Rare chromosomal deletions and duplications increase risk of schizophrenia

The International Schizophrenia Consortium

SUPPLEMENTARY INFORMATION

Outline

Supplementary Tables

1. Sample collection and ascertainment

All DNA was isolated from whole blood and no DNA samples from cell lines were used.

Aberdeen: The Scottish samples comprised a cohort of 800 schizophrenia cases and 962 controls. All participants self-identified as born in the British Isles (95% in Scotland). All cases met the Diagnostic and Statistical Manual for Mental Disorders-IV edition (DSM-IV)[1] and International Classification of Diseases 10th edition (ICD-10)[2] criteria for schizophrenia. Diagnosis was made by Operational Criteria Checklist (OPCRIT)[3]. All case participants were outpatients or stable in-patients. Detailed medical and psychiatric histories were collected. A clinical interview using the Structured Clinical Interview for DSM-IV (SCID)[4] was also performed on 723 schizophrenia cases. Controls were volunteers recruited through general practices in Scotland. Practice lists were screened for potentially suitable volunteers by age and sex and by exclusion of subjects with major mental illness or use of neuroleptic medication. Volunteers who replied to a written invitation were interviewed using a short questionnaire to exclude major mental illness in individual themselves and first degree relatives. All cases and controls gave informed consent. The study was approved by both local and multiregional academic ethical committees.

Cardiff: All cases and controls were born in Bulgaria. They were recruited and interviewed by a team of ~50 psychiatrists in a project organized by G.K. and M. O. The cases were either inpatients from five different psychiatric hospitals, or outpatients from four of the largest psychiatric dispensaries in Bulgaria. All had a history of hospitalization for a schizophrenic episode. Each proband was interviewed with an abbreviated version of the Schedules for Clinical Assessment in Neuropsychiatry (SCAN)[5]. The SCAN has been translated and validated for use in Bulgarian language by one if its authors (A. J.). Consensus best-estimate diagnoses were made according to DSM-IV[1] by two raters (G. K. and I. N.) using information from the interview and hospital discharge summaries, which were available in each case. All patients included in the study met DSM-IV[1] criteria for SCZ. Local ethics committee approval was obtained from all the regions where patients were recruited. All patients were given information sheets and provided written informed consent for participation in genetic association studies. Cases were excluded if IQ<70. DNA was extracted by standard phenol-chloroform method from peripheral blood. The controls were recruited in several settings in the two largest cities in the country: random people applying for driving licenses, non-psychiatric attendees at a GP surgery, and hospital staff. No matching for age was implemented. Although no formal interview to screen for psychiatric disorders was used, the nature of the recruitment ensured that these controls were not registered as receiving treatment for psychiatric disorders.

Dublin: Ethics Committee approval was obtained from all participating hospitals and centers. Cases provided written informed consent and met the Diagnostic and Statistical Manual for Mental Disorders-IV edition (DSM-IV)[1] and International Classification of Diseases 10th edition (ICD-10)[2] criteria for schizophrenia. Diagnosis was made by Operational Criteria Checklist (OPCRIT)[3]. All cases were over 18 years of age, of Irish origin and had been screened to exclude substance-induced psychotic disorder or psychosis due to a general medical condition. Controls were ascertained, with informed consent, from the Irish GeneBank and represented blood donors who met the same ethnicity criteria as cases. Controls were not specifically screened for psychiatric illness. Individuals taking regular prescribed medication are excluded from blood donation in Ireland and donors are not financially remunerated, making it unlikely that patient or socially disadvantaged groups (which may have higher rates of SZ) were over-represented among controls.

Edinburgh: The study was approved by the Multi-Centre Research Ethics Committee for Scotland and patients gave written informed consent for the collection of DNA samples for use in genetic studies. The sample comprised Caucasian individuals contacted through the inpatient and outpatient services of hospitals in South East Scotland. A diagnosis of SCZ was based on information from an interview with the patient using the Schedule for Affective Disorders and SCZ–Life time version (SADS-L)[6] supplemented by case note review and frequently by information from medical staff, relatives and care givers. Final diagnoses, based on DSM-IV[1] criteria were reached by consensus between two trained psychiatrists. Cases were excluded if IQ<70. Ethnically matched controls from the same region were recruited through the South of Scotland Blood Transfusion Service and from hospital staff. All controls were not directly screened to exclude those with a personal or family history of psychiatric illness; however the Blood Transfusion Service does not accept blood donations from subjects taking regular medication or with a history of a major illness.

London: The University College London case-control sample consisted of unrelated cases and ancestrally matched controls. Research subjects were selected only if both parents were English, Scottish, or Welsh, with at least three grandparents having the same origins. Subjects were also included if the fourth grandparent was of another white

European origin but were excluded if one grandparent was of Jewish or non-European Union (EU) (before the enlargement of the EU in 2004) ancestry. These data were recorded in an ancestry questionnaire, with confirmation from family histories noted on medical records. U.K. National Health Service (NHS) multicenter and local research ethics committee approval was obtained and all participating subjects signed an approved consent form after reading an information sheet. Each schizophrenic research subject had received a diagnosis and assessment by NHS psychiatrists as part of routine clinical diagnosis and treatment. Those with short-term drug-induced psychoses, psychoses with either learning disability or head injury, and other symptomatic psychoses were excluded. Schizophrenic subjects were recruited on the basis of having an ICD10 diagnosis of schizophrenia (SCZD)[2] recorded in medical case-history notes after clinical interview by NHS psychiatrists. The diagnoses were confirmed by a senior psychiatrist, usually within 1 week. SADS-L[6] interview was completed for all cases and controls by a research psychiatrist. Schizophrenic subjects were then chosen on the basis of having received a diagnosis at the "probable level" of the Research Diagnostic Criteria (RDC). Patients with schizoaffective bipolar disorder or schizomania were not included.

Portuguese Island Collection: All of the subjects that provided DNA in this study were either thoroughly-screened psychiatric controls with no personal or familial history of mental illness or probands from families segregating bipolar disorder or SCZ in the Portuguese population. This population consisted of subjects living in Portugal, the Azorean and Madeiran islands, or the direct (first or second generation) Portuguese immigrant population in the United States, as described in Sklar et al. (2004)[7]. Informed consent was obtained and a comprehensive psychiatric assessment completed from each subject using the DIGS (Diagnostic Interview for Genetic Studies)[8], Kendler's Structured Interview for Schizotypy (SIS)[9], the Schedule for the Assessment of Negative Symptoms (SANS[10]), the Schedule for the Assessment of Positive Symptoms (SAPS[10]), and the Operational Criteria Checklist (OPCRIT)[3], as translated into Portuguese and previously validated. Thorough clinical narratives were written for all subjects. Best estimate DSM-IV[1] diagnoses were made by two independent blinded researchers after review of all clinical information. Cases with disparate diagnoses were reviewed by a third senior psychiatrist, blind to the case status. For this report, we chose to examine unrelated probands with a diagnosis of SCZ.

Sweden: Cases were individuals born in Sweden or another Nordic country identified via the Swedish Hospital Discharge Register (HDR; http://www.socialstyrelsen.se/en/about/epc/) with discharge diagnoses of SCZ. 92.0% of the cases had at least 2 admissions. This register contains a nearly complete national register of all individuals hospitalized in Sweden since 1973. Each record contains the admission and discharge dates, the main discharge diagnosis, and up to eight secondary diagnoses in ICD-8 [11], ICD-9 [12], and ICD-10 codes [13]. Diagnoses were established by the attending physician. The sample is population-based covering all hospital-treated patients within three Swedish counties (Uppsala, Gästrikland and Västmanland). After ethical approval from the IRB at the Karolinska Institutet and permission from the health board to which the potential subject was registered, patients were contacted directly via an introductory letter followed by a telephone call. If they agreed, a research nurse met them at a psychiatric treatment facility or in their home, obtained written informed consent, obtained a blood sample and interviewed about other medical conditions in a lifetime perspective. For the first 121 consecutive cases, a medical record review using a structured DSM-IV[1] checklist for SCZ was conducted (C.H.) from computerized records containing notations from psychiatrists, psychologists, social workers, and nurses for inpatient and outpatient treatment. Electronic medical records were available for 111/121 subjects (obtaining hardcopy records for 10 subjects is on-going). Medical record review substantiated the presence of DSM-IV SCZ in 95.5% of subjects (=106/111). Controls were ascertained, with informed consent, frequency matched to cases by age, gender and county of residence. Controls were also identified from national population registers, and had never received a discharge diagnosis of SCZ. Controls were contacted directly in a similar procedure as the cases, gave written informed consent, were interviewed about other medical conditions and visited their family doctor or local hospital laboratory for blood donation.

