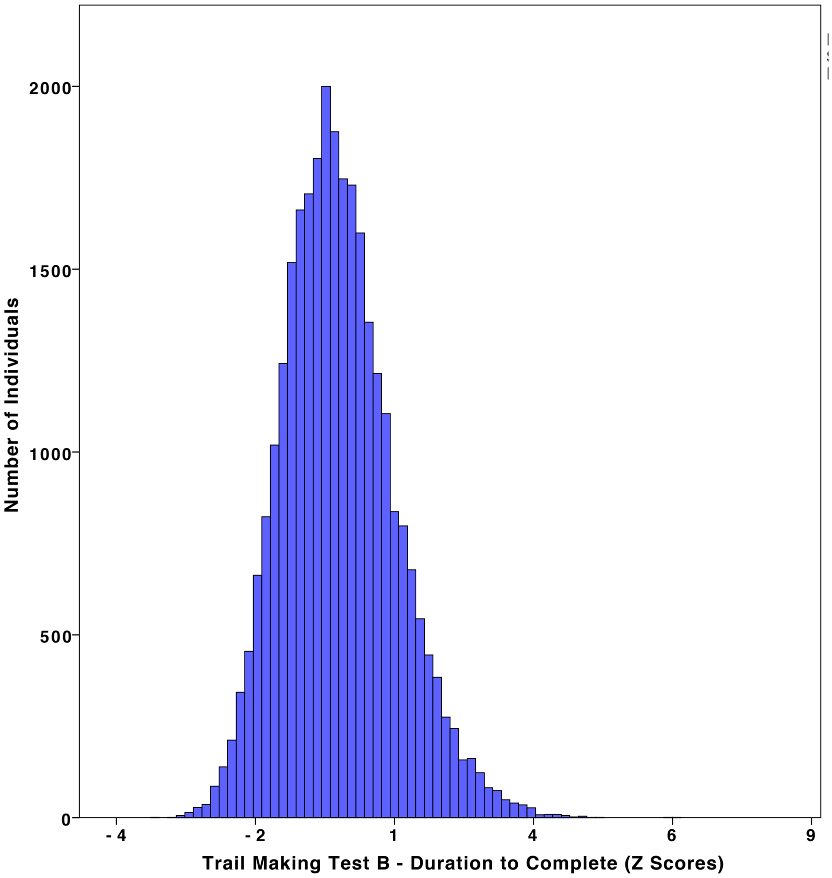


Figure S11C. Distribution of results for the duration to complete Trail Making Test B after conversion to z scores.

We compared the results (z scores) between both carriers of schizophrenia CNVs (n = 185) and other neurodevelopmental CNVs (n = 78) with non-CNV carriers (n = 28,988) in linear regression analyses corrected for age and sex (Table 1 main text). We also compared the results between individuals with schizophrenia and those without the disorder in a linear regression analysis, corrected for age and sex (Table 1 main text).

Educational Attainment

During the collection of socio-demographic information, participants were asked “which of the following qualifications do you have?” and options were presented in a list (Table S5). Due to the inexact relationship of “other professional qualifications” to the other groups, we excluded individuals who only reported such "other qualifications" (n = 7,210). We excluded those who had chosen not to answer this question (n = 1,213) and recoded -7 "none of the above" to a new group 6, assuming they didn't obtain a qualification. In this way we created an approximate gradient of qualifications.

Table S5. Coding of qualifications and our recoding of this variable.

Original UK Biobank Groupings		Recoded	
1	College/University degree	1	College/University degree
2	A/AS levels or equivalent	2	A/AS levels or equivalent
3	O levels/GCSEs or equivalent	3	O levels/GCSEs or equivalent
4	CSEs or equivalent	4	CSEs or equivalent
5	NVQ or HND or HNC or equivalent	5	NVQ or HND or HNC or equivalent
6	Other professional qualifications e.g. nursing or teaching	x	Excluded
-3	Prefer not to answer	x	Excluded
-7	None of the above	6	None of the above

A/AS levels are qualifications taken at 16-18 years of age, post-compulsory education. O levels/GCSEs are qualifications taken at 14-16 years of age at the end of compulsory education. CSEs were a predecessor to GCSEs, which included vocational subjects. NVQs/HNDs/HNCs are vocational qualifications.

Occupational Attainment

Participants were asked their job title and work history. Job codes were allocated according to the Office of National Statistics (ONS) Standard Occupational Classification (16). We used the major job categories from this classification for our analysis (Table S6).

Table S6. ONS major job categories (16).

