Copy Number Variation In Schizophrenia In Sweden

Table of Contents

Supplemental Methods

Case definition: rationale and validity

Cases were identified via the Hospital Discharge Register [1,2] which captures all public or private inpatient hospitalizations in Sweden. The register is complete from 1987 and augmented by psychiatric data from 1973-86. The register contains the dates and ICD discharge diagnoses [3-5] for each hospitalization, and capture the clinical diagnosis made by the attending physician. [6-9]

As described elsewhere, [10] our operational definition of schizophrenia includes two hospitalizations with a discharge diagnosis from the list below. The case definition of schizophrenia included the codes listed in ***Table S1*** (ICD-8 295, ICD-9 295, ICD-10 F20).

The ICD-8 and ICD-9 diagnosis of latent schizophrenia (295.5 and 295F) was excluded. Latent schizophrenia, also known as borderline, pre-psychotic or pseudo-neurotic schizophrenia, conforms more closely to a personality disorder in current psychiatric nosology.

The case definition used in most genetic studies of schizophrenia requires direct subject interview, review of medical records, and discussion with an informant (e.g., a psychiatrist familiar with the patient or a family member). This approach is effortful, and greatly increases the difficulty and expense of acquiring large samples.

Sample size is now a well-established limitation to progress in the genetic dissection of complex traits. [11] In this study, we pioneered a complementary strategy whereby we sought to establish caseness using a minimally adequate approach to diagnosis. In effect, our intent was to maximize sample size while ensuring that cases indeed had schizophrenia.

It is reasonable to ask whether the case definition used in this study corresponds to a more typical definition of schizophrenia. Given the importance of this issue, we conducted an extensive evaluation of our case definition prior to initiating sample collection. Multiple lines of evidence support the validity of our case definition.

First, many studies have conducted peer-reviewed research into the nature of schizophrenia using the Swedish Hospital Discharge Register (along with similar registers in other Scandinavian countries). In Sweden, as in other Nordic countries, the conceptualization of schizophrenia has historically been more influenced by biological theories of etiology. These factors have generally resulted in a conservative diagnostic approach (e.g. “the schizophrenia diagnosis has been given with great restriction in Swedish hospitals”). [1]

Data from Swedish and other Scandinavian population registers are generally accepted as informative for the epidemiology of schizophrenia. These registers have provided a wealth of information about risk factors for schizophrenia (***Table S2***).

Second, the Swedish Hospital Discharge Register has high agreement with medical [1,2] and psychiatric diagnoses. [12]

(a) Ekholm et al. [12] conducted a direct comparison of a Swedish register definition of schizophrenia with standard research diagnoses based on semi-structured interviews and medical records. They ascertained 143 patients with a diagnosis of schizophrenia from the Swedish Hospital Discharge Registry, abstracted medical records and conducted structured diagnostic interviews. DSM-IV diagnoses were assigned by a research psychiatrist based on all available data. Ekholm et al. concluded: *"94% of subjects … registered [≥1 time] with a diagnosis of schizophrenic psychoses (i.e. schizophrenia, schizoaffective psychosis or schizophreniform disorder) displayed a standard research DSM-IV diagnosis of these disorders."* [12] Research interviews added little new information. Thus, the Hospital Discharge Registry had a high level of agreement with research-grade diagnoses of schizophrenia.

An occasional source of disagreement was the presence of simple coding or transcription errors (e.g., incorrectly entering the ICD-9 code for schizophrenia, 295, instead of the code for short stature, 259). This is one reason why we required ≥2 admissions for schizophrenia.

(b) Co-author Dr Christina Hultman conducted a medical record review of 109 cases meeting our inclusion criteria using a structured checklist. She found that 97.2% (=106/109) met DSM-IV criteria for schizophrenia.

(c) Co-author Dr Shaun Purcell conducted an extensive evaluation of the consequences misclassification - what is the impact on statistical power if a few percent of cases are included as cases in error? Dr. Purcell evaluated the impact of misclassification rates of 2.5%, 5%, and 10%. He determined that the ratio of power with misclassification to no misclassification was 0.98, 0.95, and 0.91 for 2.5%, 5%, and 10% misclassification of cases. As anticipated for an uncommon disorder like schizophrenia (lifetime prevalence 0.4%), [13] misclassification does not substantially alter power.

Third, family history has historically been an important validator in psychiatric nosology. We conducted an extensive evaluation of our case definition of schizophrenia prior to initiating this study by combining the Hospital Discharge Register with the Multi-Generation Register [14] which allowed us to conduct a population-based, national genetic epidemiological study. [10]

Merging these Swedish national registers created a population-based cohort of 7,739,202 unique individuals of known parentage. These individuals clustered into 3,664,856 family groups encompassing first-, second-, and third-degree relatives. There were 32,536 individuals who met our criteria for schizophrenia (defined as ≥2 lifetime hospitalizations with a core schizophrenia discharge diagnosis). We noted the following findings: [10]

- The lifetime prevalence of schizophrenia was 0.407% (95% confidence interval, 0.402-0.411%), in close agreement to consensus estimates. [13]
- Of all family groups in sample, 1.267% (95% CI 1.255-1.280%) had at least one relative with schizophrenia, and the multiplex proportion was 3.81% (95% CI 3.62-4.00%) suggesting that most cases in this sample occur sporadically.
- λ_{sibs} was estimated at 8.55 without important sex differences. The recurrence risk estimates declined markedly if the definition of affection were relaxed by requiring just one admission for schizophrenia or if the definition was broadened to include schizophrenia spectrum disorders (data not shown).
- For second-degree relatives, the lowest numerical recurrence risk was for half-siblings (2.52) and the highest was for grandparents (3.80); however, there was substantial overlap of the confidence intervals for these estimates. First cousins were the only class of third-degree relatives for which we could confidently estimate recurrence risks (2.29).

Figure S7 [10] summarizes the results of this definition of schizophrenia in comparison to that taken to be true for schizophrenia. [13,15] Our results conform closely to the literature.

Fourth, our colleague Dr Paul Lichtenstein and colleagues reported in *The Lancet* estimates of the heritability of this definition of schizophrenia and its overlap with bipolar disorder in the combined Swedish Hospital Discharge Register / Multigenerational Register. The heritability of our definition of schizophrenia was 0.64 (95% CI 0.62-0.68) with small but significant common environmental effects (0.045). These results are similar to those from a far smaller meta-analysis of twin studies of schizophrenia. [16] The important overlap with bipolar disorder is now confirmed using GWAS results for both individual loci and a polygenic component. [11,17,18]

We note that few other samples in the world have direct estimates of the heritability and familiality of the schizophrenia phenotype under study.

Fifth, our definition of schizophrenia has passed peer review on multiple occasions, including two papers in *Nature* and one in *Nature Genetics*. [17-19] Our approach was also carefully vetted by the Schizophrenia Working Group of the PGC (led by Dr Kenneth Kendler) and found eligible for inclusion.

Sixth, as described in the accompanying manuscript, genomic findings in the Swedish samples are highly consistent with conventionally phenotyped cases. In particular, we note that sign tests comparing the Swedish samples with the PGC SCZ results were highly significant (0.76 or 154 of 201 SNPs with same direction of effect ,$p=8x10^{-15}$). The cases in the PGC mega-analysis were phenotyped using conventional methods (i.e., direct subject interviews, review of medical records, best estimated conferences). In addition, the Swedish results are similar to the PGC SCZ results in terms of rare CNV prevalences, CNV burden, common variation effect sizes, and polygenic profiles.

In summary, the validity of the definition of schizophrenia used in this study is strongly supported.

Diagnostic refinement

We attempted to improve upon the basic definition of SCZ. HDR data were obtained from all subjects considered eligible for this study. The base inclusion criterion for the study was ≥ 2 admissions with a diagnosis compatible with schizophrenia.

The data included admission/discharge dates, a primary diagnostic code plus up to seven additional diagnoses as ICD8, ICD9, or ICD10 codes. Diagnostic codes were assigned by the treating physician. These data were cleaned, examined for errors, and ICD codes converted to text.

These data were then matched against a manually curated list of flags for all ICD diagnoses. ***Table S15*** lists the core diagnostic flags for SCZ (34 diagnoses), schizoaffective disorder (SAD, 5), and bipolar disorder (BIP, 29). SCZ and SAD were used for the case definition. BIP is a key part of the differential diagnosis.

In addition, the discharge diagnoses were matched to a list of general medical conditions that serve as “organic” flags for psychosis (1,393 diagnoses). Psychosis can occur secondary to a general medical condition. This list was inclusive and had a comprehensive set of general medical conditions that could flag the presence of non-idiopathic SCZ (infections, neoplasms, endocrine, vascular disease, etc.).

HDR records for all potentially eligible cases (almost 400,000 discharge diagnoses across all subjects) were then reviewed.

First, all admissions and diagnoses with the “organic” flag set were manually reviewed by PFS (~30,000 diagnoses). The goal was to identify subjects to remove given the clear presence of a medical condition incompatible with idiopathic SCZ. This required the following conditions to be

met: (a) Plausible, the presence of a condition that medical judgment suggests is incompatible with idiopathic SCZ; (b) Not a risk factor. The presence of factors like cannabis use did not lead to exclusion (cannabis use is a risk factors for psychosis, but a causal path is not established); (c) Temporality. The condition preceded the development of psychosis (i.e., present since birth or present at first admission). Some conditions that developed well after onset of psychosis were allowed (e.g., the occurrence of stroke after multiple admissions for SCZ over decades); and (d) Consistent positive evidence in the HDR. Examples of general medical exclusions: congenital hypothyroidism, congenital syphilis, Mendelian diseases like Huntington's disease and porphyria, and myxedema. In addition, potential cases were excluded if the initial diagnosis was of a plausible medical condition which was then followed by admissions for SCZ (e.g., an initial diagnosis of frontal lobe neoplasm or encephalitis).

Some conditions were allowed, and did not lead to exclusion. Structural variants were were allowed. Brain structural abnormalities were allowed as some may result from SCZ (e.g., ventricular enlargement). Epilepsy was allowed as its relation to psychosis is complex. Non-specific congenital abnormalities were allowed. Head trauma/concussion was allowed unless there was evidence that it was devastating and present at initial admission. Thyroid disease was allowed unless consistently noted and present at all admissions.

Second, the timing and pattern of admissions were reviewed and descriptively evaluated at some length. This led to the following algorithm for diagnostic refinement:

- remove potential cases with manually-curated general medical conditions
- remove cases with < 2 admissions for SCZ or SAD after accounting for contiguous admissions
- remove cases with total inpatient stay < 7 days
- remove cases where bipolar disorder was the dominant discharge diagnosis
- remove cases where drug/alcohol predominated

These exclusions led to the removal of 3.4% of eligible cases due to the primacy of another psychiatric disorder (0.9%) or a general medical condition (0.3%) or uncertainties in the Hospital Discharge Register (e.g., contiguous admissions with brief total duration, 2.2%).

Subject ascertainment

Cases were ascertained from all of Sweden using the Hospital Discharge Register from 2005-11, and the sampling frame is thus population-based and covers all hospital-treated patients.

All procedures were approved by ethical committees in Sweden and in the US, and all subjects provided written informed consent (or legal guardian consent and subject assent). We also obtained permissions from the area health board to which potential subjects were registered.

Potential cases 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 conducted a brief interview about other medical conditions in a lifetime.

Controls were also identified from national population registers, and had never received a discharge diagnosis of SCZ or bipolar disorder. 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.

Quality Controls

SNP-based subject QC

Genotypes were called using Birdsuite (Affymetrix) or BeadStudio (Illumina). Multi-step quality control (QC) procedures were carried out using SNP genotypes. The exclusionary measures were: SNP missingness ≥ 0.05 (before sample removal); subject missingness ≥ 0.02; autosomal heterozygosity deviation; SNP missingness ≥ 0.02 (after sample removal); difference in SNP missingness between cases and controls ≥ 0.02; and deviation from Hardy-Weinberg equilibrium ($P < 10^{-6}$ in controls or $P < 10^{-10}$ in cases).