2. Genotyping and CNV detection

Samples were genotyped by the Genetic Analysis Platform at The Broad Institute of Harvard and MIT according to standard protocols. Controls from UCL were previously genotyped at The Broad Institute as part of a recently published genome-wide association study of bipolar disorder[14]. The bipolar study was completed using the 500k Nsp/Sty Affymetrix Genome-Wide Human SNP Arrays, a first generation genotyping platform without specific copy number probes, and so were not included in this study.

The Affymetrix 5.0 array includes 470,000 probes for SNP genotyping and 420,000 copy number probes. The Affymetrix 6.0 array has greater probe density, with 906,600 SNP and 940,000 copy number probes[15]. Of those, 800,000 are evenly spaced along the genome, while the rest are targeted to 3,700 known CNV regions culled from a variety of sources[15].

CNVs were identified using Birdseye[16], which identifies rare CNVs by integrating intensity data from neighboring probes using a hidden Markov model (HMM) on a per-individual basis. Performance is dependent on a number of factors including SNP and copy number probe density, mean intra-individual probe variance and CNV frequency. For each CNV a LOD score was generated that describes the likelihood of the CNV relative to no CNV over the given interval.

3. Quality control evaluation

The data presented here were collected as part of a whole-genome association study of SCZ. QC was completed using SNPs to remove duplicate samples, poorly genotyped and/or contaminated samples, as part of standard quality control metrics used for whole-genome association studies[14] (data not shown).

For the 6,606 individuals passing QC (and excluding control samples from the UCL site, for whom no CNV data were available), we observed 1,493,335 unfiltered regions of copy number other than 2. In order to obtain a high-quality CNV dataset, we restricted analysis to the 56,838 with a LOD > 10 and physical length greater than 100kb. In addition, we removed 34 individuals who were outliers with respect to the number or total kb span of CNVs (more than 30 events, or total events spanning more than 10Mb), leaving 6,572 individuals (3,391 cases, 3,181 controls; 3,721 males, 2,851 females) and 52,328 events. These 34 outlier samples each had on average over 100 events greater than 100kb.

We imposed a ~1% frequency threshold, by removing any CNV with greater than 50% of its length spanning a region with more than 65 CNVs in the total sample, which left 7,365 CNVs. We also removed events spanning regions of common CNV (>3%) in the CEU HapMap individuals identified using Affymetrix array data generated in the same laboratory and called with the same analytic pipeline. As a final step, we joined any CNVs that appeared to be artificially split by the HMM and also removed any CNVs that spanned known large gaps in hg17 (greater than 200kb) or regions of known rearrangement (hg17: chr2:88695164..95087413; chr14:104644530..106268819; chr22:20797557..21512883).

The final CNV list comprised 6,753 events in 6,572 individuals. The mean event length was 301kb. There were 2,652 deletions (mean length 284.0kb, median 166.3kb) and 4,101 duplications (mean 312.0kb, median 194.4kb). That duplications are longer on average than deletions may reflect underlying biology and/or differential detection sensitivity with respect to event size and type.

Comparing 5.0 and 6.0 arrays, differences are to be expected given the different coverage of SNP and copy number probes. Between the 2,899 individuals with 5.0 array data and 3,673 with 6.0 array data, we observed differences in the number of deletions (0.354 versus 0.442 events per person; P=6.7×10^{-8}) but not mean size (P=0.42). For duplications, we observed differences in both the rate (0.534 versus 0.695; P=2.4×10^{-10}) and mean size (125.9kb versus 149.8kb; P=0.001). These results underscore the importance of controlling for array type (as well as site, plate and other potential confounding effects), as we did in subsequent analyses with respect to disease status. Table S1 shows the frequency distribution of CNV count per individual, stratified by array type and disease status. For the 5.0 array, the mean per person total CNV extent for cases and controls were 301.0kb and 223.2kb respectively; for the 6.0 array, these figures were 359.0kb and 323.6kb. The full set of 6,753 QC passing CNVs is available from http://pngu.mgh.harvard.edu/isc/ which can be loaded as a UCSC Genome Browser custom track (BED file format).

Table S1: *Distribution of CNVs per individual by array type and phenotype*

CNV per person	*Cases*	*Controls*	*Total*
Affymetrix 5.0			
0	755 (0.42)	519 (0.48)	1274 (0.44)
1	648 (0.36)	369 (0.34)	1017 (0.35)
2	275 (0.15)	144 (0.13)	419 (0.14)
3	94 (0.05)	42 (0.04)	136 (0.05)
4	23 (0.01)	12 (0.01)	35 (0.01)
5	3 (0.00)	2 (0.00)	5 (0.00)
>5	12 (0.01)	1 (0.00)	13 (0.00)
Affymetrix 6.0			
0	552 (0.35)	725 (0.35)	1277 (0.35)
1	542 (0.34)	754 (0.36)	1296 (0.35)
2	284 (0.18)	396 (0.19)	680 (0.19)
3	126 (0.08)	151 (0.07)	277 (0.08)
4	50 (0.03)	46 (0.02)	96 (0.03)
5	11 (0.01)	12 (0.01)	23 (0.01)
>5	16 (0.01)	8 (0.00)	24 (0.01)

Counts (and proportions calculated within array type and phenotypic class) of all CNVs passing QC. Considering separately individuals genotyped on Affymetrix 5.0 and Affymetrix 6.0 arrays, in both groups cases have a significantly higher rate of CNVs compared to controls (P<0.05).

4. Generation of auxiliary CNV datasets

For specific analyses considering the global CNV burden with respect to CNV frequency, type and proximity to a gene, we created the following auxiliary datasets:

- **Single-occurrences.** CNVs which were only observed once in our data, in either a case or control. These were conservatively defined as having no overlap with any other CNVs.

- **2 to 6 occurrences.** CNVs which had greater than 50% of their length spanning any one consecutive region containing 6 or fewer CNVs in the total sample; single-occurrence CNVs were then removed, as above.

These auxiliary datasets were generated for deletions and duplications combined. Additionally, we generated deletion-only and duplication-only versions. In these type-specific datasets, a deletion, for example, was still considered a single-occurrence even if it was spanned by a duplication. As such, the number of single-occurrence deletions and single-occurrence duplications does not sum to the total number of single-occurrence CNVs. There were 890 single-occurrence CNVs (470 deletions and 734 duplications); in the 2 to 6 frequency range, there were 2,465 CNVs (994 deletions and 1,532 duplications).

- **Genic CNVs.** CNVs that at least partially overlapped at least one gene including a 20kb region upstream and downstream, based on UCSC hg17 genomic coordinates.
- **Non-genic CNVs.** Any CNV that did not meet the criteria of being a genic CNV.

The major CNV lists described above were each partitioned into genic and non-genic lists. Results for all categories are given in Table S2, below. We also considered two alternative definitions for a gene being "involved" in a given CNV: first, that a CNV "disrupts" a gene (that is, *only partially* deletes or duplicates it) and second, that a CNV completely deletes or duplicates it. The default definition ("intersection") covers both these scenarios. These alternate definitions did not significantly alter the pattern of results (data not shown).

5. CNV Burden Analysis

The basic CNV burden analyses were conducted using a permutation procedure to assess statistical significance for a series of 1-sided tests (hypothesizing that cases will show greater burden of rare CNVs than controls):

- Excess of average number of CNVs per case individual compared to the average number per control individual, referred to as *CNV burden (number)*.
- Excess of average number of genes intersected by CNVs per case compared to average per control individual, referred to as *CNV burden (gene-count).*

Cases and controls were permuted only within two groups defined by their array type (Affymetrix 5.0 versus Affymetrix 6.0), thereby controlling for differences in the CNV rate between the two arrays. This is particularly important, as the ratio of cases to controls also differs between array types. As described below, we also employed a series of control procedures and analytic methods to verify the primary result of increased CNV burden in SCZ cases compared to controls. All tests reported in the main text used 1 million permutations to derive empirical *P*-values.

Table S2 (below) shows the results of the burden analyses stratifying by genic and non-genic CNVs. Note that these analyses are related but distinct from the "gene-count" burden analysis.

- The gene-count analyses (as presented in Tables 2 & 3 of the main text) compare the actual number of genes that are intersected by CNVs between cases and controls.

- In contrast, the analyses presented in Table S2 (and referenced in the main text) compare the number of CNVs between cases and controls, but limit the analysis either to the set of CNVs that intersect at least one gene ("genic CNVs") and those that do not ("non-genic CNVs").