Group Number	Major Job Group
1	Managers and Senior Officials
2	Professional Occupations
3	Associate Professional and Technical Occupations
4	Administrative and Secretarial Occupations

5	Skilled Trades Occupations
6	Personal Service Occupations
7	Sales and Customer Service Occupations
8	Process, Plant and Machine Operatives
9	Elementary Occupations

In putting together the Standard Occupational Classification, the ONS grouped jobs according to skill level. The length of time necessary to become competent at a job was used as an approximation of skill level. Within the broad structure of the classification, the ONS made reference to four skill levels:

1. Competence associated with general education, usually acquired by the time a person completes their compulsory education.
2. Knowledge provided via a good general education (as for 1) but with a longer period of work-related training.
3. Knowledge associated with a period of post-compulsory education but not to degree level.
4. Degree or equivalent work experience (16).

We wondered whether the reductions in educational/occupational attainment associated with carrying a CNV (schizophrenia or other neurodevelopmental) might be explained by reduced cognitive performance. To answer this question, we carried out logistic regression analyses of the effect of CNV carrier status on both educational and occupational attainment, with and without Fluid Intelligence Test score included as a covariate (as it has the largest effect size) (Table S7). Prior to analysis, we dichotomized educational attainment, putting college/university degree into a single group and the remaining categories into a single group. For occupational attainment, we put managerial and professional occupations into a single group and the remaining occupational categories into a single group. This was done in order to allow logistic regression analysis (although we are aware that this causes some loss of power, reflected in the more modest p-values, compared to the results from ordinal regression presented in the main text). In the case of educational attainment, approximately a half to two thirds of the adverse effect of CNV status was explained by the effect of the Fluid Intelligence Test score. In the case of occupational attainment, the results differed a bit between the carriers of schizophrenia CNVs (about two thirds explained) and the “other CNVs (about one third explained), perhaps due to the smaller sample size. In any case, these results indicate that other phenotypic consequences of having a pathogenic CNV (e.g. medical problems (or cognitive consequences not captured by the Fluid Intelligence Test) also affect educational and occupational attainment.

Table S7. Regression results for the effect of carrying a CNV on educational and occupational attainment with and without Fluid Intelligence Test (FIT) score used as a covariate.

	Without FIT Score Covariate			With FIT Score Covariate		
	B	SE	p	B	SE	p
Educational attainment Schizophrenia CNVs	0.48	0.07	$1.59\text{x}10^{-12}$	0.19	0.13	0.15
Occupational attainment Schizophrenia CNVs	0.39	0.09	$6.0\text{x}10^{-6}$	0.11	0.14	0.42
Educational attainment Other Neurodevelopmental CNVs	0.62	0.10	$1.74\text{x}10^{-9}$	0.18	0.19	0.35
Occupational attainment Other Neurodevelopmental CNVs	0.55	0.14	$6.2\text{x}10^{-5}$	0.40	0.23	0.08

Analysis of Phenotypes of Carriers of Deletions and Duplications at 16p11.2 and 17p12

Table S8. BMI, height and weight in 16p11.2 CNV carriers. The results are expressed as differences in weight and height (kg and cm), after linear regression analyses controlled for age and sex. Deletion carriers have increased BMI, while duplication carriers have reduced BMI.

	16p11.2 deletions (n = 44)			16p11.2 duplications (n = 42)		
	B	95% CI	p	B	95% CI	p
BMI	7.25	5.82 to 8.68	$2.8x10^{-23}$	-1.81	-3.27 to -0.35	0.015
Height (cm)	-7.29	-9.19 to -5.39	$6.2x10^{-14}$	3.0	1.05 to 4.95	0.003
Weight (kg)	12.36	8.10 to 16.61	$1.3x10^{-8}$	-2.51	-6.87 to 1.85	0.26

Table S9. Linear regression analyses for cognitive test results and 16p11.2 CNV carrier status.

Cognitive Test	**16p11.2 del**			**16p11.2 dup**		
	n	B (SE)	P	n	B (SE)	P
Pairs Matching Test	34	0.325 (0.169)	0.055	42	-0.028 (0.152)	0.854
Reaction Time Test	37	0.636 (0.155)	**$4.2×10^{-5}$**	42	0.684 (0.146)	**$3.0×10^{-6}$**
Fluid Intelligence Test	10	-0.635 (0.315)	**0.044**	12	-1.028 (0.288)	**$3.51×10^{-4}$**
Digit Span	4	-0.823 (0.496)	0.097	3	-1.165 (0.572)	**0.042**
Symbol Digit Substitution	1	-1.096 (0.900)	n.a.	8	-0.477 (0.318)	0.134
TMTA	1	1.057 (0.949)	n.a.	7	1.217 (0.359)	**0.001**
TMTB	1	0.863 (0.923)	n.a.	7	1.309 (0.349)	**$1.77×10^{-4}$**

n – number of CNV carriers, B – unstandardized coefficient (z score difference), SE – standard error, P-value. Significant results are in bold. No p-values are shown where only one person completed the test.