After basic quality control, 77,986 autosomal SNPs directly genotyped on all three GWAS platforms were extracted and pruned to remove SNPs in LD ($r^2 > 0.05$) or with minor allele frequency < 0.05, leaving 39,239 SNPs suitable for robust relatedness testing. Relatedness testing was done with PLINK [20] and pairs of subjects with $\hat{\pi} > 0.2$ were identified and one member of each relative pair removed at random.

Intensity-based subject QC

The SNP-based QC excluded most gross sample failures. Nonetheless, to measure whether an assay is useful for CNV analysis, probe-intensity-based metrics have been established, including MAPD and waviness. MAPD is a measure of probe variance and is defined as the median of the absolute values of all pairwise differences between $\log_2$ ratios for a given genotyping array. MAPD is robust against high biological variability in $\log_2$ ratios induced by large CNVs. GC wave or waviness describes a spatial "wave" pattern in $\log_2$ ratios and is a systematic technical artifact observed in various array platforms. As shown in ***Table S2***, we removed problematic arrays with high MAPD or waviness using empirically derived thresholds (i.e., exceeding 3 standard deviations from the sample mean per array type). Furthermore, we visualized pseudo-color images of all excluded arrays and a random sample of the good performing arrays to ensure the intensity-based QC worked properly. Example array images are shown in ***Figure S1***.

CNV-load-based subject QC

In addition, we removed 14 individuals who were outliers with respect to the total number or length of CNVs (>40 CNVs or total CNV spanning >6Mb). Thresholds were empirically derived as mean + 3xSD in the post-QC sample and by observing the distributions of these metrics across the entire dataset.

CNV validation with Illumina Human Exome BeadChips

Exome array. The 250K SNPs on Illumina Human Exome BeadChips were derived from exome sequencing of 12,028 European subjects (including ~500 subjects from this study), and met the following criteria: exonic or splice site variant of predicted functionality, minor allele observed a total of ≥ 3 times, minor allele observed in ≥ 2 different cohorts, passed sequencing quality control, and high Illumina SNP design scores. The exome array includes at least one SNP in 79% of all genes, comparable to GWAS arrays (81% Affymetrix 6.0, 82% Illumina Omni Express).

Exome array genotyping and quality control. Genotyping was done at the Broad Institute. We used 96-well plates for processing using the Illumina Infinium HumanExome BeadChip v1.0. The majority of Exome genotypes were called using GenomeStudio v2010.3 with the calling algorithm/genotyping module version 1.8.4 using the custom cluster file

StanCtrExChp_CEPH.egt. Subsequent processing of genotype calling was done by zCall[21]. The Broad Institute did not filter any SNPs based off of technical quality control metrics. Only samples passing an overall call rate of 98% criteria and standard identity check were released.

CNV calling and quality control from exome arrays. CNV calling began with raw intensity data processing. A custom cluster file was created using the GenCall algorithm based on all samples. Normalized intensity values were obtained using Illumina's GenomeStudio (v2010.3) with the calling algorithm/genotyping module (v1.8.4). For CNV calling, PennCNV (June 2011 version) [22] was applied to the log R ratios (LRR) and B allele frequencies (BAF) calculated from the normalized intensity values. The default waviness correction and customized PennCNV parameters were used. [23] Low-confidence CNVs were excluded (confidence scores < 10). Low quality samples were excluded if they had extreme values for probe variance (i.e. LRR_standard deviation > 0.2, 95th percentile or BAF_drift > 0.01, 95th percentile), or were outliers with respect to the total number of CNV calls (>152, 95th percentile).

CNV validation. For each CNV, we checked whether a CNV was also detected from the exome arrays in the same sample (defined by ≥1 bp overlap).

Replication samples

Overview of the Replication Samples

We obtained replication association results from 6,882 schizophrenia cases and 11,255 controls. Cases were from the United Kingdom CLOZUK [24] and CardiffCOGS samples. Cases were genotyped at the Broad Institute using Illumina OmniExpress or OmniCombo arrays.

Controls were from four external studies of non-psychiatric disorders. The control datasets were chosen as they were genotyped on Illumina arrays similar to those used for the cases (Illumina Human Omni2.5, Illumina HumanOmni1_Quad, or Illumina 1.2M).

- The Genetic Architecture of Smoking and Smoking Cessation (dbGaP, phs000404.v1.p1)
- High Density SNP Association Analysis of Melanoma: Case-Control and Outcomes Investigation (dbGaP, phs000187.v1.p1)
- Genetic Epidemiology of Refractive Error in the KORA Study (dbGaP, phs000303.v1.p1)
- WTCCC2 project samples from National Blood Donors Cohort (European Genome-Phenome Archive, EGAD00000000024)
- WTCCC2 project samples from 1958 British Birth Cohort (European Genome-Phenome Archive, EGAD00000000022).

dbGaP http://www.ncbi.nlm.nih.gov/gap. European Genome-Phenome Archive https://www.ebi.ac.uk/ega.

CNV Calling and Quality Control in the Replication Samples

Principal component analysis (PCA) was performed to derive ethnicities of the samples. Identity by decent (IBD) was performed to identify and remove duplicate individuals. All coordinates are according to UCSC build 37, hg19.

Raw intensity data from each case/control dataset were independently processed and analysed to account for potential batch effects. Log2ratios and B-allele frequencies were generated using Illumina Genome Studio software (v2011.1). CNVs were called using the PennCNV calling algorithm, following the standard protocol and adjusting for GC content. CNVs were called using the 520,766 probes common to all discovery arrays to void a cross-platform CNV locus

detection bias. Samples were excluded if for any one of the following QC metrics they represented an outlier in their source dataset: Log2ratio standard deviation, B-allele frequency drift, wave factor and total number of CNVs called per person.

Following the exclusion of poorly performing samples, we performed quality control on the called CNVs. Firstly, CNVs in the same individual were joined if the distance separating them was less than 50% of their combined length using a custom developed open source programme (http://x004.psycm.uwcm.ac.uk/~dobril/combine_CNVs/). All CNVs were then excluded if they were covered by less than 10 probes, were less than 10kb in length, overlapped with low copy repeats by more than 50% of their length, or had a probe density (calculated by dividing the size of the CNV by the number of probes covering it) greater than 1 probe/20kb. CNV loci with a frequency > 1% of the total discovery sample were excluded using PLINK.

The remaining rare CNVs were required to pass a median Z-score outlier method of validation. This method is detailed in Kirov et al (2012). [PMID: 22083728] Briefly, each probe intensity within an individual is converted to a Z-score, which is the probe intensity standardised across all probes within that individual, and then standardised for that probe across all individuals. These rounds of standardisation help reduce noise created by natural fluctuations in probe intensity. A median Z-score value for all probes within a putative CNV region is used to assess copy number, with true deletions and duplications represented as outliers in the samples median Z-score distribution. Each CNV in every individual was assigned a Z-score. CNVs with Z-scores of <-6 were accepted as true deletions, while those with Z-sores of >+3 were accepted as duplications. The Z-score histograms of CNVs with marginal Z-Scores (deletion Z-score between -4 and -6 and duplication Z-score between +2 and +3) were manually inspected, and from these CNVs the Log2ratios and B-allele frequencies of those with ambiguous Z-scores were visually imspected with the Illumina GenomeStudio v2011.1 software. This resulted in 2,569 CNVs being filtered out from the data.

Supplemental Tables

Table S1. Diagnostic codes

ICD	Code	ICD class	ICD subclass	SCZ	SAD	BIP
8	295	Schizophrenia	Schizophrenia	1	0	0
8	295.0	Schizophrenia	Simple type	1	0	0
8	295.1	Schizophrenia	Hebephrenic type	1	0	0
8	295.2	Schizophrenia	Catatonic type	1	0	0
8	295.3	Schizophrenia	Paranoid type	1	0	0
8	295.4	Schizophrenia	Acute schizophrenia episode	1	0	0
8	295.6	Schizophrenia	Residual schizophrenia	1	0	0
8	295.7	Schizophrenia	Schizo-affective type	0	1	0
8	295.8	Schizophrenia	Other	1	0	0
8	295.9	Schizophrenia	Unspecified type	1	0	0
8	296.1	Affective psychoses	Manic-depression psychosis, manic type	0	0	1
8	296.2	Affective psychoses	Manic depressive psychosis, depressed type	0	0	1
8	296.3	Affective psychoses	Manic-depressive psychosis, circular type	0	0	1
8	296.8	Affective psychoses	Other	0	0	1
8	296.9	Affective psychoses	Unspecified	0	0	1
9	295	Schizophrenic dis	Schizophrenic dis	1	0	0
9	295.0	Schizophrenic dis	Simple type	1	0	0
9	295.1	Schizophrenic dis	Disorganized type	1	0	0
9	295.2	Schizophrenic dis	Catatonic type	1	0	0
9	295.3	Schizophrenic dis	Paranoid type	1	0	0
9	295.4	Schizophrenic dis	Schizophreniform dis	1	0	0
9	295.6	Schizophrenic dis	Residual type	1	0	0
9	295.7	Schizophrenic dis	Schizoaffective dis	0	1	0
9	295.8	Schizophrenic dis	Other specified types of schizophrenia	1	0	0
9	295.9	Schizophrenic dis	Unspecified schizophrenia	1	0	0
9	296.0	Episodic mood dis	Bipolar I dis, single manic episode	0	0	1
9	296.1	Episodic mood dis	Manic dis, recurrent episode	0	0	1
9	296.4	Episodic mood dis	Bipolar I dis, most recent episode (or current) manic	0	0	1
9	296.5	Episodic mood dis	Bipolar I dis, most recent episode (or current) dep	0	0	1
9	296.6	Episodic mood dis	Bipolar I dis, most recent episode (or current) mixed	0	0	1
9	296.7	Episodic mood dis	Bipolar I dis, most recent epis (or current) unspecified	0	0	1
9	296.89	Episodic mood dis	Other & unspec bipolar dis; atyp manic, atyp depress	0	0	1
9	296.99	Episodic mood dis	Other & unspec episodic mood dis	0	0	1
9	V11.0	Personal hx ment dis	Schizophrenia	1	0	0
10	F20	Schizophrenia	Schizophrenia	1	0	0
10	F20.0	Schizophrenia	Paranoid schizophrenia	1	0	0
10	F20.1	Schizophrenia	Hebephrenic schizophrenia	1	0	0
10	F20.2	Schizophrenia	Catatonic schizophrenia	1	0	0
10	F20.3	Schizophrenia	Undifferentiated schizophrenia	1	0	0
10	F20.4	Schizophrenia	Postschizophrenic depression	1	0	0
10	F20.5	Schizophrenia	Residual schizophrenia	1	0	0
10	F20.6	Schizophrenia	Simple schizophrenia	1	0	0
10	F20.8	Schizophrenia	Other schizophrenia	1	0	0
10	F20.9	Schizophrenia	Schizophrenia, unspecified	1	0	0
10	F23.1	Acute/trans psychotic dis	Acute polymorphic psychotic dis with sx of SCZ	1	0	0
10	F23.2	Acute/trans psychotic dis	Acute schizophrenia-like psychotic dis	1	0	0
10	F25	Schizoaffective dis	Schizoaffective dis	0	1	0
10	F25.0	Schizoaffective dis	Schizoaffective dis, manic type	0	1	0
10	F25.1	Schizoaffective dis	Schizoaffective dis, depressive type	1	0	0
10	F25.2	Schizoaffective dis	Schizoaffective dis, mixed type	0	1	0
10	F25.8	Schizoaffective dis	Other schizoaffective dis	1	0	0
10	F25.9	Schizoaffective dis	Schizoaffective dis, unspecified	1	0	0
10	F30	Manic episode	Manic episode	0	0	1
10	F30.1	Manic episode	Mania without psychotic symptoms	0	0	1
10	F30.2	Manic episode	Mania with psychotic symptoms	0	0	1
10	F30.8	Manic episode	Other manic episodes	0	0	1
10	F30.9	Manic episode	Manic episode, unspecified	0	0	1
10	F31	Bipolar affective dis	Bipolar affective dis	0	0	1
10	F31.0	Bipolar affective dis	Bipolar affective dis, current epi hypomanic	0	0	1
10	F31.1	Bipolar affective dis	Bipolar affective dis, current epi manic not psychotic	0	0	1
10	F31.2	Bipolar affective dis	Bipolar affective dis, current epi manic with psychotic	0	0	1
10	F31.3	Bipolar affective dis	Bipolar affective dis, current epi mild or mod dep	0	0	1
10	F31.4	Bipolar affective dis	Bipolar aff dis, current epi sev dep without psychotic	0	0	1
10	F31.5	Bipolar affective dis	Bipolar aff dis, current epi sev dep with psychotic	0	0	1
10	F31.6	Bipolar affective dis	Bipolar affective dis, current episode mixed	0	0	1
10	F31.7	Bipolar affective dis	Bipolar affective dis, currently in remission	0	0	1
10	F31.8	Bipolar affective dis	Other bipolar affective dis	0	0	1
10	F31.9	Bipolar affective dis	Bipolar affective dis, unspecified	0	0	1