Table S2: *Global CNV burden analysis: comparison of genic and non-genic CNVs*

				CNV burden (number)		
CNV Type	***Frequency***	***Genic/ Non-genic***	***CNV (N)***	***P***	***Case/control ratio***	***Baseline rate (controls)***
Deletions & duplications	All events	Genic	4377	5×10^{-6}	1.184	0.628
		Non-genic	2376	0.16	1.098	0.362
	Single occurrences	Genic	629	6×10^{-4}	1.384	0.083
		Non-genic	261	6×10^{-4}	1.612	0.031
	2 - 6 occurrences	Genic	1617	7×10^{-4}	1.201	0.226
		Non-genic	848	0.19	1.102	0.127
Deletions only	All events	Genic	1406	0.0051	1.155	0.202
		Non-genic	1246	0.83	1.006	0.200
	Single occurrences	Genic	283	0.024	1.299	0.038
		Non-genic	187	0.13	1.267	0.026
	2 - 6 occurrences	Genic	455	0.02	1.242	0.063
		Non-genic	539	0.37	1.084	0.083
Duplications only	All events	Genic	2971	1×10^{-4}	1.200	0.427
		Non-genic	1130	0.0140	1.211	0.162
	Single occurrences	Genic	547	4×10^{-4}	1.480	0.070
		Non-genic	187	4×10^{-4}	1.920	0.020
	2 - 6 occurrences	Genic	1171	0.026	1.152	0.169
		Non-genic	361	0.083	1.201	0.052

We explored different classes of CNVs by type and frequency to investigate the presence of an increased burden in genic and non-genic intervals. P-values were estimated by permutation, controlling for array type. Note that the number of deletions and duplications will not sum to the number of "Deletions & duplications" in the frequency-filtered datasets – see Supplementary Information for explanation.

6. Technical Controls for CNV Burden Analysis

We performed the following analyses to investigate possible sources of bias and confounding in the global CNV burden analysis.

Use of additional experimental covariates

The original CNV burden analysis controlled for between-array effects (by permuting cases and controls within array type). Alternatively, we controlled for either sample collection site (7 sites, dropping UCL) or 96-well genotyping plate membership (88 plates, the majority of which contain both cases and controls from a single site) and still observed highly significant empirical significance values for the primary CNV burden test (Table S3).

Table S3: *Global CNV burden analysis: different covariates and measures of CNV burden.*

Covariate	***N (individuals)***	***CNV burden (number)***	***CNV burden (gene-count)***	***Total CNV size (kb)***	***Mean CNV size (kb)***
Array type (original analysis)	6572	0.000027	0.000002	0.000109	0.010784
Genotyping plate	6572	0.000056	0.000001	0.000921	0.037904
Sample collection site	6025	0.000145	0.000001	0.000819	0.025344

Based on 1 million permutations in each case, controlling for different potential confounding variables by permutation (permuting cases and controls only with each covariate class when calculating the empirical significance of the test statistic).

In addition, we repeated the primary global CNV burden analysis using logistic regression instead of the permutation procedure; similar results were obtained, but this facilitated the inclusion of continuously-distributed covariates such as the intra-individual probe variance for SNPs (V_S) and copy number probes (V_C). The results are shown below in Table S4. To summarise the Table: we show that the global CNV burden result holds (for both number and gene-count metrics):

- controlling for array type, both for all events and for single-occurrence CNVs
- restricting analysis only to individuals with at least 1 CNV genome-wide
- restricting analysis only to individuals with less than 2Mb of total CNV burden
- controlling for sex effects (additionally demonstrating there is no interaction between sex and CNV burden)
- including the intra-individual QC metrics of SNP and probe variance (V_S and V_C) in the model
- restricting analysis to a 90% subset of the sample based on these metrics (probe variance analyses described below in more detail).

Table S4: *Global CNV burden analysis: controlling for further potential confounders.*

Dataset	*N (individuals)*	*Covariate*	*Test variable*	*P, 1-sided, asymptotic*
Full sample	6572	Array type	CNV burden (number)	0.0000357
Full sample	6572	Array type	CNV burden (gene count)	0.0000006
Single-occurrence	6572	Array type	CNV burden (number)	0.0000092
Single-occurrence	6572	Array type	CNV burden (gene count)	0.0112000
Individuals with 1+ events	4021	Array type	CNV burden (number)	0.0001230
Individuals with 1+ events	4021	Array type	CNV burden (gene count)	0.0000033
Individuals < 2Mb total CNV	6477	Array type	CNV burden (number)	0.0027000
Individuals < 2Mb total CNV	6477	Array type	CNV burden (gene count)	0.0000064
Full sample	6572	Array type & sex	CNV burden (number)	0.0000217
Full sample	6572	Array type & sex	CNV burden (gene count)	0.0000005
Full sample	6572	Array type & sex	Interaction of CNV burden (number) by sex	0.3513750
Full sample	6572	Array type & sex	Interaction of CNV burden (gene count) by sex	0.0955000
Full sample	6572	Probe variance (V_S, V_C) & array type	CNV burden (number)	0.0000450
Full sample	6572	Probe variance (V_S, V_C) & array type	CNV burden (gene count)	0.0000010
90% probe variance subset	5883	Array type	CNV burden (number)	0.0004600
90% probe variance subset	5883	Array type	CNV burden (gene count)	0.0000288

Sample collection site sensitivity analysis

We investigated whether the primary CNV burden resulted from a single collection site, which might be suggestive of possible bias and confounding. Repeating the primary analysis, controlling for chip type, we performed a sensitivity analysis, by dropping each of the 8 sites, one at a time (Table S5). In each case, the association observed in the remaining sub-sample was highly significant, suggesting that a single site did not drive the observed association between CNV burden and phenotype.

Table S5: *Global CNV burden analysis: sample collection site sensitivity analysis.*

Excluded site	*Site N*	*CNV burden (number)*	*CNV burden (gene count)*
Aberdeen	1421	0.002	0.024
Cardiff	1125	0.001	0.002
Dublin	1194	0.002	0.001
Edinburgh	693	0.001	0.041
Portugal	533	0.001	0.015
Sweden1	398	0.001	0.001
Sweden2	661	0.001	0.003
UCL	547	0.001	0.004

Values are empirical P-values controlling for array type, calculated in the whole sample after the removal of one site.

Analysis of a restricted sub-sample based on intra-individual prove variance

We explored the possibility that subtle differences in DNA quality or experimental protocol could be responsible for our observation of more CNVs in cases than controls (the primary CNV burden analysis). We used the intra-individual variance in probe intensity for autosomal SNP and CN probes (denoted V_S and V_C) as surrogate measures of overall DNA sample quality and/or experimental performance.

Given that all samples in the present study already passed stringent QC procedures for the SNP component of this study, we expected most gross sample failures to have already been excluded. Nonetheless, we repeated the original CNV burden analysis after removing ~10% of the sample with the highest variance measures (threshold based on visual inspection of probe variance distribution). This approach is conservative: in contrast to removing only extreme outliers, here we aimed to create a sub-sample of individuals that was more homogeneous with respect to these metrics. As shown in Table S6, in the total sample (N=6,572), V_S and V_C were in fact moderately associated with disease state, with cases having, on average, slightly higher variances (rows 1&2). The metrics V_S and V_C were also moderately, positively associated with the total kilobase span of CNVs per individual (rows 3&4) although not the number of CNVs (rows 5&6). Nonetheless, the association between CNV burden and disease state remains when controlling for both V_S and V_C (rows 7&8), suggesting that the global burden analyses are not in fact being driven by bias in sample quality as indexed by the two probe variance measures.

To further illustrate the robustness of the CNV burden-disease association, Table S6 also shows results for the same analyses but restricted to the ~90% of the sample with lower probe variance scores. In this case, the association between probe variance and disease is no longer observed (rows 1&2); nor is the association of probe variance with CNV burden (rows 3-6). However, the CNV-disease association remains (P=0.00086 and P=0.00075 for tests of total kb burden and number of events, respectively) in this subset of the sample (N=5,883), controlling for V_S and V_C.

Table S6: *Global CNV burden analysis: controlling for intra-individual probe variance.*

Dependent variable	*Covariates*	*Predictor*	*Entire sample N=6,572 (P-value)*	*Low probe variance N=5,883 (P-value)*
Disease status	Array type, sex	V_S	0.07300	0.21800
Disease status	Array type, sex	V_C	0.00200	0.29600
CNV burden (total kb)	Array type, sex	V_S	0.09500	0.57700
CNV burden (total kb)	Array type, sex	V_C	0.02140	0.19300
CNV burden (number)	Array type, sex	V_S	0.23850	0.53400
CNV burden (number)	Array type, sex	V_C	0.10770	0.81400
Disease status	Array type, sex, V_S, V_C	CNV burden (total kb)	0.00003	0.00086
Disease status	Array type, sex, V_S, V_C	CNV burden (number)	0.00005	0.00075

Asymptotic, 2-sided P-values from logistic and linear regression analysis. V_S is the individual SNP probe variance, V_C is individual copy number probe variance. Also includes covariates of array type (5.0 or 6.0) and sex.

Genic CNV enrichment after matching for overall CNV burden

Our final approach to establish the robustness of the association between CNV burden and SCZ focused on the observation that the burden appeared to be greatest for genic CNVs. Given this, we matched cases and controls based on their general level of CNV burden and then asked whether CNV burden in cases and controls was differentially distributed with respect to genic regions versus non-genic regions. We postulated that, whereas general CNV burden can be influenced by confounding factors not controlled for, it is harder to imagine mechanisms by which confounding factors might, on a genome-wide scale, differentially influence whether or not CNVs tend to fall in genic regions.