Table S10. Rates of 17p12 deletions and duplications and the frequency of peripheral neuropathy among the genotyped UK Biobank participants. The rate of self-reported peripheral neuropathy in deletion carriers is 32 times higher than in the rest of the population, and it is 133 times higher among duplication carriers.

	17p12 del (HNPP)			17p12 dup (CMT1A)		
	N	N Peripheral Neuropathy (%)	Fisher's Exact P	N	N Peripheral Neuropathy (%)	Fisher's Exact P
Carriers	84	4 (4.8%)	1.0×10^{-6}	45	9 (20%)	3×10^{-17}
Non-carriers	151,575	226 (0.15%)		151,614	221 (0.15%)	

Table S11. Linear regression analyses for cognitive test results in 17p12 CNV carriers. Neither deletion nor duplication carriers show cognitive deficits and a higher proportion of them completed the tests, compared to 16p11.2 carriers.

Cognitive Test	**17p12 del (HNPP)**			**17p12 dup (CMT1A)**		
	n	B (SE)	P	n	B (SE)	P
Pairs Matching Test (Z)	77	0.001 (0.002)	0.62	37	4.37×10^{-5} (0.002)	0.98
Reaction Time Test (Z)	76	-0.001 (0.001)	0.31	39	0.002 (0.002)	0.38
Fluid Intelligence Test (Z)	23	-0.004 (0.003)	0.15	14	-0.002 (0.004)	0.55
Digit Span (Z)	7	0.001 (0.005)	0.79	5	0.009 (0.006)	0.12
Symbol Digit Substitution (Z)	18	-0.001 (0.003)	0.80	13	-0.005 (0.003)	0.14
TMTA (Z)	17	0.005 (0.003)	0.083	11	0.005 (0.004)	0.22
TMTB (Z)	17	0.005 (0.003)	0.085	11	0.002 (0.004)	0.61

n = number of CNV carriers who have completed the test, B – unstandardized coefficient (z scores), SE – standard error, P-value.

Table S12. Numbers of CNVs found in each batch. Batches -1 to -11 are the BiLEVE arrays. Results are shown for all 93 targeted loci but if no CNVs were found at a locus, this locus is not shown. No filtering for ethnicity has been applied here.

See Supplement 2 (Excel file) for Table S12.

Supplemental References

1. Murphy S (2015): Genotyping of 500,000 UK Biobank participants - description of sample processing workflow and preparation of DNA for genotyping. https://biobank.ctsu.ox.ac.uk/crystal/docs/genotyping_sample_workflow.pdf
2. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. (2007): PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res*. 17:1665-1674.
3. Rees E, Walters JT, Chambert KD, O'Dushlaine C, Szatkiewicz J, Richards AL, et al. (2014): CNV analysis in a large schizophrenia sample implicates deletions at 16p12.1 and SLC1A1 and duplications at 1p36.33 and CGNL1. *Hum Mol Genet*. 23:1669-1676.
4. Coe BP, Witherspoon K, Rosenfeld JA, van Bon BW, Vulto-van Silfhout AT, Bosco P, et al. (2014): Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet*. 46:1063-1071.
5. Dittwald P, Gambin T, Szafranski P, Li J, Amato S, Divon MY, et al. (2013): NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits. *Genome Res*. 23:1395-1409.
6. ISC (2008): Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*. 455:237-241.
7. Levinson DF, Duan J, Oh S, Wang K, Sanders AR, Shi J, et al. (2011): Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. *Am J Psychiatry*. 168:302-316.
8. Rees E, Walters JT, Georgieva L, Isles AR, Chambert KD, Richards AL, et al. (2014): Analysis of copy number variations at 15 schizophrenia-associated loci. *Br J Psychiatry*. 204:108-114.
9. Rees E, Kendall K, Pardinas A, Legge S, Pocklington A, Escott-Price A, et al. (2016): Analysis of intellectual disability copy number variants ifor association with schizophrenia. *JAMA Psychiatry.* Published online 17 August 2016.
10. Kirov G, Pocklington AJ, Holmans P, Ivanov D, Ikeda M, Ruderfer D, et al. (2012): De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol Psychiatry*. 17:142-153.
11. International Schizophrenia Consortium (2008): Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*. 455:237-241.
12. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, et al. (2008): Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet*. 40:1253-1260.
13. American Psychiatric Association (1994): *Diagnostic and Statistical Manual of Mental Disorders*. Washington DC: American Psychiatric Press.

14. Sanders AR, Duan J, Levinson DF, Shi J, He D, Hou C, et al. (2008): No significant association of 14 candidate genes with schizophrenia in a large European ancestry sample: implications for psychiatric genetics. *Am J Psychiatry*. 165:497-506.

15. George D, Mallery P (2010): *SPSS for Windows Step by Step: A Simple Guide and Reference (17.0 update)*. 10a ed. Boston: Pearson.

16. Office of National Statistics (2000): *Standard occupational classification 2000*. London: The Stationery Office.