Table S2. Risk factors for schizophrenia using Swedish national register data

Type	Risk factor	Review	Schizophrenia risk in Sweden
Natal	Seasonality Urban birth Paternal age SES	25,26 27,28 29,30 31	⇑ Jan-Apr births (OR 1.4) [6] ⇑ urban (hazard ratio 1.3) [32] ⇑ older (hazard ratio 1.47/decade) [33] ⇑ lower [34]
Obstetric	Pregnancy Abnormal growth Delivery	35,36	⇑ bleeding (OR 2.0) [6] ⇑ low birth weight (OR 1.5-2.0) [37] ⇑ preeclampsia (OR 2.5) [38]
Cognition	Intelligence	39	⇓ pre-morbid IQ (0.5 SD) [40]

Table S3. Metrics for intensity-based QC

Genotyping batch	Sw1	Sw2,3,4	Sw5,6
GWAS array types	Affymetrix 5.0	Affymetrix 6.0	IlluminaOmni Express
Total probe sets	443,816 (SNP)	909,622 (SNP) 945,826 (CN)	733,202 (SNP)
Software	Affymetrix Power Tools apt-copynumber-workflow	Affymetrix Power Tools apt-copynumber-workflow	PennCNV Genomic_wave.pl
Major parameters and high quality reference values[1]	MAPD<0.4 Waviness.SD<0.11	MAPD<0.4 Waviness.SD<0.11	MAPD<0.19 |Waviness.factor|<0.05

Thresholds are based on mean + 3xSD per array type. For Illumina arrays, log R ratios (LRR) were used to compute MAPD.

Table S4. Summary of Subject Quality Control

Feature	Sw1	Sw2	Sw3	Sw4	Sw5	Sw6	Total
Subjects (pre-QC)	464	694	1,498	2,388	4,461	2,345	11,850
After SNP-based QC	427	643	1,356	2,261	4,361	2,194	11,242
After intensity-based QC	413	620	1,312	2,062	4,132	2,111	10,650
After CNV-load-based QC	413	616	1,310	2,058	4,130	2,109	10,636
With eligible exome arrays	330	508	1,196	1,869	3,905	1,900	9,708

Case/control distribution	Sw1	Sw2	Sw3	Sw4	Sw5	Sw6	Total
After CNV-load-based QC	413	616	1,310	2,058	4,130	2,109	10,636
# controls	206	233	830	1,074	2,456	1,118	5,917
# cases with SCZ	207	383	480	984	1,674	991	4,719

Note: Data collection for this study took six years (2005-2011). GWAS genotyping was conducted in six separate batches (denoted Sw1-Sw6) using three GWAS chips (Affymetrix 5.0, Affymetrix 6.0, and Illumina Omni Express). Genotypes were generated as sufficient numbers of samples accumulated from the field work in Sweden. Thus there were six genotyping batches and there were slight differences in case control ratio between batches.

Table S5 Validation using exome arrays for genic CNVs ≥400kb

Wave	Array platform	# Subjects (% Total)	# Subjects with exome array validation	# DEL	# (%) Validated DEL	# DUP	# (%) Validated DUP	Average validation rate
Sw1	Affymetrix 5.0	413 (3.9%)	307	2	2 (100%)	7	6 (86%)	89%
Sw2,3,4	Affymetrix 6.0	3,984 (37.4%)	3,030	25	24 (96%)	143	116 (82%)	83%
Sw5,6	Illumina Omni Express	6,239 (58.7%)	5,763	65	65 (100%)	174	154 (89%)	92%

The same DNA samples from all cases and controls were genotyped on both GWAS arrays and Illumina exome arrays. Previously, we developed CNV calling procedures for exome array data (essentially, an exon-focused set of 250K probes), and have shown that the exome array has high sensitivity and specificity to identify genic CNVs ≥ 400kb. Therefore, we used these additional data for large-scale validation. We contrasted the genome-wide array CNVs used in this paper to exome array CNVs, stratified by array type. A CNV is considered validated if it is ≥400kb and it is overlapped by an exome array CNV in the same sample by 50% of its length. ***Table S16*** displays the results for GWAS array deletions (DEL) and duplications (DUP) separately and combined.

Table S6. Validation of genomic outliers

Sample ID	Wave	Array	Status	Affected chromosome	Size (mb)	Exome array
PT-8K83	Sw6	ILMN	Control	chr3 trisomy †	123.7	Bad chip
PT-289M	Sw2	Affy 6	SCZ	chr8 trisomy †, chr15 del	97.6; 1.3	Chr8 confirmed
PT-8U7L	Sw3	Affy 6	SCZ	chr4 dup, chr15 dup	10.0; 2.0	Chr4 confirmed
PT-ES7Q	Sw5	ILMN	SCZ	chr9 dup	10.4	Confirmed
PT-BPII	Sw4	Affy 6	Control	chr13 dup	11.5	Confirmed
PT-2M38	Sw2	Affy 6	Control	chr15 dup	2.0	No coverage
PT-9ZDU	Sw4	Affy 6	SCZ	Potential mosaic chr13, chr11	1.4	Confirmed

† Confirmed using qPCR. [41]

Table S7. Linear models of CNV burden: batch, ancestry, sex, age.

We fit multiple linear regression models where the dependent variable is CNV burden (total number, total KB, or gene counts) and the independent variables are phenotype (case/control status), batch, ancestry (the first 4 principal components), sex, and age. The ANOVA table from each regression is displayed below. Two predictors are significant (P<0.002 multiple-testing-adjusted cutoff), namely “phenotype” as expected and “batch” as a significant confounder.

We note that principal component 3 (PC3/c3) should not confound our analysis for the following reasons: (1) it is not significant after multiple-testing adjustment. (2) The variance explained by PC3 is much smaller compared to the variance explained by genotyping batch, which we included as a covariate in our analysis. (3) Critically, even when we included PC3 as a covariate, the qualitative results do not change.

Response: Total.number.CNV

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Phenotype	1	15.0	14.978	14.8230	0.0001188	***
Batch	5	615.0	123.007	121.7327	< 2.2e-16	***
c1	1	0.9	0.857	0.8479	0.3571578	
c2	1	0.1	0.051	0.0501	0.8228973	
c3	1	5.0	5.034	4.9818	0.0256365	
c4	1	0.5	0.541	0.5357	0.4642313	
sex	1	1.8	1.773	1.7549	0.1852879	
age.at.sampling	1	3.0	2.979	2.9483	0.0859985	
Residuals	10417	10526.0	1.010			

Response: Total.CNV.KB

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Phenotype	1	7224345	7224345	30.1311	4.133e-08	***
Batch	5	57601352	11520270	48.0484	< 2.2e-16	***
c1	1	38080	38080	0.1588	0.69025	
c2	1	588288	588288	2.4536	0.11728	
c3	1	286154	286154	1.1935	0.27465	
c4	1	90736	90736	0.3784	0.53845	
sex	1	3869	3869	0.0161	0.89892	
age.at.sampling	1	729098	729098	3.0409	0.08122	
Residuals	10417	2497617849	239764			

Response: Gene Count

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Phenotype	1	485	485.11	24.8614	6.26e-07	***
Batch	5	4774	954.74	48.9293	< 2.2e-16	***
c1	1	13	13.17	0.6750	0.4113	
c2	1	0	0.03	0.0014	0.9703	
c3	1	1	0.65	0.0333	0.8552	
c4	1	3	2.90	0.1487	0.6998	
sex	1	10	9.51	0.4873	0.4851	
age.at.sampling	1	28	28.36	1.4532	0.2280	
Residuals	10417	203264	19.51			

Table S8 CNV characteristics

Sample characteristics	Cases	Controls
Subjects (after quality control)		
Sw1 (Affymetrix 5.0)	207	206
Sw2-4 (Affymetrix 6.0)	1,847	2,137
Sw5-6 (Illumina Omni Express)	2,665	3,574
Total sample	4,719	5,917
Mean Number of CNVs per subject		
Sw1	1.058	0.845
Sw2-4	1.236	1.209
Sw5-6	0.758	0.684
Total sample	0.958	0.879
Mean number of >500Kb CNVs per subject		
Sw1	0.184	0.141
Sw2-4	0.176	0.147
Sw5-6	0.107	0.080
Total sample	0.137	0.106
Mean number of singleton CNVs per subject		
Sw1	0.097	0.058
Sw2-4	0.103	0.086
Sw5-6	0.078	0.059
Total sample	0.089	0.069

Table S9 Global CNV burden analysis (number, gene count, length): event type and frequency

S9a: CNV number

CNV type	Frequency	CNV (n)	Case/Control ratio	Baseline rate (controls)	OR	95% CI	Empirical P value
Deletions and duplications	All	9723	1.09	0.88	1.07	(1.03,1.11)	0.0003
	1x	827	1.29	0.07	1.27	(1.12,1.45)	0.0002
	2–6x	7747	1.06	0.71	1.05	(1,1.09)	0.02
	≥7x	5115	1.06	0.45	1.04	(0.987,1.1)	0.07
Deletions	All	3727	1.09	0.34	1.08	(1.02,1.16)	0.007
	1x	490	1.32	0.04	1.34	(1.12,1.6)	0.0006
	2–6x	2804	1.03	0.26	1.02	(0.948,1.1)	0.306
	≥7x	1757	0.97	0.17	0.96	(0.878,1.06)	0.78
Duplications	All	5996	1.09	0.54	1.07	(1.02,1.12)	0.005
	1x	663	1.16	0.06	1.14	(0.989,1.31)	0.039
	2–6x	4588	1.09	0.41	1.06	(1.01,1.12)	0.016
	≥7x	2925	1.09	0.26	1.07	(0.996,1.15)	0.03

CNVs are <1%, ≥100kb, and spanning ≥15 probes. Empirical P values were obtained in PLINK by 100,000 permutations and permuting phenotype labels within genotyping batches. A total of 81 burden tests were conducted in Tables S9, S10, S11, and S12, thus the multiple-testing-adjusted P value cutoff based on Bonferroni method is 0.0006. Odds ratios were computed in R by fitting a logistic regression model of logit(Prb(case)) ~ burden + batch, which indicate

increase in risk for SCZ per unit increase of CNV burden. Allele categories. “x” meaning occurrence. The allele frequency is computed in PLINK using the default regional-based method and overlapping parameter (--cnv-overlap 0). Allele frequency in the Swedish sample: 1x (single occurrence, <0.0001): 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. 2-6x (2 to 6 occurrences; 0.0001-0.0005): CNVs which had ≥1bp of their length spanning any one consecutive region containing 2 to 6 CNVs in the total sample. ≥7x (7 or more occurrences; 0.0005-0.01): CNVs which had ≥1bp of their length spanning any one consecutive region containing 7 or more CNVs in the total sample.