We controlled for confounding factors in two ways, with similar results. First, we matched cases and controls based on their genome-wide number of CNVs and then, using permutation within these matched groups, assessed whether the count of intersected genes differs between cases and controls. The primary burden analysis (CNV number) is necessarily correlated with burden as defined by gene-count, since the greater number of events an individual has, the higher the gene-count is likely to be by chance. In contrast, this more conservative analysis focused specifically on the rate of genic CNVs in cases and controls, independent of any differences in the number of CNVs genome-wide. After matching cases and controls for the number of CNVs genome-wide, the empirical significance for the gene-count test was *P*=0.0003.

We also adopted a second procedure that additionally controlled for any case/control difference in total kilobase distance covered by CNVs (rather than the number). Entering both gene-count and total CNV length as independent variables in a logistic regression, we observed that the disease-association with gene-count remains significant (2 sided, asymptotic *P* = 0.0071) whereas the effect of total CNV kb is not (*P*=0.53).

A similar pattern is obtained (that the gene-count term remains significant) when entering other covariates in this framework:

- the number of events (gene-count *P*=0.0011)

- the number, average size and total kb span of all events (gene-count *P*=0.0061)
- all the above, array type and the two probe variance metrics (gene-count *P*=0.0064).

In contrast, and as a proof-of-principle experiment, we took array type (5.0 versus 6.0) as the dependent variable instead of disease status (i.e. an "outcome" for which we know there are differences in CNV burden solely due to technical bias), and applied this same procedure. As expected, we observed an association between array type and gene-count alone (P=0.00231), although controlling for total kilobase span removes this effect completely (P=0.895). As true disease state is correlated with array type, we also performed this same proof-of-principle experiment additionally controlling for disease state, with the same results (unadjusted gene-count P=8×10^{-5}, but adjusted P = 0.64).

In conclusion, these statistical control procedures demonstrate an association between genic CNV burden and SCZ, controlling for differences in overall CNV burden between cases and controls.

7. Mapping Specific CNV Loci

To identify specific loci harboring CNVs associated with SCZ, we considered every position of the genome with a distinct set of CNVs overlapping that particular position. Given the observed number of case and control CNVs at that location, and the total case and control sample N, we calculated a standard chi-squared test for independence, but evaluated significance via permutation. As well as allowing for us to control for important covariates such as array type, genotyping plate or sample collection site, and providing a test robust to very small cell sizes, using permutation also provides a natural way of correcting for genome-wide multiple testing, by comparing each observed statistic to the maximum across the genome per replicate when calculating the corrected empirical *P*-values. All tests were 1-sided, given the model of rare, moderately or highly penetrant variants impacting risk for disease. As shown in Table S10, our results for the three loci identified (22q11.2, 15q13.3 and 1q21.1) were invariant to controlling for genotyping plate or sample collection site instead of array type. Given the very small cell sizes (e.g. for a 10:0 event), approximate odds ratios were calculated following Gart's (1966)[17] practice of replacing n_{ij} with $\{ n_{ij} + 0.5 \}$ in the estimator of the odds ratio. The genome-wide mapping analyses were performed after removing the 13 individuals with 22q11.2 deletions.

For the three loci identified, Table S7 provides additional information on the deletions as well as phenotypic details for the individuals carrying them. In addition to the large deletions reported in this Table, we observed duplications >500kb at these three loci: at

22q11.2, three controls; at 15q13.3, ten cases and nine controls; at 1q21.1 four cases and two controls.

Table S8 provides an annotated list of the genes in these three regions. These data were compiled using SLEP (Sullivan Lab Evidence Project) (URL: https://slep.unc.edu/evidence/).

Table S9 provides a comparison between the CNVs observed by Walsh et al[18] and our study.

Table S7: *List of individuals with deletions at 22q11.2, 15q13.3 and 1q21.1.*

Sample ID (Fig. 1)	*Sex*	*Site*	*Chip*	*Chr*	*Start (Mb)*	*Stop (Mb)*	*Size (Mb)*	*SCZ type*	*Age of Onset*	*Cognitive Deficits*	*Family History*
1	F	UCL	5.0	22	17.26	19.92	2.66	U	20	None	No
2	M	UCL	5.0	22	17.26	19.79	2.53	U	22	7-9 years of schooling	No
3	F	Sw	6.0	22	17.25	18.69	1.44	D	22[a]		Brother: schizophrenia
4	F	Sw	6.0	22	17.11	18.69	1.58	D	29[a]		
5	M	Port	5.0	22	17.24	18.68	1.44	U	17		No
6	M	Dub	6.0	22	17.25	18.77	1.52		18	None	No
7	F	Card	6.0	22	19.04	19.79	0.75	P	35	Poor school results	No
8	M	Ab	5.0	22	17.26	19.79	2.53	P	21		
9	F	Ab	5.0	22	17.26	19.79	2.53		21	None	Father: “nervous illness”
10	M	Ab	5.0	22	17.24	19.79	2.55	P	24	Learning Disability	Mother: depressive psychosis
11	M	Ab	5.0	22	17.26	18.68	1.43				
12	F	Ab	5.0	22	19.03	19.79	0.76				
13	M	Ab	5.0	22	17.24	19.92	2.68		21	None	Grandmother: depression
14	M	UCL	5.0	15	28.72	30.49	1.77	P	38	Less than 7 years schooling	No
15	M	UCL	5.0	15	29.81	30.34	0.53	P	17	7-9 years of schooling	No
16[b]	M	Sw	6.0	15	28.68	30.23	1.56	U	28[a]	Mild MR	
17	M	Sw	6.0	15	28.68	30.23	1.56	D/ U	19[a]	Mild MR	No
18	M	Port	5.0	15	28.72	30.57	1.85	U	20	None	No
19	F	Dub	6.0	15	28.17	30.65	2.47		25	None	No
20	F	Card	6.0	15	28.71	30.25	1.54	P	32	None	No
21	M	Card	6.0	15	28.71	30.25	1.54	P	27	Poor school results	No
22	M	Ab	5.0	15	28.69	30.57	1.89		20	IQ=83	
23	M	UCL	5.0	1	143.52	145.04	1.53	P	28	IQ=67	Mother: SCZ, Sister: learning disability, possible SCZ, Cousin: hypomania
24	M	UCL	5.0	1	143.51	145.04	1.53	P	21	IQ=90 Left school w/o qualification	Mother: multiple sclerosis parents: unspecified mental illness and "possible" mental illness in siblings
25	F	Sw	6.0	1	143.40	144.95	1.56	P	52[a]	None	
26	F	Port	5.0	1	143.71	145.13	1.42	U	15	None	No
27	M	Port	5.0	1	143.72	144.58	0.86	U	31	None	No
28	F	Ab	5.0	1	143.51	145.12	1.60			IQ=79	
29	M	Ab	5.0	1	143.72	145.04	1.32		27		No
30	M	Ab	5.0	1	143.72	144.95	1.23		15	None	Adopted
31	F	Ab	5.0	1	142.54	145.35	2.81		33	Borderline learning disability	None
32[b]	M	Ab	5.0	1	142.74	145.02	2.27		32	None	None
33*	M	Ed	6.0	1	143.57	144.95	1.38	NA	NA		

Ab: Aberdeen; Card: Cardiff; Dub: Dublin; Ed: Edinburgh; Port: Portuguese; Sw: Swedish; UCL: University College London. D: Disorganized Type; P: Paranoid Type; U: Undifferentiated Type.

[a] Age at first hospitalization; [b] Patient also had report of epilepsy. All positions are based on hg17. *Control sample. A blank cell indicates that the data were not available.