S9b: Gene count

CNV type	*Frequency*	*Case/Control ratio*	*Control rate (number genes)*	*OR (per genes)*	*95% CI*	*Empirical P value*
Deletions and duplications	All	1.23	1.88	1.02	(1.01,1.03)	1E-05
	1x	1.32	0.15	1.04	(1.01,1.08)	0.009
	2–6x	1.19	1.45	1.02	(1.01,1.03)	0.0005
	≥7x	1.11	0.99	1.01	(0.998,1.02)	0.05
Deletions	All	1.38	0.52	1.04	(1.02,1.05)	1E-05
	1x	1.69	0.06	1.07	(1.01,1.12)	0.005
	2–6x	1.28	0.39	1.03	(1.01,1.05)	0.002
	≥7x	1.06	0.25	1.01	(0.972,1.04)	0.37
Duplications	All	1.18	1.36	1.02	(1.01,1.03)	0.005
	1x	1.10	0.15	1.01	(0.98,1.05)	0.241
	2–6x	1.15	0.99	1.01	(1,1.03)	0.021
	≥7x	1.07	0.61	1.01	(0.989,1.02)	0.25

ORs indicate increase in risk for SCZ per gene affected by CNVs.

S9c: Total CNV length

CNV type	Frequency	Case/Control ratio	Control rate (Mb)	OR (per 100kb)	95% CI	Empirical P value
Deletions and duplications	All	1.14	0.468	1.02	(1.01,1.03)	3E-05
	1x	1.16	0.256	1.08	(1.03,1.12)	0.022
	2–6x	1.11	0.406	1.02	(1.01,1.03)	0.0006
	≥7x	1.09	0.334	1.02	(1.01,1.04)	0.004
Deletions	All	1.19	0.320	1.03	(1.02,1.04)	0.0001
	1x	1.38	0.234	1.09	(1.04,1.15)	0.005
	2–6x	1.04	0.293	1.02	(0.999,1.04)	0.013
	≥7x	1.02	0.266	0.995	(0.969,1.02)	0.35
Duplications	All	1.10	0.431	1.02	(1.01,1.03)	0.002
	1x	1.09	0.325	1.02	(0.986,1.05)	0.204
	2–6x	1.09	0.380	1.02	(1.01,1.03)	0.027
	≥7x	1.11	0.297	1.03	(1.01,1.05)	0.006

ORs indicate increase in risk for SCZ per 100kb of CNV.

Table S10 Global CNV burden analysis of CNV number: event type and size

S10a: CNV number

CNV type	Frequency	CNV (n)	Case/Control ratio	Control rate	OR	95% CI	Empirical P value
Deletions and duplications	100-200kb	5283	1.04	0.487	1.03	(0.978,1.09)	0.126
	200-500kb	3163	1.09	0.285	1.08	(1.01,1.16)	0.0104
	500kb+	1277	1.29	0.106	1.26	(1.13,1.4)	1.00E-05
Deletions	100-200kb	2272	1.05	0.209	1.05	(0.962,1.14)	0.153
	200-500kb	1060	1.06	0.097	1.07	(0.944,1.21)	0.16
	500kb+	395	1.42	0.0313	1.4	(1.16,1.7)	0.00044
Duplications	100-200kb	3011	1.04	0.278	1.02	(0.953,1.1)	0.273
	200-500kb	2103	1.11	0.188	1.09	(1.01,1.19)	0.0176
	500kb+	882	1.24	0.075	1.19	(1.05,1.36)	0.00427

S10b: Gene count

CNV type	Frequency	Case/Control ratio	Control rate (number genes)	OR (per gene)	95% CI	Empirical P value
Deletions and duplications	100-200kb	1.03	0.696	1.00	(0.98,1.03)	0.384
	200-500kb	0.996	0.615	0.997	(0.977,1.02)	0.602
	500kb+	1.74	0.568	1.03	(1.02,1.05)	1.00E-05
Deletions	100-200kb	1.11	0.231	1.02	(0.983,1.06)	0.153
	200-500kb	0.872	0.151	0.967	(0.921,1.01)	0.916
	500kb+	2.38	0.14	1.06	(1.04,1.09)	1.00E-05
Duplications	100-200kb	0.993	0.465	0.993	(0.964,1.02)	0.674
	200-500kb	1.04	0.464	1.00	(0.982,1.03)	0.355
	500kb+	1.53	0.428	1.02	(1.01,1.04)	0.00011

ORs indicate increase in risk for SCZ per gene affected by CNVs.

S10c: Total CNV length

CNV type	Frequency	Case/Control ratio	Control rate (MB)	OR (per 100kb)	95% CI	Empirical P value
Deletions and duplications	100-200kb	1.01	0.180	1.02	(0.985,1.06)	0.455
	200-500kb	1.01	0.381	1.03	(1.01,1.05)	0.275
	500kb+	1.09	1.050	1.02	(1.01,1.03)	0.0279
Deletions	100-200kb	1.02	0.153	1.04	(0.976,1.1)	0.165
	200-500kb	0.985	0.330	1.02	(0.981,1.06)	0.769
	500kb+	1.23	1.030	1.03	(1.02,1.05)	0.0072
Duplications	100-200kb	1.02	0.163	1.01	(0.965,1.07)	0.209
	200-500kb	1.04	0.367	1.03	(1,1.05)	0.048
	500kb+	1.02	1.030	1.02	(1,1.03)	0.328

ORs indicate increase in risk for SCZ per 100kb of CNV.

Table S11 Global CNV burden analysis of single-occurrence CNVs: event type and size

CNV number							
CNV type	**Frequency**	**CNV (n)**	**Case/Control ratio**	**Control rate**	**OR**	**95% CI**	**Empirical P value**
Deletions and duplications	100-200kb	478	1.16	0.042	1.16	(0.975,1.39)	0.052
	200-500kb	278	1.51	0.021	1.51	(1.2,1.9)	0.00032
	500kb+	71	1.36	0.006	1.34	(0.842,2.13)	0.133
Deletions	100-200kb	204	1.33	0.017	1.34	(1.02,1.76)	0.022
	200-500kb	89	1.68	0.006	1.72	(1.13,2.62)	0.007
	500kb+	13	2.01	0.001	2.07	(0.679,6.34)	0.155
Duplications	100-200kb	275	1.05	0.025	1.06	(0.836,1.33)	0.350
	200-500kb	189	1.44	0.015	1.43	(1.09,1.88)	0.007
	500kb+	58	1.25	0.005	1.22	(0.73,2.03)	0.266

Single-occurrence CNVs are those only observed once in our data, in either a case or control. These were conservatively defined as having no overlap with any other CNVs. PLINK command used: --cnv-overlap 0 --cnv-freq-exclude-above 1. The allele frequency of the single-occurrence CNVs in the Swedish sample is 0.000094.

Table S12 Global CNV burden analysis of >500Kb CNVs: event type and frequency

CNV number							
CNV type	**Frequency**	**CNV (n)**	**Case/Control ratio**	**Baseline rate (controls)**	**OR**	**95% CI**	**Empirical P value**
Deletions and duplications	1x	180	1.23	0.015	1.22	(0.91,1.63)	0.105
	2-6x	868	1.3	0.072	1.26	(1.1,1.44)	0.0003
	≥7x	438	1.31	0.036	1.27	(1.06,1.53)	0.007
Deletions	1x	74	1.74	0.005	1.77	(1.11,2.82)	0.011
	2-6x	257	1.22	0.022	1.22	(0.96,1.56)	0.057
	≥7x	113	1	0.011	0.984	(0.683,1.42)	0.570
Duplications	1x	164	1.06	0.015	1.04	(0.763,1.41)	0.441
	2-6x	559	1.33	0.046	1.27	(1.08,1.49)	0.002
	≥7x	220	1.4	0.018	1.34	(1.03,1.74)	0.019

Table S13 Duplications at 17q12 and 22q11.2 from both GWAS and exome arrays

(1) For 17q12, PT-L191 was detected from GWAS array but did not have eligible exome array. (2) For 22q11.2, all events >500Kb are shown. PT-BQOL was detected from exome array but did not have any eligible GWAS array.

17q12							
GWAS arrays							
FID	**IID**	**PHE**	**CHR**	**BP1**	**BP2**	**TYPE**	**KB**
PT-9Z95	1	2	17	35057020	36295000	DUP	1238
PT-L1MG	1	2	17	34917969	36195934	DUP	1278
PT-M8JR	1	2	17	35113905	36195934	DUP	1082
PT-OOQ4	1	2	17	34597521	36195934	DUP	1598
PT-OPSR	1	1	17	34815551	36295000	DUP	1479
PT-L191	1	2	17	35512253	36140519	DUP	628.3
Exome Arrays							
FID	**IID**	**PHE**	**CHR**	**BP1**	**BP2**	**TYPE**	**KB**
PT-9Z95	1	2	17	34819191	36194230	DUP	1375
PT-L1MG	1	2	17	34842808	36194230	DUP	1351
PT-M8JR	1	2	17	34819191	36194230	DUP	1375
PT-OOQ4	1	2	17	34854121	36194230	DUP	1340
PT-OPSR	1	1	17	34819191	36194230	DUP	1375

22q11.2							
GWAS arrays							
FID	**IID**	**PHE**	**CHR**	**BP1**	**BP2**	**TYPE**	**KB**
PT-8UDI	1	1	22	18770421	21611349	DUP	2841
PT-BSGF	1	1	22	18876428	21465835	DUP	2589
PT-ERVX	1	1	22	20733495	21608479	DUP	875
PT-ETKH	1	1	22	20733495	21463730	DUP	730.2
PT-ONUZ	1	1	22	19407597	21463730	DUP	2056
Exome Arrays							
FID	**IID**	**PHE**	**CHR**	**BP1**	**BP2**	**TYPE**	**KB**
PT-8UDI	1	1	22	18918528	20302992	DUP	1384
PT-BSGF	1	1	22	18900755	21408430	DUP	2508
PT-ERVX	1	1	22	20748934	21408430	DUP	659.5
PT-ETKH	1	1	22	20760977	21377650	DUP	616.7
PT-ONUZ	1	1	22	19026403	21408430	DUP	2382
PT-BQOL	1	1	22	18918528	21408430	DUP	2490

Table S14 Novel association regions and replication results

TYPE	CHR	BP1	BP2	Genes	Swedish sample (case:control)	UK sample (case:control)
DEL	5	104269333	104278677		9:02	7:08
DEL	5	104269333	104394274		9:01	7:11
DEL	9	122619031	122833468		6:00	0:00
DUP	7	62358419	62413547		28:16	0:00
DUP	8	87205568	87314758	SLC7A13	25:12	2:09
DUP	10	45237208	45321852		39:25	0:00
DUP	12	100278040	100402187		14:06	1:04
DUP	8	13947372	15095848	SGCZ	5:00	4:04
DUP	16	69796208	69975644	WWP2	5:00	0:00
DUP	22	23800000	24951903	many	6:00	2:04
DUP	17	34800000	36200000	17q12, many genes	5:01	5:02
DUP	22	18700000	21800000	22q11.2, many genes	0:05	0:10

For each novel association region, we applied matching procedures to count the number of CNV events in the UK samples. Specifically, for single-gene loci (SLC7A13, SGCZ, WWP2), we computed the counts of CNV events disrupting the gene (≥1bp overlap). For all other region, we computed the counts of CNV events that overlapped the region by >50% of its length.

Table S15 Overlap between genes affected by common variants and rare CNVs in the shared pathways

Gene set Name	#genes in set	#genes in set overlapped by associated GWAS loci (A)	# genes in set overlapped by 500kb CNVs (B)	# genes in set overlapped by 500kb case CNVs (C)	#genes in set shared between (A) and (B)	#genes in set shared between (A) and (C)	genes in set shared between (A) and (B)	genes in set shared between (A) and (C)
Calcium signaling (hsa04020)	178	19	19	11	1	0	GRIN2A	-
FMRP targets	810	128	31	28	7	7	APP,B3GAT1,IGSF9B,MAGI2,MFHAS1,TNKS,YWHAG	APP,B3GAT1,IGSF9B,MAGI2,MFHAS1,TNKS,YWHAG

1. The association P values from GWAS were reported in Ripke et al (2013). Based on Hapmap 3 imputed Swedish data, we defined linkage disequilibrium (LD) intervals around index SNPs with $P < 10^{-3}$ to include all SNPs with $P < 0.05$ in $R^2 > 0.2$, within 500kb. Conservatively, any interval spanning the MHC region (broadly defined as 25-35Mb, hg19) was removed due to the extensive LD in this region and high gene count. A total of 2121 genomic intervals representing nominally associated GWAS loci were identified and enclosed a total 1791 of genes. We then identified genes overlapped by both gene-sets of interests and the associated GWAS genes, designated as (A).
2. In Table 2 of the main texts, we found that genes affected by >500kb CNVs (deletions and duplications combined) were significantly enriched for genes in Calcium signaling channel in SCZ cases than in controls; and that genes affected by >500kb deletions were significantly enriched for FMRP targets in SCZ cases than in controls. We identified genes overlapped by both gene-sets of interests and 500kb CNVs (or deletions) in both cases and controls, designated as (B), and in cases only, designated as (C).

Table S16 Geneset association results using additional expert curated geneset

Gene Set			CNVs >100Kb						CNVs >500Kb					
Name	**Genes**	**Type**	***#CNVs***	***#Genes***	***OR***	**95% CI**	P_{emp}	***Adj_P***	***#CNVs***	***#Genes***	***OR***	**95% CI**	P_{emp}	***Adj_P***
Autism	102	DEL & DUP	143	16	1.03	(0.738 , 1.43)	0.4117	1	18	5	1.11	(0.441 , 2.81)	0.3714	1
		DEL	107	9	1.14	(0.774 , 1.69)	0.2186	1	10	2	16.3	(0.792 , 335)	0.00069	0.07
		DUP	36	12	0.8	(0.416 , 1.52)	0.7886	1	8	5	0.15	(0.0177 , 1.2)	0.9835	1
Mental Retardation	503	DEL & DUP	311	120	1.19	(0.962 , 1.47)	0.03969	1	91	48	2.17	(1.41 , 3.34)	4.00E-05	**0.0049**
		DEL	64	44	0.9	(0.562 , 1.44)	0.5935	1	15	12	1.86	(0.61 , 5.65)	0.1276	1
		DUP	247	91	1.29	(1.01 , 1.64)	0.0177	1	76	40	2.35	(1.46 , 3.77)	3.00E-05	**0.0038**
Synaptic genes (Ruano et al)	718	DEL & DUP	941	255	1.1	(0.981 , 1.24)	0.04505	1	351	104	1.23	(1.03 , 1.48)	0.00497	0.432
		DEL	279	98	1.35	(1.08 , 1.67)	0.00157	0.15	135	41	1.67	(1.15 , 2.43)	0.00031	**0.036**
		DUP	662	209	1.01	(0.877 , 1.16)	0.4756	1	216	80	1.06	(0.846 , 1.33)	0.2348	1
Synapse Proteome (G2Cdb)	1023	DEL & DUP	708	186	1.23	(1.07 , 1.41)	0.00193	0.19	317	82	1.3	(1.05 , 1.6)	0.00631	0.54
		DEL	196	73	1.18	(0.904 , 1.54)	0.1183	1	104	33	1.42	(0.949 , 2.11)	0.04254	1
		DUP	512	145	1.25	(1.06 , 1.47)	0.00432	0.38	213	59	1.24	(0.961 , 1.61)	0.04298	1
KO mouse behavior (JAX)	2019	DEL & DUP	1415	452	1.07	(0.975 , 1.17)	0.06433	1	411	182	1.22	(1.05 , 1.42)	0.00173	0.17
		DEL	387	167	1.26	(1.06 , 1.49)	0.00206	0.20	152	68	1.53	(1.11 , 2.1)	0.00099	0.10
		DUP	1028	344	0.99	(0.884 , 1.11)	0.5581	1	259	134	1.12	(0.933 , 1.35)	0.0698	1
Cytoplasm (Kirov et al)	266	DEL & DUP	165	58	1.61	(1.19 , 2.17)	0.0014	0.14	42	18	3.11	(1.5 , 6.45)	8.00E-05	**0.0096**
		DEL	72	16	1.42	(0.886 , 2.28)	0.08238	1	8	6	2.43	(0.484 , 12.2)	0.09496	1
		DUP	93	46	1.78	(1.2 , 2.65)	0.0021	0.20	34	13	3.38	(1.49 , 7.68)	1.00E-04	**0.012**
Nucleus (Kirov et al)	143	DEL & DUP	111	21	0.93	(0.643 , 1.35)	0.6485	1	75	5	0.91	(0.57 , 1.46)	0.6453	1
		DEL	36	2	0.83	(0.425 , 1.61)	0.7294	1	32	2	0.77	(0.373 , 1.6)	0.7288	1
		DUP	75	20	1.01	(0.647 , 1.58)	0.4718	1	43	4	0.99	(0.536 , 1.83)	0.524	1
Pre-synaptic (Kirov et al)	421	DEL & DUP	389	92	1.02	(0.851 , 1.23)	0.459	1	191	36	1.15	(0.885 , 1.5)	0.1221	1
		DEL	122	34	1.09	(0.791 , 1.49)	0.3083	1	75	14	1.3	(0.854 , 1.97)	0.08559	1
		DUP	267	76	1	(0.803 , 1.25)	0.5602	1	116	29	1.04	(0.732 , 1.49)	0.3878	1
Pre-synaptic active zone (Kirov et al)	171	DEL & DUP	96	29	1.44	(0.964 , 2.14)	0.04787	1	38	12	2.95	(1.38 , 6.31)	0.00055	0.061
		DEL	20	7	2.06	(0.771 , 5.5)	0.096	1	13	5	8.15	(1.04 , 64)	0.00132	0.133
		DUP	76	26	1.34	(0.865 , 2.07)	0.1001	1	25	8	2.25	(0.961 , 5.28)	0.01827	1
Synaptic vesicle (Kirov et al)	335	DEL & DUP	350	76	1	(0.824 , 1.22)	0.513	1	185	30	1.14	(0.876 , 1.49)	0.1341	1
		DEL	115	31	1.08	(0.782 , 1.49)	0.2985	1	72	11	1.3	(0.85 , 1.97)	0.08293	1
		DUP	235	62	0.97	(0.759 , 1.25)	0.6554	1	113	26	1.03	(0.72 , 1.47)	0.4167	1

DEL: deletions. DUP: duplications. All tests were one-sided assuming enrichment in cases using genic CNVs. #CNV = the number of events that overlapped any gene in the geneset by ≥1bp. #genes = the number of unique genes in the geneset that had at least 1 CNV hit (≥1bp overlap). OR=odds ratio, indicating the increase in risk for schizophrenia correcting for rate and size of genic CNVs and genotyping batch effect (a continuity correction applied if necessary). CI: confidence interval. P_{emp}, empirical P values were obtained in PLINK by 100,000 permutations and permuting phenotype labels within genotyping batches. Adj_P: Holm-Bonferroni multiple-testing adjusted P values considering all 126 tests performed in Table 3 and Table S16.

Table S17 Loci and genes with case CNV hits in major genesets with significant enrichment

Geneset	>500Kb case CNVs		Relationship with known loci		Relationship with genes	
Name	**Type**	**#Total events**	**#total events within known loci (proportion)**	**known SCZ loci (#event within)**	**#genes with case CNV hits**	**Genes with case CNV hits**
Calcium signaling (hsa04020)	DEL & DUP	56	17 (0.30)	15q13.3(8),16p11.2distal (3), 22q11.21(6)	11	ADCY2,ADORA2A,ADORA2B,ATP2A1,CHRNA7,P2RX2,P2RX6,PDE1C,PPP3CA,PRKX,SPECC1L,
FMRP target	DEL	76	36 (0.47)	3q29(1),22q11.21(35)	28	ADAP1,AGAP1,ANAPC1,APP,ARVCF,B3GAT1,DGCR2,DLGAP1,DTNA,FAM115A,IGSF9B,INTS1,KIAA0226,KIAA0430,KLHL22,LPIN2,MAGI2,MFHAS1,NAV3,PI4KA,PKP4,PPP3CA,RTN4R,SEC23A,SEPT5,SYT1,TNKS,YWHAG,
PSD/mGluR5	DEL & DUP	20	16 (0.80)	16p11.2(10),22q11.21(6)	3	ALDOA,SEPT5,YWHAG,
PSD/NMDAR	DEL & DUP	24	17 (0.71)	3q29(7),16p11.2(10)	8	DLG1,DLGAP1,MAPK3,PPP3CA,RAB6A,STX1A,SYT1,YWHAG,
PSD/PSD-95	DEL	18	12 (0.67)	3q29(6)22q11.21(6)	5	DLG1,DLGAP1,PPP3CA,SEPT5,SYT1,
Mitochondrion (Kirov et al)	DEL	35	20 (0.57)	3q29(6),16p11.2distal(3), 22q11.21(11)	13	ACAD8,AMACR,ATP5J,BDH1,GOT2,HSD17B4,KIAA0564,MDH2,MLYCD,SLC25A1,TUFM,TXNRD2,UQCRC2,
Mitocarta	DEL	87	47 (0.54)	1q21.1(5),3q29(6)16p11.2distal(3),22q11.21(33)	31	ACAD8,ACP6,AGXT2,AIFM3,AMACR,ATP5J,BDH1,COMT,COX10,CYB5B,GOT2,HSD17B4,HSDL1,LYRM2,MDH2,MIPEP,MLYCD,MRPL39,MRPL40,MSRA,MTRF1,NT5C3,PRODH,PXMP2,RG9MTD1,RNMTL1,SLC25A1,TOMM70A,TUFM,TXNRD2,UQCRC2,
Cytoplasm	DEL & DUP	34	13 (0.38)	16p11.2(10),16p11.2distal(3)	13	ANXA5,CA2,CA3,DDT,EIF4H,MSRA,MVP,NAT1,PGM1,SULT1A1,TARS,UPB1,YWHAG,
Cytoplasm	DUP	28	10 (0.36)	16p11.2(10)	9	ANXA5,CA2,CA3,DDT,EIF4H,MVP,NAT1,PGM1,UPB1,
Synaptic genes (Ruano et al)	DEL	100	45 (0.45)	3q29(6),15q13.3(7),16p11.2distal(3),22q11.21(29)	34	ALCAM,APP,CDH13,CDH8,CHRNA7,CNTN4,CNTN6,CYFIP1,DLG1,DLGAP1,DOC2B,HSD17B4,HTR1B,LIN7A,NLGN4X,OPCML,P2RX6,PI4KA,PKP4,POR,PPFIA2,PPFIBP1,PPP3CA,PRKAR1B,RAB3C,RABEP2,RAP1GDS1,RPH3AL,SACS,SEPT5,SLC25A1,SNAP29,SYT1,YWHAG,
Mental Retardation	DEL & DUP	67	22 (0.33)	16p11.2(20),16p11.2distal(2)	36	ABAT,ACBD6,ADRA2B,ALDOA,ALG6,ASAH1,ASS1,BBS7,CA2,CA8,CCNA2,CDK8,CHD2,CLCNKB,CLN3,COL18A1,COX10,CRBN,ERCC6,ERCC8,LIMK1,LINS,MLYCD,MOGS,MTRR,NBN,NLGN4X,PEX10,PHACTR1,PHIP,PRRT2,PSMA7,SHH,SMARCB1,TGIF1,TUSC3,
Mental Retardation	DUP	56	20 (0.36)	16p11.2(20)	28	ABAT,ALDOA,ALG6,ASAH1,ASS1,BBS7,CA2,CA8,CCNA2,CDK8,CHD2,CLCNKB,COL18A1,CRBN,ERCC6,LIMK1,LINS,MOGS,MTRR,NBN,NLGN4X,PEX10,PHACTR1,PRRT2,PSMA7,SHH,SMARCB1,TUSC3,

Only genesets with adj_P<0.05 in Table 3 and Table S16 and with #genes < 40 are shown. Only case CNVs are concerned in Table S17. #Total >500kb case CNV events = the total number of SCZ case events that overlapped any gene in the geneset by ≥1bp. #Events within: A case CNV event is considered to be within a known locus (Table S7) if >50% of its length overlapped by the known locus associated with SCZ. #genes = the number of unique genes in the geneset that had at least 1 case CNV hit (≥1bp overlap).