Table S8: *Genes within deletion regions on 22q11.2, 15q13.3 and 1q21.1.*

Chr	*Start (kb)*	*Stop (kb)*	*Gene Name*	*Gene Product*	*Gene Aliases*	*Novartis_ ExpBrainP75*	*SCZ_ studies*	*Notes and References*
22	17014.98	17034.14	*USP18*	ubiquitin specific protease 18		No	--	
22	17268.44	17273.76	*DGCR6*	DiGeorge syndrome critical region protein 6		Yes	2	
22	17275.24	17298.35	*PRODH*	proline dehydrogenase (oxidase) 1	HSPOX2, PRODH1, PIG6, PRODH2, TP53I6	Yes	10	Hyperprolinemia, type I, 239500 (3) [OMIM=606810]/[19] 600850 (3) [OMIM=606810]
22	17493.47	17494.54	*TSSK2*	spermiogenesis associated serine/threonine	SPOGA2, FLJ38613	No	--	
22	17496.26	17506.71	*DGCR14*	DiGeorge syndrome critical region protein 14	DGSI, Es2el, ES2, DGS-H	Yes	4	
22	17511.06	17512.35	*GSCL*	goosecoid-like		No	2	
22	17538.20	17540.74	*SLC25A1*	solute carrier family 25 (mitochondrial carrier; citrate transporter), member 1	CTP	Yes	1	
22	17541.54	17653.79	*CLTCL1*	clathrin, heavy polypeptide-like 1		Yes	1	
22	17693.52	17793.55	*HIRA*	HIR histone cell cycle regulation defective homolog A (S. cerevisiae)		Yes	2	
22	17794.63	17798.04	*MRPL40*	mitochondrial ribosomal protein L40	MRP-L22	Yes	--	
22	17812.75	17841.16	*UFD1L*	ubiquitin fusion degradation 1-like isoform A	UFD1	Yes	4	
22	17842.05	17880.99	*CDC45L*	CDC45-like		No	2	
22	17885.68	17886.33	*CLDN5*	claudin 5	CPETRL1, BEC1	Yes	5	
22	18076.67	18084.56	*SEPT5*	septin 5	HCDCREL-1, H5	Yes	--	
22	18085.62	18086.85	*GP1BB*	glycoprotein Ib, beta polypeptide precursor	CD42c	Yes	--	Bernard-Soulier syndrome, type B, 231200 (3) [OMIM=138720]/Giant platelet disorder, isolated (3) [OMIM=138720]
22	18121.72	18128.94	*TBX1*	T-box 1 isoform C		No	4	Conotruncal anomaly face syndrome, 217095 (3) [OMIM=602054]/DiGeorge syndrome, 188400 (3) [OMIM=602054]/Velocardiofacial syndrome, 192430 (3) [OMIM=602054]
22	18150.79	18183.43	*GNB1L*	guanine nucleotide binding protein	GY2	no data	2	? Schizophrenia [PMID=18003636]
22	18213.24	18214.34	*C22orf29*	hypothetical protein LOC79680	FLJ21125	No	--	
22	18237.59	18303.91	*TXNRD2*	thioredoxin reductase 2 precursor	TR, TRXR2, TR3	Yes	1	
22	18324.60	18330.81	*COMT*	catechol-O-methyltransferase isoform MB-COMT		Yes	73	? Neuropsychiatric disorders, the "usual suspects" [PMID=N/A]/? Schizophrenia [PMID=16033310]/evidence of monoallelic expression

								[PMID=18006746]/{ Panic disorder, susceptibility to}, 167870 (3) [OMIM=116790]/[19], 181500 (3) [OMIM=116790]
22	18333.30	18352.87	*ARVCF*	armadillo repeat protein		Yes	7	
22	18395.22	18395.30	*hsa-mir-185*	RNA object miRna hsa-mir-185		no data	--	
22	18398.88	18426.74	*C22orf25*	hypothetical protein LOC128989	DKFZp761P1121	no data	--	
22	18448.04	18472.19	*DGCR8*	DiGeorge syndrome critical region gene 8	DGCRK6, Gy1	Yes	2	
22	18474.64	18478.98	*HTF9C*	HpaII tiny fragments locus 9C		Yes	1	
22	18479.73	18489.13	*RANBP1*	RAN binding protein 1		Yes	1	
22	18494.02	18507.48	*ZDHHC8*	zinc finger, DHHC domain containing 8	ZNF378, KIAA1292	Yes	8	
22	18603.79	18630.17	*RTN4R*	reticulon 4 receptor precursor	NOGOR	no data	4	[19], 181500 (3) [OMIM=605566]
22	18676.75	18682.07	*DGCR6L*	DiGeorge syndrome critical region gene 6 like	FLJ10666	No	1	
22	19073.03	19087.31	*ZNF74*	zinc finger protein 74 (Cos52)	Cos52, Zfp520, ZNF520	No	3	evidence of monoallelic expression [PMID=18006746]
22	19104.22	19116.60	*SCARF2*	scavenger receptor class F, member 2 isoform 1	SREC-II, SREC2, HUMZD58C02	No	--	
22	19120.91	19168.05	*KLHL22*	kelch-like	FLJ14360, KELCHL	Yes	--	
22	19186.52	19265.55	*PCQAP*	positive cofactor 2, glutamine/Q-rich-associated		Yes	3	
22	19380.03	19380.51	*DKFZp434N035*	Hypothetical protein DKFZp434N035 (Em:AC007050.6 protein).		no data	--	
22	19386.53	19413.51	*PIK4CA*	phosphatidylinositol 4-kinase, catalytic, alpha		Yes	4	
22	19458.15	19465.91	*SERPIND1*	heparin cofactor II precursor	HC-II, HLS2, HC2, D22S673	No	--	Thrombophilia due to heparin cofactor II deficiency (3) [OMIM=142360]
22	19537.95	19566.68	*SNAP29*	synaptosomal-associated protein 29		No	3	Cerebral dysgenesis, neuropathy, ichthyosis, and palmoplantar keratoderma syndrome, 609528 (3) [OMIM=604202]
22	19596.78	19628.69	*CRKL*	v-crk sarcoma virus CT10 oncogene homolog		Yes	--	
22	19646.79	19659.87	*AIFM3*	apoptosis-inducing factor like isoform 1	AIFL, FLJ30473	Yes	--	
22	19678.72	19680.75	*THAP7*	THAP domain containing 7 isoform a	MGC10963	Yes	--	
22	19694.05	19705.46	*P2RXL1*	purinergic receptor P2X-like 1, orphan receptor	P2XM	Yes	--	
22	20121.67	20125.59	*HIC2*	hypermethylated in cancer 2	KIAA1020, HRG22, ZBTB30	Yes	--	
22	20246.59	20300.51	*UBE2L3*	ubiquitin-conjugating enzyme E2L 3 isoform 1	UBCH7	Yes	--	
15	28440.73	28473.16	*CHRFAM7A*	CHRNA7-FAM7A fusion	D-10, CHRNA7-DR1	no data	7	

15	29018.44	29071.10	*MTMR10*	myotubularin related protein 10	FLJ20313	No	--	
15	29080.84	29181.22	*TRPM1*	transient receptor potential cation channel,	LTRPC1	No	--	
15	29144.53	29144.64	*hsa-mir-211*	RNA object miRna hsa-mir-211		no data	--	
15	29406.71	29451.79	*KLF13*	Kruppel-like factor 13	RFLAT-1, BTEB3, NSLP1, FKLF-2	No	--	
15	29562.79	29734.74	*OTUD7A*	OTU domain containing 7A	CEZANNE2	No	--	
15	30110.09	30247.95	*CHRNA7*	cholinergic receptor, nicotinic, alpha 7		No	11	? Schizophrenia [PMID=16843094]/ Schizophrenia, neurophysiologic defect in (2) [OMIM=118511]
1	142585.45	142605.97	*SEC22B*	SEC22 vesicle trafficking protein homolog B	sec22b, ERS-24	Yes	--	
1	142737.90	142771.08	*NOTCH2NL*	Notch homolog 2 N-terminal like protein	N2N	Yes	--	
1	142905.38	142905.98	*HFE2*	hemojuvelin isoform c	JH, HFE2A, RGMC, HJV, hemojuvelin	Yes	--	Hemochromatosis, juvenile, 602390 (3) [OMIM=606464]/Hemochromatosis, juvenile, digenic, 602390 (3) [OMIM=606464]
1	142927.85	142930.26	*TXNIP*	thioredoxin interacting protein	VDUP1, EST01027, HHCPA78, THIF	Yes	--	
1	142945.68	142949.27	*POLR3GL*	polymerase (RNA) III (DNA directed) polypeptide	flj32422, MGC3200	Yes	--	
1	142966.20	142987.82	*LIX1L*	Lix1 homolog (mouse) like	MGC46719	No	--	
1	142996.71	142998.26	*RBM8A*	RNA binding motif protein 8A	ZNRP, BOV-1A, BOV-1B, BOV-1C, RBM8B, Y14	Yes	--	
1	143005.44	143011.96	*PEX11B*	peroxisomal biogenesis factor 11B		Yes	--	
1	143014.11	143031.32	*ITGA10*	integrin, alpha 10 precursor		Yes	--	
1	143038.36	143056.80	*ANKRD35*	ankyrin repeat domain 35	FLJ25124	Yes	--	
1	143067.11	143074.67	*PIAS3*	protein inhibitor of activated STAT, 3	FLJ14651, ZMIZ5	Yes	--	
1	143081.73	143098.38	*POLR3C*	polymerase (RNA) III (DNA directed) polypeptide	RPC62, RPC3	Yes	--	
1	143100.28	143177.26	*ZNF364*	Rabring 7	CL469780, RNF115	Yes	--	
1	143185.64	143195.80	*CD160*	CD160 antigen	BY55, NK1, NK28	No	--	
1	143236.09	143252.67	*PDZK1*	PDZ domain containing 1	PDZD1	No	--	
1	143502.09	143537.79	*NBPF11*	Hypothetical protein LOC200030		no data	--	
1	143545.02	143549.78	*FAM108A3*	hypothetical protein LOC653401	C1orf47	no data	--	
1	143855.96	143868.54	*PRKAB2*	AMP-activated protein kinase beta 2		Yes	--	
1	143883.29	143921.44	*FMO5*	flavin containing monooxygenase 5		Yes	--	
1	143939.17	143992.01	*CHD1L*	chromodomain helicase DNA binding protein	ALC1	Yes	--	
1	144238.00	144322.82	*BCL9*	B-cell CCL/lymphoma 9		Yes	0	mutated in colorectal cancer [PMID=17932254]
1	144344.04	144366.99	*ACP6*	acid phosphatase 6, lysophosphatidic	LPAP, ACPL1	Yes	--	