Table S18 Logistic regression with rare CNV burden and SNP burden: Odds Ratio

		SNP			CNV		
CNV type	**Frequency**	**Odds Ratio (per unit of RPS)**	**95% CI**	**P-value**	**Odds Ratio (per CNV)**	**95% CI**	**P-value**
Deletions and duplications	All	1.093	(1.086,1.101)	<1E-20	1.07	(1.03,1.11)	0.0009
	1x	1.093	(1.086,1.101)	<1E-20	1.31	(1.14,1.49)	0.0001
	2–6x	1.093	(1.086,1.101)	<1E-20	1.04	(0.995,1.09)	0.0854
	≥7x	1.093	(1.086,1.101)	<1E-20	1.04	(0.98,1.1)	0.212
	100-200kb	1.093	(1.086,1.101)	<1E-20	1.03	(0.976,1.09)	0.327
	200-500kb	1.093	(1.086,1.101)	<1E-20	1.07	(0.998,1.15)	0.06
	500kb+	1.093	(1.086,1.101)	<1E-20	1.26	(1.13,1.42)	5.13E-05
Deletions only	All	1.093	(1.086,1.101)	<1E-20	1.09	(1.02,1.16)	0.0117
	1x	1.093	(1.086,1.101)	<1E-20	1.35	(1.13,1.62)	0.0013
	2–6x	1.093	(1.086,1.101)	<1E-20	1.02	(0.942,1.1)	0.674
	≥7x	1.093	(1.086,1.101)	<1E-20	0.96	(0.871,1.06)	0.408
	100-200kb	1.093	(1.086,1.101)	<1E-20	1.06	(0.963,1.14)	0.234
	200-500kb	1.093	(1.086,1.101)	<1E-20	1.06	(0.926,1.19)	0.329
	500kb+	1.093	(1.086,1.101)	<1E-20	1.45	(1.22,1.81)	0.0002
Duplications only	All	1.093	(1.086,1.101)	<1E-20	1.06	(1.01,1.11)	0.0245
	1x	1.093	(1.086,1.101)	<1E-20	1.13	(0.973,1.3)	0.112
	2–6x	1.093	(1.086,1.101)	<1E-20	1.05	(0.995,1.11)	0.075
	≥7x	1.093	(1.086,1.101)	<1E-20	1.07	(0.992,1.15)	0.083
	100-200kb	1.093	(1.086,1.101)	<1E-20	1.01	(0.949,1.1)	0.791
	200-500kb	1.093	(1.086,1.101)	<1E-20	1.07	(0.993,1.18)	0.104
	500kb+	1.093	(1.086,1.101)	<1E-20	1.17	(1.03,1.34)	0.019

SNP burden is based on risk profile scores (RPS) and CNV burden is based on the number of CNVs for each frequency and size category. For each test (corresponding to a row in the table), we fit a multiple logistic regression model: logit(Prb(case)) ~ RPS burden + CNV burden + genotyping batch. The two types of genetic burden (rare CNV and common SNP) are independent and combined in an additive model (interaction term not significant). For CNV, OR measures increase of disease likelihood per CNV. For SNP, OR measures increase of disease likelihood per unit of RPS.

Table S19 Proportion of variance explained by RPS burden and burden of known SCZ-associated CNVs

Locus	Type	Frequency (case:control)	% Variance by RPS burden	% Variance by CNV burden
1q21.1	Deletion	5:1	8.752	0.076
3q29	Deletion	6:0	8.752	0.132
15q13.3	Deletion	7:2	8.752	0.041
22q11.2	Deletion	6:0	8.752	0.200
16p11.2	Duplication	10:2	8.752	0.142
All of above		34:5	8.752	0.526

CNVs events were identified the same way as the regional tests performed in *Table 2* of the main texts.

For each test (corresponding to a row in the table), we fit the following logistic regression models.

(1) logit(Pr(case)) ~ RPS burden + CNV burden + genotyping batch.
(2) logit(Pr(case)) ~ RPS burden + genotyping batch.
(3) logit(Pr(case)) ~ genotyping batch.

To estimate the proportion of variance of case-control status accounted for by RPS and CNV burden, we computed the difference in the Nagelkerke pseudo R^2 score contrasting a full model with a reduced model. For RPS, the pseudo R^2 contrast model (2) with (3) and their values are listed under column (% Variance by RPS burden). For CNV burden, the pseudo R^2 contrast models (1) with (2) and their values are listed under column (% Variance by CNV burden).

Table S20 Proportion of variance explained by RPS burden and rare CNV burden

CNV Type	Class	% Variance by RPS burden	% Variance by CNV burden
Deletions & Duplications	All	8.752	0.143
	1x	8.752	0.191
	2-6x	8.752	0.038
	7+x	8.752	0.020
	100-500kb	8.752	0.052
	>500kb	8.752	0.221
Deletions	All	8.752	0.082
	1x	8.752	0.134
	2-6x	8.752	0.002
	7+x	8.752	0.009
	100-500kb	8.752	0.022
	>500kb	8.752	0.194
Duplications	All	8.752	0.066
	1x	8.752	0.033
	2-6x	8.752	0.041
	7+x	8.752	0.039
	100-500kb	8.752	0.030
	>500kb	8.752	0.071

In *Table S20*, estimates of the proportion of variance were obtained the same way as in *Table S19*. CNV burden was measured by the number of events stratified by CNV type, size, and frequency. RPS accounted for at least an order of magnitude more variance than rare CNVs in this sample. Similar results were obtained when using gene count and total length as burden metrics (data not shown).

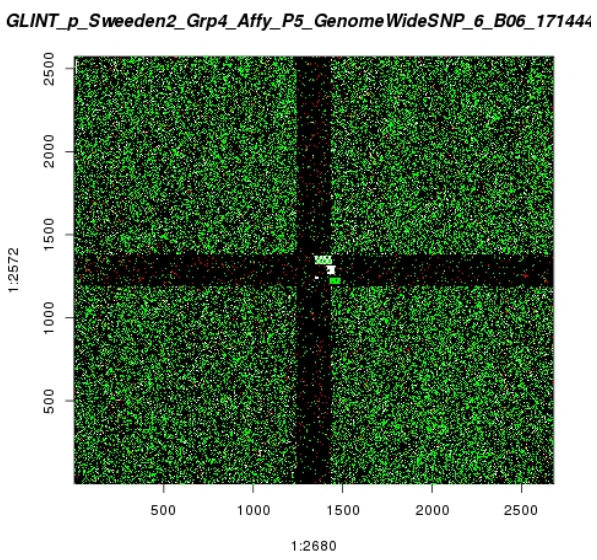
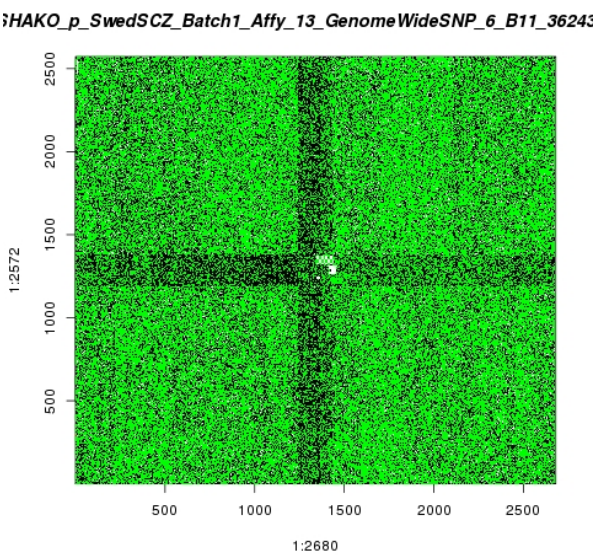
Table S21 CNV burden in individuals with or without BLM mutations.

Burden	Deletions and duplications				Deletions only				Duplications only			
Metrics	Without	With	Beta	P	Without	With	Beta	P	Without	With	Beta	P
Total number	1.07	1.17	0.089	0.259	0.41	0.52	0.114	0.081	0.66	0.65	-0.025	0.588
Total KB	332.5	380.86	42.3	0.26	117.97	164.74	45.8	0.122	214.53	216.12	-3.49	0.526
Gene count	2.5	3.44	0.89	0.0789	**0.72**	**1.7**	**0.96**	**0.0019**	1.78	1.75	-0.074	0.555

Among the 4,500 subsamples with exome sequencing, a total of 63 individuals had at least one disruptive exonic mutation in *BLM*, and the remaining 4,437 individuals do not harbor *BLM* mutations. A total of 52 deletions and 41 duplications were identified in individuals with *BLM* mutations. Burden metrics: the number of CNVs (total number), the genomic length impacted by CNVs (total KB), and the number of genes impacted by CNVs (gene count). Beta: regression coefficients representing the mean change in CNV burden for a *BLM* mutation while accounting for batch effect. P: P values associated with beta, one-sided assuming higher burden in individuals with *BLM* mutations. Red font: P<0.005 (the multiple-testing-adjusted cutoff based on Bonferroni method).

Supplemental Figures

Figure S1. Example Affymetrix 6.0 array images

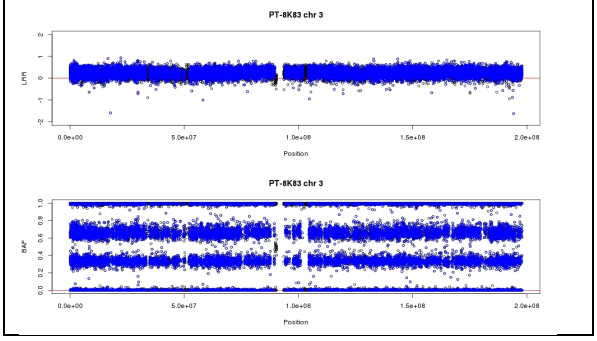
(a) Acceptable array: MAPD=0.2, Waviness.SD=0.07	(b) Problematic array: MAPD=0.51, Waviness.SD=0.11
	

For Affymetrix 6.0, copy number probes are located within the middle cross and SNP probes are located in the 4 quadrants. Each chip has 2680 cols x 2572 rows, or 1,856,069 units. Each unit has > 1 million copies of a 25 bp probe. The images were produced using R scripts where the color scale is per standard gene expression color schemes (ranging from green=low-intensity to black=medium intensity to red=high intensity). These images can be quickly inspected for large problems such as spots, bubbles, scratches, and gradients (as in ***Figure S1b***).

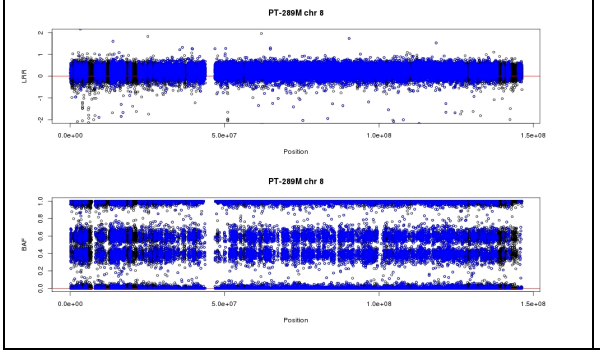
Figure S2 Intensity plots of genomic outliers

The title of each sub-figure indicates sample ID and genotyping platform. The x-axis indicates genomic position of the probes and y-axis indicates the values of LRR (top) or BAF (bottom). red dots indicate probes predicted to be involved in a deletion, and blue dots probes predicted to be involved in a duplication.

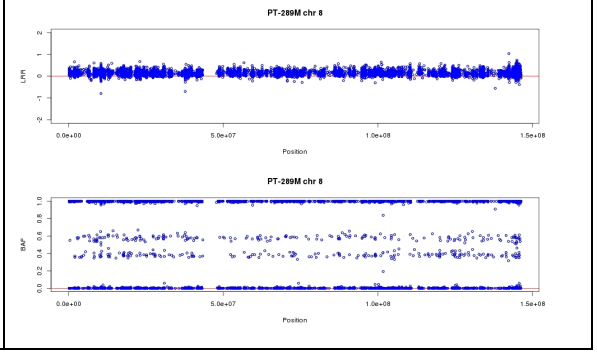
(S2a) PT-8K83: ILMN Omni Express



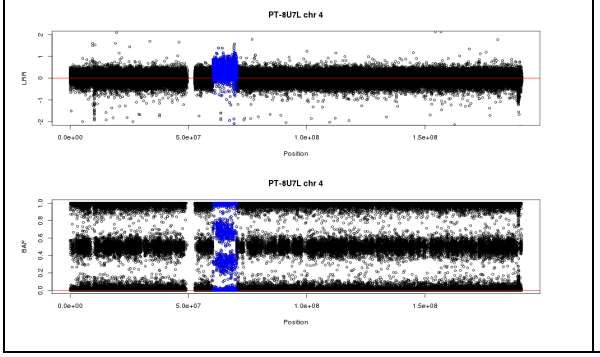
(S2b) PT-289M: Affy SNP 6.0



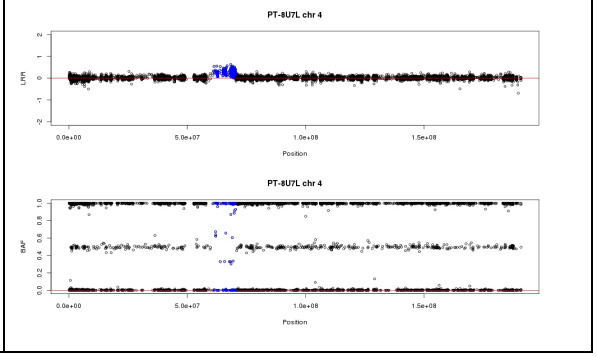
PT-289M: Illumina exome array

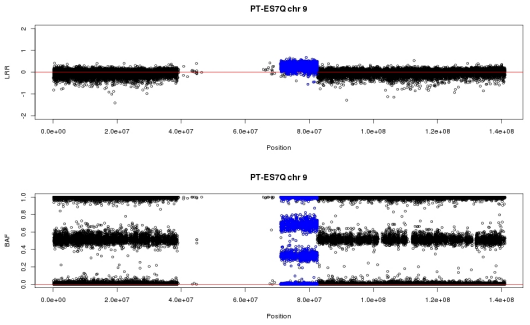
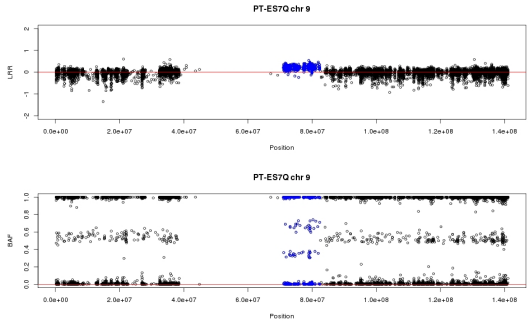


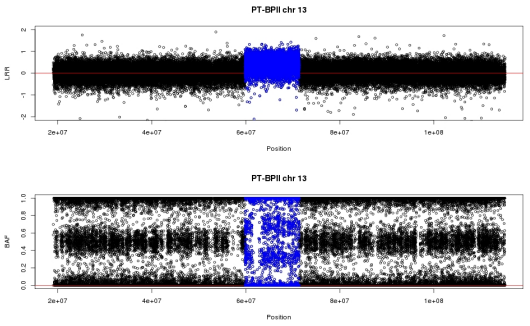
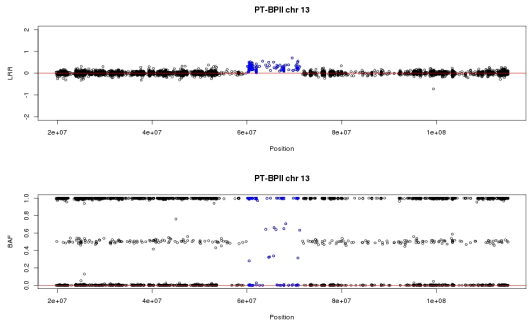
(S2c) PT-8U7L: Affy SNP 6.0

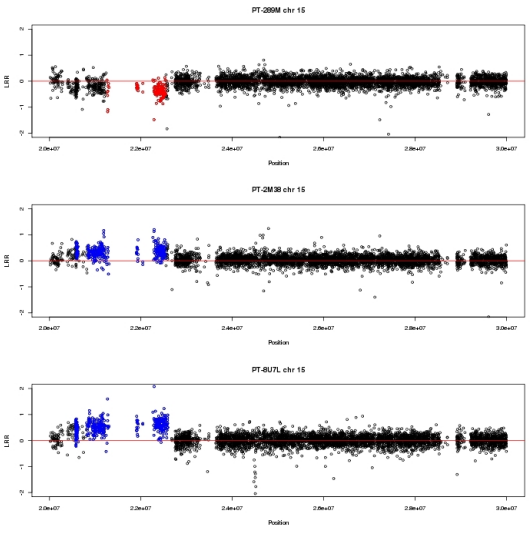


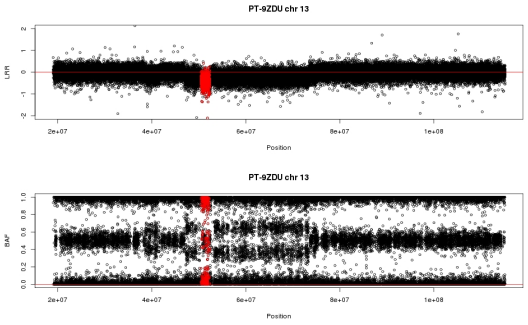
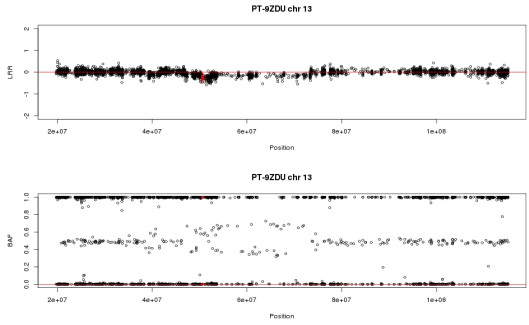
PT-8U7L: Illumina exome array



(S2d) PT-ES7Q: Illumina Omni Express	PT-ES7Q: Illumina exome array
	

(S2e) PT-BPII: Affy SNP 6.0	PT-BPII: Illumina exome array
	

(S2f) Chr15q11.2 deletion and duplications


(S2g) PT-9ZDU: Affy6.0	PT-9ZDU: Illumina exome array
	

For PT-9ZDU, the BAF suggests two long regions over which the allelic ratios are skewed (one is at 38-42 Mb, the other at 44-72Mb). This might normally suggest a duplication, but the LRR suggests that the copy number is actually reduced. These observations suggest deletions that are mosaic within the individual's sample -- i.e. that are present in many but not all of the cells that were sampled.

Figure S3 Intensity plots of example duplication at 17q12 from GWAS and exome arrays

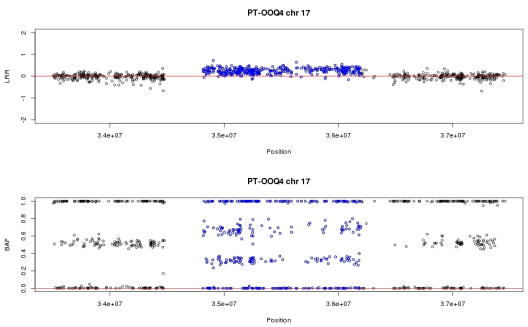
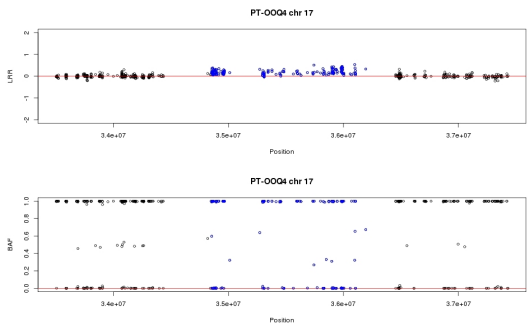
PT-OOQ4, GWAS array	PT-OOQ4, exome array
	

Figure S4 Intensity plots of example duplication at 22q11 from GWAS and exome arrays

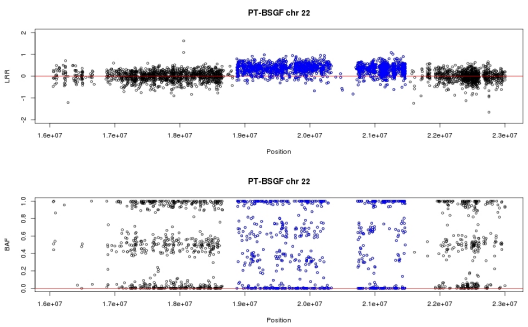
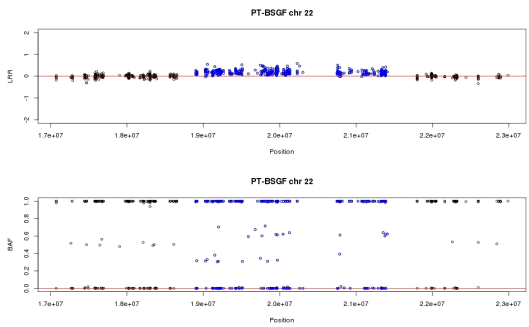
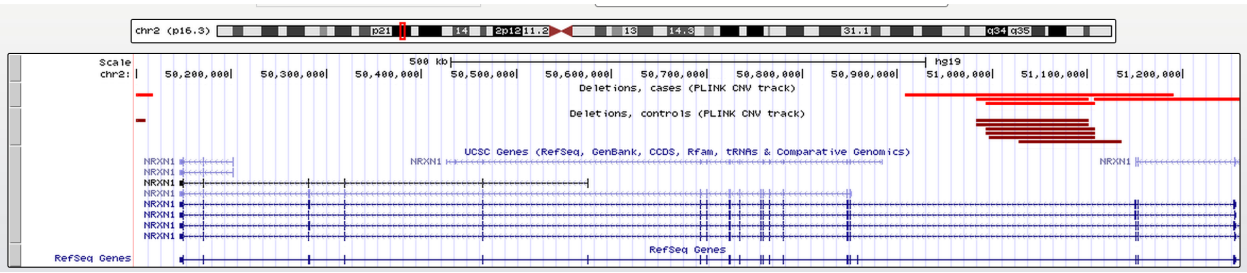
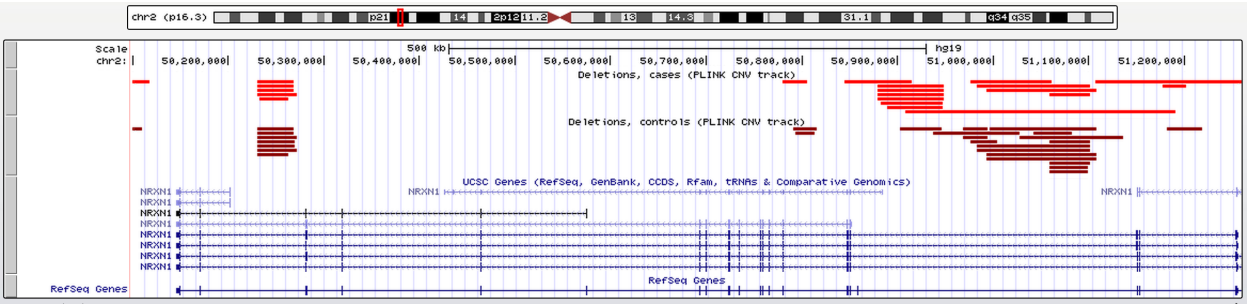
PT-BSGF, GWAS array	PT-BSGF, exome array
	

Figure S5 *NRXN1* deletions in >100Kb and >20Kb CNV datasets.