1	144455.09	144456.16	*GJA5*	connexin 40	CX40	No	1	Atrial fibrillation, familial 4, 108770 (3) [OMIM=121013]/Atrial standstill, 108770 (3) [OMIM=121013]
1	144599.75	144606.21	*GJA8*	connexin 50	CX50, CAE1, CZP1, CAE	No	1	Cataract, zonular pulverulent-1, 116200 [OMIM=600897]
1	144625.47	144689.97	*GPR89A*	G protein-coupled receptor 89		no data	--	
1	144966.12	145046.72	*NM_207400*	CDNA FLJ39739 fis, clone SMINT2016440.		no data	--	
1	145357.24	145374.23	*NBPF15*	hypothetical protein LOC284565	MGC8902	no data	--	

- Novartis_ExpBrainP75, expressed in >0 human brain regions at 75th percentile in the Novartis transcriptomic experiment
- SCZ_studies: Number of SCZ association studies for this gene
- Notes/References: a mix of annotations - OMIM, curated GWAS findings, usually has PMID as citation

Table S9: *Copy number variable regions listed in Table 2 in Walsh et al.*

Affection	*Chr.*	*Start* (Mb)*	*Stop* (Mb)*	*Region Size (kb)*	*Type*	*# genes*	*Disrupted genes*	*Events in the current study*			
								CASES		*CONTROLS*	
								Del	Dup	Del	Dup
Case	1	143.70	144.94	1349.14	Del	11	NBPF10	9	3	1	1
Case	1	230.70	230.91	209.10	Dup	3	SLC35F3, TARBP1				
Case	2	48.71	49.38	670.93	Dup	3	STON1-GTF2A1L				1
Case	2	211.91	212.31	399.16	Del	1	ERBB4				1
Case	3	7.18	7.31	136.52	Del	1	GRM7				
Case	3	53.06	53.19	135.18	Dup	2	PRKCD				
Case	3	197.23	198.58	1348.55	Del	20	-	2			
Case	5	36.19	36.69	502.68	Del	4	SKP2, SLC1A3				
Case	7	77.17	77.66	498.45	Dup	2	MAGI2, PHTF2				
Case	7	100.10	115.77	15668.29	Dup	82	SLC12A9, CAV1				
Case	7	150.88	151.34	461.99	Dup	4	PRKAG2, MLL3				
Case	8	142.03	142.39	368.52	Dup	3	PTK2				
Case	9	2.01	3.12	1105.06	Dup	4	SMARCA2				
Case	9	3.10	3.54	440.09	Del	1	-				
Case	9	25.33	25.85	526.89	Dup	1	-				
Case	11	33.26	33.54	279.21	Dup	2	HIPK3, C11orf41				1
Case	11	83.69	83.94	242.11	Del	1	DLG2			1	
Case	14	53.49	53.77	282.77	Dup	1	-				
Case	18	7.07	7.57	495.02	Dup	2	LAMA1, PTPRM		1		1
Case	19	59.05	59.36	317.74	Dup	13	TMC4				
Case	22	32.04	32.71	666.71	Dup	1	LARGE				
Case	Y	0.0	57.70	57,700.00	Dup	Entire Y	-	NA	NA	NA	NA
Case	Y	57.49	57.64	152.39	Dup	1	-	NA	NA	NA	NA
Ctrl	1	99.92	100.14	219.76	Del	2	FRRS1				
Ctrl	2	8.20	8.61	410.03	Del	1	-				
Ctrl	3	79.67	81.41	1740.70	Del	1	ROBO1				
Ctrl	6	95.73	96.14	413.04	Del	1	MANEA				
Ctrl	7	84.14	84.51	363.92	Del	1	-				
Ctrl	7	110.84	112.14	1300.90	Dup	6	FLJ31818				1
Ctrl	7	126.89	127.23	338.84	Del	1	SND1				
Ctrl	8	8.14	11.76	3611.15	Del	20	CTSB	1			
Ctrl	9	12.64	13.22	575.53	Dup	3	MPDZ				
Ctrl	12	24.58	25.25	665.30	Dup	7	SOX5, LYRM5		1		
Ctrl	12	29.79	29.99	196.99	Del	1	TMTC1				
Ctrl	16	69.40	69.75	352.89	Del	2	HYDIN		2		1
Ctrl	22	30.89	31.19	292.66	Dup	5	BPIL2				

CNVs regions listed in Walsh et al.[18] Table 2 are listed. For each interval, events identified in the present study are tallied for cases and controls, deletions (Del) and duplications (Dup), as marked. In order for an event in the current study to be considered overlapping with an event presented in Walsh et al., the two CNV regions being compared were required to overlap by at least 50% of the union of their total spanned length. All coordinates are given based on hg17. Chr=chromosome. Number of genes and disrupted genes are recreated from Walsh et al. table 2 and represent the authors' data and terminology.

Table S10: *Controlling for potential confounders for CNVs at 22q11.2, 15q13.3 and 1q21.1*

			Permutation within:		
Locus	***Case CNVs (N)***	***Control CNVs (N)***	***Array type***	***Genotyping plate***	***Sample collection site***
1q21	10	1	0.0076 (0.046)	0.006 (0.034)	0.01 (0.043)
15q13	9	0	0.0029 (0.046)	0.011 (0.034)	0.002 (0.043)
22q11	11	0	0.0017 (0.005)	0.0018 (0.005)	0.002 (0.008)

Empirical significance values obtained by 100,000 permutations, within either array type, genotyping plate or sample collection site, to control for potential confounders. In each case, the pointwise and genome-wide corrected significance values are presented. This analysis looked only at >500kb deletions; the genome-wide analysis identifying 1q21 and 15q13 loci was performed after removing the 13 individuals identified with 22q11 deletions.

8. CNV Validation

Primers and probes were designed within CNV regions using Primer3 software (Amplicon size range: 100-120 bp; Primer Tm: 59-60°C; Prob Tm:68-70°C). Probes were labeled with 5'FAM, 3'BHQ1, which allowed each assay to be multiplexed with the control assay, PMP22 labeled with 5'VIC, 3'MGBNFQ. An absolute quantification real time PCR was performed for each test assay, multiplexed with the control assay, on all samples with an expected CNV event at that site and minimum of 40 controls. All samples were plated in triplicate using 3ng of DNA with a final reaction volume of 20ul. An analysis of qPCR results was performed by obtaining a ΔCT for each reaction and normalizing based on the median control ΔCT and then averaging across the triplicate experiments.

Table S11 shows the primers used; Table S12 shows the results of validation (with specific probe positions corresponding to Figure 1 in the main text)

Table S11: *qPCR primers for CNVs at 22q11.2, 15q13.3 and 1q21.1*

Ch.	***Primer Label in Figure 1***	***Right Primer Sequence (5'→3')***	***Left Primer Sequence (5'→3')***
22	A	TCCTCCAAGAGTCACCCATC	GTTGCTGTCAGGAAGCATCA
22	B	GCTGCAGGAGTAAGGACAGG	ACTGACAGGGCTAAGGAGCA
22	C	GACCGCCACCTCTATGTGTT	GGTCTGGAAGTCCACGTCAT
15	D	GAAGAACAGAGGGTGGGTGA	CTTTGGACACAGCGAGTGAA
15	E	CTGTGGATGAGCTGTCCTGA	GCTCCTTCCCTCTTCAGCTT
15	F	GGCACTGGAGTTCCCTGATA	GGGGGTTCTGTCTTGCACTA
1	G	GAGCTTTTGGTTTGCTGAGG	GACCTCTGTCCTGCTTCCTG
1	H	CTTTCCCAGACCCCACTGTA	CCTTCCAAATCTCCCAGTCA
1	I	GTGTTGTTCTCCCGTCCAGT	GGCCCTAGCCTCTTGGTATC

Table S12: *qPCR validation results for CNVs at 22q11.2, 15q13.3 and 1q21.1*

	Chromosome 22 Primers		
Sample	***A***	***B***	***C***
1	Del	Del	Del
2	Del	Del	Del
3	Del	Del	2-copies
4	Del	Del	2-copies
5	Del	Del	2-copies
6	Del	Del	2-copies
7	Dup	2-copies	Del
8	Del	Del	Del
9	Del	Del	Del
10	Del	Del	Del
11	Del	Del	2-copies
12	2-copies	2-copies	Del
13	Del	Del	Del

	Chromosome 15 Primers		
Sample	***D***	***E***	***F***
14	2-copies	Del	Del
15	2-copies	Del	Del
16	2-copies	Del	2-copies
17	2-copies	Del	Not called
18	2-copies	Del	Not called
19	Del	Del	Del
20	2-copies	Del	2-copies
21	2-copies	Del	Not called
22	2-copies	Del	Not called

	Chromosome 1 Primers		
Sample	***G***	***H***	***I***
23	2-copies	Del	Del
24	2-copies	Del	Del
25	2-copies	Del	Del
26	2-copies	Del	Del
27	2-copies	Del	Del
28	2-copies	Del	Del
29	2-copies	Del	Del
30	2-copies	Del	Del
31	Not Tested	Del	Not Tested
32	Del	Del	Del
33*	Del	Del	Del

*Control sample.