(S5a) Results using the >100Kb CNV dataset

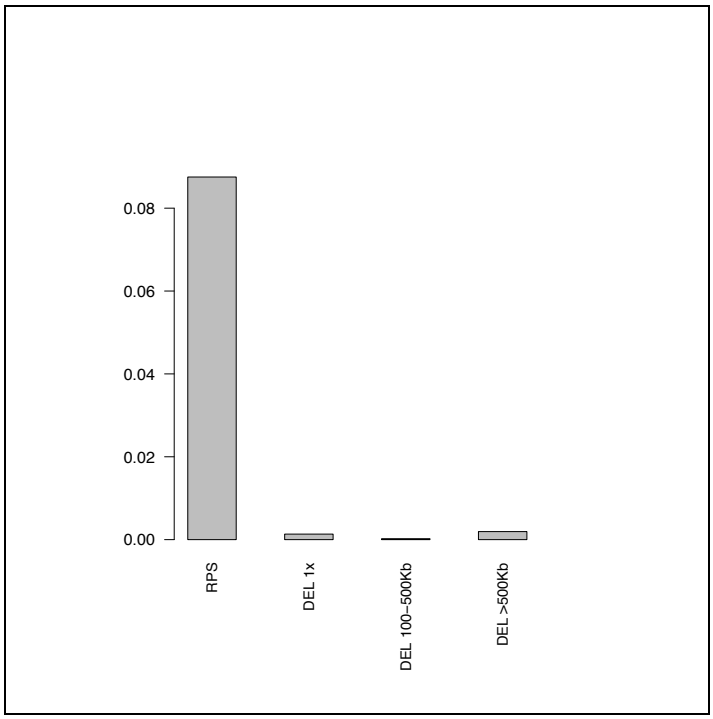


(S5b) Results using the >20Kb CNV dataset



We *Figure S5a*, we report the results using the >100Kb CNV dataset. We observed 4 deletions in cases and 6 deletions in controls when all >100kb deletions were considered. We observed two deletions in SCZ cases disrupting one exon of NRXN1 but no exonic deletion was found in controls. In *Figure S5b,* we report the results using the >20Kb CNV dataset, where we found 20 deletions in cases and 26 deletions in controls when all >20kb deletions were considered. We observed four deletions in SCZ cases disrupting NRXN1 exons but no exonic deletion was found in controls.

Figure S6 Relative impact of rare CNV burden and common variant allelic burden



We computed the difference in the Nagelkerke pseudo R^2 score to estimate the proportion of variance of case-control status in the Swedish samples accounted for by the common variant allelic burden (risk profile scores, RPS) and by the rare CNV burden (as measured by the number of CNV for all >100kb CNVs stratified by type, size, and frequency). The Y-axis of the barplot shows the estimates of effect size (i.e. Nagelkerke pseudo R^2). This barplot shows selected CNV class to illustrate that RPS accounted for at least an order of magnitude more variance than rare CNVs in this sample. Complete results are shown in ***Table S20*** of this supplement.

Figure S7: Prevalence and recurrence risks in comparison to the literature

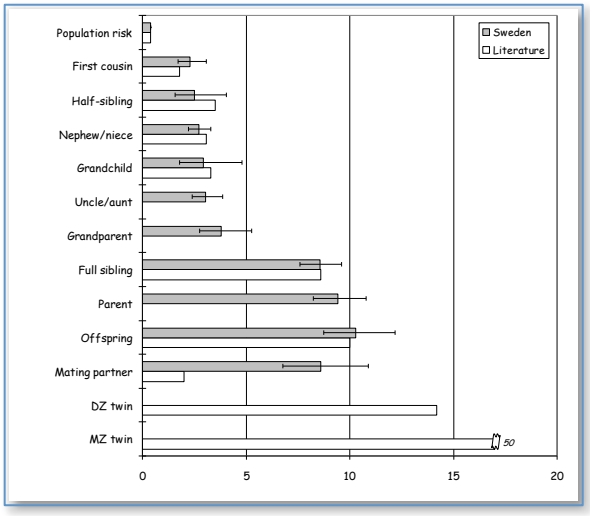
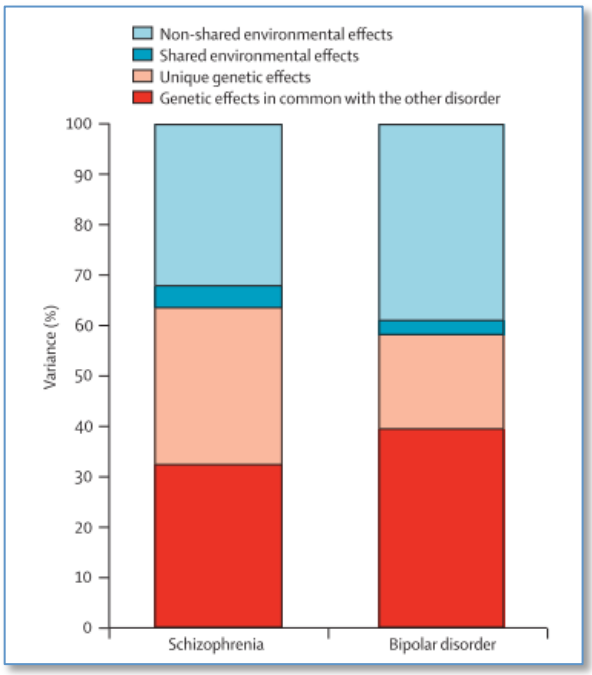


Figure S8: Heritability of schizophrenia and bipolar disorder in Sweden



References

1. Kristjansson, E., Allebeck, P. & Wistedt, B. Validity of the diagnosis of schizophrenia in a psychiatric inpatient register. *Nordisk Psykiatrik Tidsskrift* **41**, 229-34 (1987).

2. Dalman, C., Broms, J., Cullberg, J. & Allebeck, P. Young cases of schizophrenia identified in a national inpatient register--are the diagnoses valid? *Social Psychiatry and Psychiatric Epidemiology* **37**, 527-31 (2002).

3. World Health Organization. *International Classification of Diseases*, (World Health Organization, Geneva, 1967).

4. World Health Organization. *International Classification of Diseases*, (World Health Organization, Geneva, 1978).

5. World Health Organization. *International Classification of Diseases*, (World Health Organization, Geneva, 1992).

6. Hultman, C.M., Sparen, P., Takei, N., Murray, R.M. & Cnattingius, S. Prenatal and perinatal risk factors for schizophrenia, affective psychosis, and reactive psychosis of early onset: case-control study. *Bmj* **318**, 421-6 (1999).

7. Zammit, S. *et al.* Investigating the association between cigarette smoking and schizophrenia in a cohort study. *Am J Psychiatry* **160**, 2216-21 (2003).

8. Andersson, R.E., Olaison, G., Tysk, C. & Ekbom, A. Appendectomy and protection against ulcerative colitis. *N Engl J Med* **344**, 808-14 (2001).

9. Hansson, L.E. *et al.* The risk of stomach cancer in patients with gastric or duodenal ulcer disease. *N Engl J Med* **335**, 242-9 (1996).

10. Lichtenstein, P. *et al.* Recurrence risks for schizophrenia in a Swedish national cohort. *Psychological Medicine* **36**, 1417-26 (2006).

11. Sullivan, P.F., Daly, M.J. & O'Donovan, M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Reviews Genetics* **13**, 537-51 (2012).

12. Ekholm, B. *et al.* Evaluation of diagnostic procedures in Swedish patients with schizophrenia and related psychoses. *Nordic Journal of Psychiatry* **59**, 457-64 (2005).

13. Saha, S., Chant, D., Welham, J. & McGrath, J. A systematic review of the prevalence of schizophrenia. *PLoS Medicine* **2**, e141 (2005).

14. Statistics Sweden. Multi-Generation Register 2002: A description of contents and quality. (Statistics Sweden,, Örebro, Sweden, 2003).

15. Gottesman, I.I. *Schizophrenia Genesis: The Origins of Madness*, (New York, WH Freeman,, 1991).

16. Sullivan, P.F., Kendler, K.S. & Neale, M.C. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Archives of General Psychiatry* **60**, 1187-92 (2003).

17. International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-52 (2009).

18. Schizophrenia Psychiatric Genome-Wide Association Study Consortium. Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics* **43**, 969-76 (2011).

19. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237-41 (2008).

20. Purcell, S. *et al.* PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* **81**, 559-75 (2007).

21. Goldstein, J.I. *et al.* zCall: a rare variant caller for array-based genotyping. *Bioinformatics* **28**, 2543-5 (2012).

22. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**, 1665-74 (2007).

23. Szatkiewicz, J.P. *et al.* Detecting large copy number variants using exome genotyping arrays. *Molecular Psychiatry* (In press).

24. Hamshere, M.L. *et al.* Genome-wide significant associations in schizophrenia to ITIH3/4, CACNA1C and SDCCAG8, and extensive replication of associations reported by the Schizophrenia PGC. *Molecular psychiatry* (2012).

25. Davies, G., Welham, J., Chant, D., Torrey, E.F. & McGrath, J. A systematic review and meta-analysis of Northern Hemisphere season of birth studies in schizophrenia. *Schizophr Bull* **29**, 587-93 (2003).

26. Torrey, E.F., Rawlings, R.R., Ennis, J.M., Merrill, D.D. & Flores, D.S. Birth seasonality in bipolar disorder, schizophrenia, schizoaffective disorder and stillbirths. *Schizophr Res* **21**, 141-9 (1996).

27. Mortensen, P.B. *et al.* Effects of family history and place and season of birth on the risk of schizophrenia. *N Engl J Med* **340**, 603-8 (1999).

28. Kelly, B.D. *et al.* Schizophrenia and the city: A review of literature and prospective study of psychosis and urbanicity in Ireland. *Schizophr Res* **116**, 75-89 (2010).

29. Miller, B. *et al.* Meta-analysis of Paternal Age and Schizophrenia Risk in Male Versus Female Offspring. *Schizophr Bull* (2010).

30. Torrey, E.F. *et al.* Paternal age as a risk factor for schizophrenia: how important is it? *Schizophr Res* **114**, 1-5 (2009).

31. Cantor-Graae, E. The contribution of social factors to the development of schizophrenia: a review of recent findings. *Can J Psychiatry* **52**, 277-86 (2007).

32. Harrison, G. *et al.* Association between psychotic disorder and urban place of birth is not mediated by obstetric complications or childhood socio-economic position: a cohort study. *Psychol Med* **33**, 723-31 (2003).

33. Sipos, A. *et al.* Paternal age and schizophrenia: a population based cohort study. *Bmj* **329**, 1070 (2004).

34. Wicks, S., Hjern, A. & Dalman, C. Social risk or genetic liability for psychosis? A study of children born in Sweden and reared by adoptive parents. *Am J Psychiatry* **167**, 1240-6 (2010).

35. Cannon, M., Jones, P.B. & Murray, R.M. Obstetric complications and schizophrenia: Historical and meta-analytic review. *American Journal of Psychiatry* **159**, 1080-92 (2002).

36. Geddes, J.R. & Lawrie, S.M. Obstetric complications and schizophrenia: a meta-analysis. *Br J Psychiatry* **167**, 786-93 (1995).

37. Abel, K.M. *et al.* Birth weight, schizophrenia, and adult mental disorder: is risk confined to the smallest babies? *Arch Gen Psychiatry* **67**, 923-30 (2010).

38. Dalman, C., Allebeck, P., Cullberg, J., Grunewald, C. & Koster, M. Obstetric complications and the risk of schizophrenia: a longitudinal study of a national birth cohort. *Arch Gen Psychiatry* **56**, 234-40 (1999).

39. Aylward, E., Walker, E. & Bettes, B. Intelligence in schizophrenia: meta-analysis of the research. *Schizophr Bull* **10**, 430-59 (1984).

40. Woodberry, K.A., Giuliano, A.J. & Seidman, L.J. Premorbid IQ in schizophrenia: a meta-analytic review. *Am J Psychiatry* **165**, 579-87 (2008).

41. Ruderfer, D.M. *et al.* Mosaic copy number variation in schizophrenia. *European journal of human genetics : EJHG* (2013).