Results for qPCR validation at probes A-I (Figure 1, main text, mapping to the grey lines left to right, A-C, D-F and G-I). 2-copies: evidence of normal copy number using specified primer pair. Del=evidence of decreased copy; Dup=evidence of increased copy number. Not tested=sample was not tested for that primer pair. Not called=qPCR results were inconclusive

9. Power simulation

By simulation, we approximated the statistical power to detect a locus similar to 22q11. We assumed a population frequency for the deletion of 1/8000 (i.e. so it would be observed in 1/4000 live births), set the relative risk to 20.0 and the population disease prevalence to 1/100. We simulated 10,000 datasets for 3,391 cases and 3,181 controls under this model. Using Fisher's exact test to account for small cell sizes, for a type I error rate of 0.01 (1-sided test) we had 97.7% power. The mean case frequency was ~0.5%, the mean control frequency was ~0.02%.

Type I error rate	Power
0.05	100%
0.01	98%
0.001	87%
0.0001	63%

For a similarly rare variant but with a relative risk of 10.0, the average case frequency was ~0.25% (control frequency still ~0.02%) and power was lower:

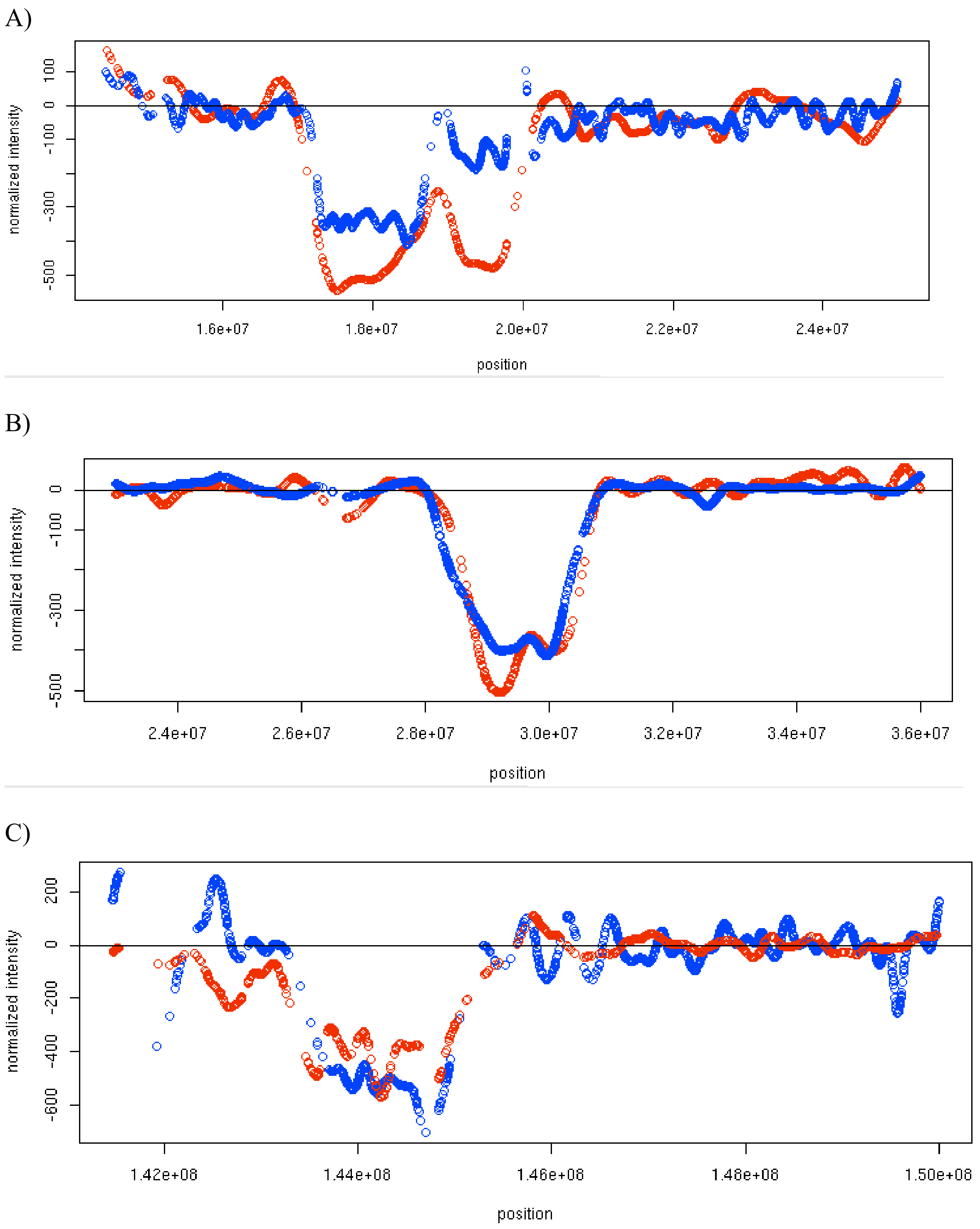
Type I error rate	Power
0.05	78%
0.01	50%
0.001	14%
0.0001	3%

Given that we required a *P*-value of ~0.01 to withstand correction for multiple testing in our primary scan, this suggests we had good power to detect loci with effects as large as the 22q11 deletion, although this assumes perfect sensitivity and specificity for detection. We observed 100% specificity when experimentally validating all 33 large deletions implicated at the three loci in this study; also, for such large deletions, we might expect sensitivity to (at least partially) detect such a CNVs will also be quite high. However, we certainly could have missed additional loci with slightly less penetrant or rarer variants, or with lower sensitivity and specificity (e.g. smaller variants). Additionally, heterogeneity introduced by the different array types and sites could also reduce power.

10. Supplementary Figures and Legends

Supplementary Figure Legend 1: Normalized average difference of probe intensity of samples with deletion events from all other samples. A) Chromosome 22: The normalized average difference in intensity of individuals carrying large deletion events versus all other samples is plotted as a function of base-pair position. Blue = 6.0 data, red = 5.0 data. B) Chromosome 15: same as A. C) Chromosome 1: same as A.

Supplemental Figure 1: Normalized average difference of intensity between samples with deletion events and those without.



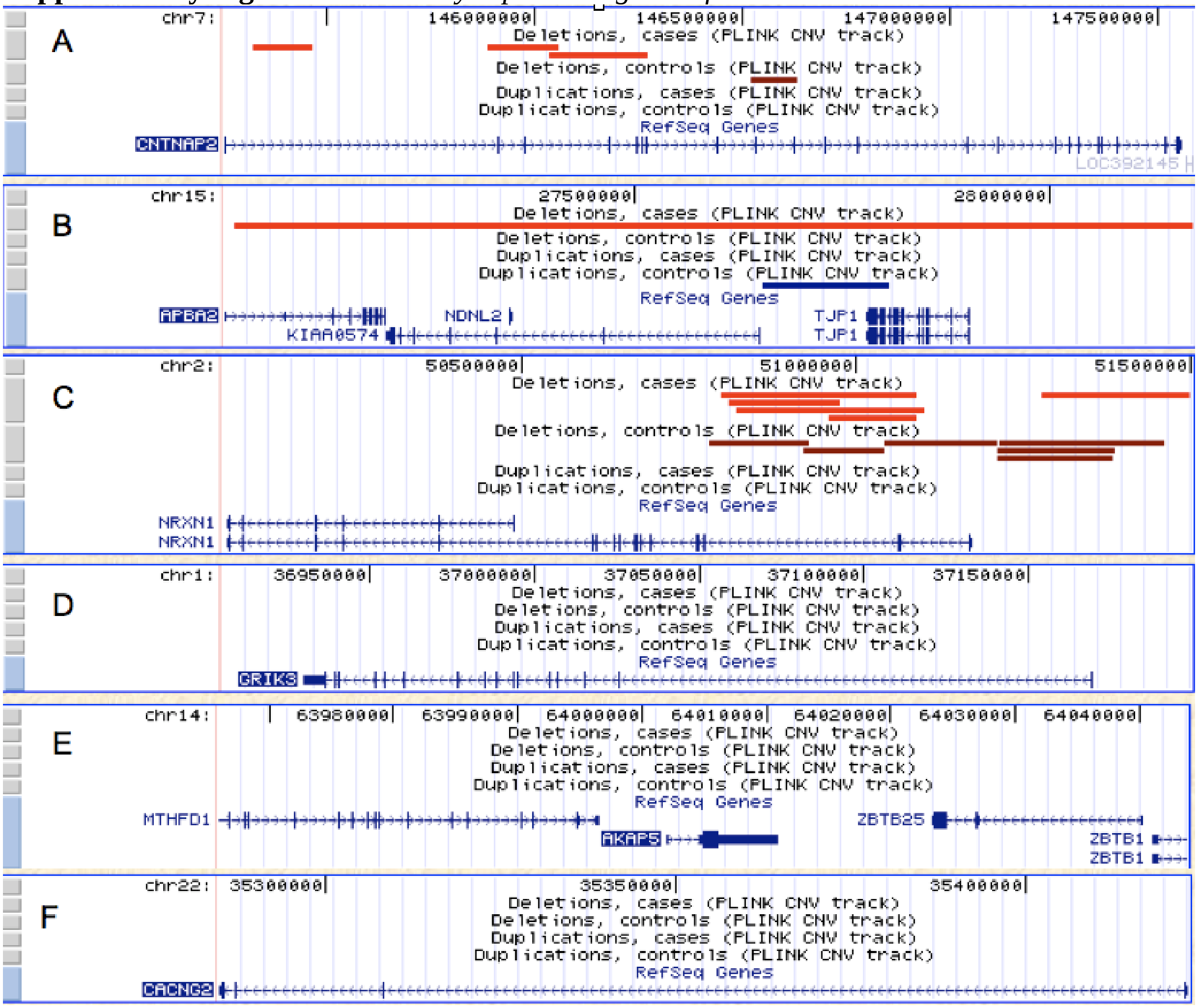
A)

B)

C)

Supplementary Figure Legend 2: Genomic interval of each of six regions previously reported to harbor CNVs in schizophrenia patients. In each plot, chromosome and base pair position are labeled in black along the top (hg17 coordinates). The bottom track of each plot includes RefSeq genes located within the interval of interest marked as blue horizontal lines with arrows to denote strand. Copy number events found in the current study are separated by cases and controls, deletions and duplications. Events are marked by horizontal colored bars (red = deletions, blue = duplications). A) Contactin associated 2 (*CNTNAP2*) region[20]; B) Chromosome 15, 27Mb to 28.4Mb[21]; C) Neurexin 1 (*NRXN1*) gene region[21]; D) Glutamate receptor (*GRIK3*, *GLUR7*) gene region[22]; E) A-kinase anchor protein 5 (*AKAP5*) gene region[22]; F) voltage-dependent calcium channel gamma-2 (*CACNG2*) gene region[22]

Supplementary Figure 2: *Previously reported regions of CNV in SCZ*


A
chr7:
146000000
146500000
147000000
147500000
Deletions, cases (PLINK CNV track)
Deletions, controls (PLINK CNV track)
Duplications, cases (PLINK CNV track)
Duplications, controls (PLINK CNV track)
RefSeq Genes
CNTNAP2
LOC392145
B
chr15:
27500000
28000000
Deletions, cases (PLINK CNV track)
Deletions, controls (PLINK CNV track)
Duplications, cases (PLINK CNV track)
Duplications, controls (PLINK CNV track)
RefSeq Genes
APBA2
NDNL2
KIAA0574
TJP1
TJP1
C
chr2:
50500000
51000000
51500000
Deletions, cases (PLINK CNV track)
Deletions, controls (PLINK CNV track)
Duplications, cases (PLINK CNV track)
Duplications, controls (PLINK CNV track)
RefSeq Genes
NRXN1
NRXN1
D
chr1:
36950000
37000000
37050000
37100000
37150000
Deletions, cases (PLINK CNV track)
Deletions, controls (PLINK CNV track)
Duplications, cases (PLINK CNV track)
Duplications, controls (PLINK CNV track)
RefSeq Genes
GRIK3
E
chr14:
63980000
63990000
64000000
64010000
64020000
64030000
64040000
Deletions, cases (PLINK CNV track)
Deletions, controls (PLINK CNV track)
Duplications, cases (PLINK CNV track)
Duplications, controls (PLINK CNV track)
RefSeq Genes
MTHFD1
ZBTB25
AKAP5
ZBTB1
ZBTB1
F
chr22:
35300000
35350000
35400000
Deletions, cases (PLINK CNV track)
Deletions, controls (PLINK CNV track)
Duplications, cases (PLINK CNV track)
Duplications, controls (PLINK CNV track)
RefSeq Genes
CACNG2

11. Supplementary References

1. American Psychiatry Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV), 4th edn.*, (American Psychiatric Association, Washington, DC, 2000).
2. Janca, A., Ustun, T.B., Early, T.S. & Sartorius, N. The ICD-10 symptom checklist: a companion to the ICD-10 classification of mental and behavioural disorders. *Soc Psychiatry Psychiatr Epidemiol* **28**, 239-42 (1993).
3. McGuffin, P., Farmer, A.E., Gottesman, II, Murray, R.M. & Reveley, A.M. Twin concordance for operationally defined schizophrenia. Confirmation of familiality and heritability. *Arch Gen Psychiatry* **41**, 541-5 (1984).
4. First, M.B., Spitzer, R. L., Gibbon, M., et al *Structured Clinical Interview for Axis I DSM–IV Disorders.*, (Biometrics Research, New York, 1994).
5. Wing, J.K. et al. SCAN. Schedules for Clinical Assessment in Neuropsychiatry. *Arch Gen Psychiatry* **47**, 589-93 (1990).
6. Endicott, J. & Spitzer, R.L. A diagnostic interview: the schedule for affective disorders and schizophrenia. *Arch Gen Psychiatry* **35**, 837-44 (1978).
7. Sklar, P. et al. Genome-wide scan in Portuguese Island families identifies 5q31-5q35 as a susceptibility locus for schizophrenia and psychosis. *Mol Psychiatry* **9**, 213-8 (2004).
8. Nurnberger, J.I., Jr. et al. Diagnostic interview for genetic studies. Rationale, unique features, and training. NIMH Genetics Initiative. *Arch Gen Psychiatry* **51**, 849-59; discussion 863-4 (1994).
9. Kendler, K.S., Lieberman, J.A. & Walsh, D. The Structured Interview for Schizotypy (SIS): a preliminary report. *Schizophr Bull* **15**, 559-71 (1989).
10. Rosen, W.G. et al. Positive and negative symptoms in schizophrenia. *Psychiatry Res* **13**, 277-84 (1984).
11. World Health Organization. *International Classification of Diseases*, (World Health Organization, Geneva, 1967).
12. World Health Organization. *International Classification of Diseases*, (World Health Organization, Geneva, 1978).
13. World Health Organization. *International Classification of Diseases*, (World Health Organization, Geneva, 1992).
14. Sklar, P. et al. Whole-genome association study of bipolar disorder. *Mol Psychiatry* (2008).
15. McCarroll, S., Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, deBakker PW, Maller J, Kirby A, Elliott AE, Parkin M, Hubbell E, Webster T, Mei R, Handsaker R, Lincoln S, Nizzari M, Blume J, Jones K, Rava R, Daly MJ, Gabriel SB, and Altshuler D Integrated detection and population-genetic analysis of SNPs and copy number variation *Nat Genet, in press* (2008).
16. Korn, J., Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Nizzari MM, Gabriel SB, Purcell SM, Daly MJ, Altshuler D. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms, and rare CNVs *Nat Genet, in press* (2008).
17. Gart, J. Alternative analyses of contingency tables. *J. Roy. Statist. Soc. Ser. B.* **28**, 164-179 (1966).

18. Walsh, T. et al. Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia. *Science* (2008).
19. Murphy, K.C. Schizophrenia and velo-cardio-facial syndrome. *Lancet* **359**, 426-30 (2002).
20. Friedman, J.I. et al. CNTNAP2 gene dosage variation is associated with schizophrenia and epilepsy. *Mol Psychiatry* **13**, 261-6 (2008).
21. Kirov, G. et al. Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. *Hum Mol Genet* **17**, 458-65 (2008).
22. Wilson, G.M. et al. DNA copy-number analysis in bipolar disorder and schizophrenia reveals aberrations in genes involved in glutamate signaling. *Hum Mol Genet* **15**, 743-9 (2006).