Correction: 17 April 2008

www.sciencemag.org/cgi/content/full/1155174/DC1

# Supporting Online Material for

## Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia

Tom Walsh, Jon M. McClellan,* Shane E. McCarthy, Anjené M. Addington, Sarah B. Pierce, Greg M. Cooper, Alex S. Nord, Mary Kusenda, Dheeraj Malhotra, Abhishek Bhandari, Sunday M. Stray, Caitlin F. Rippey, Patricia Roccanova, Vlad Makarov, B. Lakshmi, Robert L. Findling, Linmarie Sikich, Thomas Stromberg, Barry Merriman, Nitin Gogtay, Philip Butler, Kristen Eckstrand, Laila Noory, Peter Gochman, Robert Long, Zugen Chen, Sean Davis, Carl Baker, Evan E. Eichler, Paul S. Meltzer, Stanley F. Nelson, Andrew B. Singleton, Ming K. Lee, Judith L. Rapoport, Mary-Claire King, Jonathan Sebat

*To whom correspondence should be addressed. E-mail: drjack@u.washington.edu

Published 27 March 2008 on *Science* Express

DOI: 10.1126/science.1155174

**This PDF file includes:**

Materials and Methods
Figs. S1 to S3
Tables S1 to S4
References

**Correction (17 April 2008):** On page 7, an additional statement was provided regarding the Database of Genomic Variants. Also, a new reference (S17) was added, and selected references and their respective callouts were renumbered accordingly. The *x*-axis label of fig. S2 was also corrected. Finally, the name of the 10th author (Abhishek Bhandari) was originally misspelled and has been corrected.

# SUPPORTING ONLINE MATERIAL: MATERIALS AND METHODS

## SUBJECTS

Cases in the original series were individuals with schizophrenia or schizoaffective disorder meeting DSM-IV criteria. Institutional approval was obtained at each participating site. Of 150 subjects with schizophrenia, 120 were adults recruited as inpatients from Western State Hospital in Lakewood, Washington, the largest public psychiatric hospital in Washington State. Referrals were made by treating clinicians based on a primary diagnosis of schizophrenia, irrespective of syndromic features or cognitive delays. Individuals known to carry the VCFS deletion on chromosome 22q11 were excluded. There was no exclusion for cognitive deficits, so long as patients could provide fully informed consent. Many patients had a history of forensic involvement, either deemed not guilty by reason of insanity through the criminal court system or referred from correctional institutions or law enforcement for mental health care. Any pending criminal charges had to be resolved before patients could provide informed consent to be recruited to the study.

We also recruited 30 youth with early onset schizophrenia spectrum disorders. These patients were age 19 or younger at time of recruitment and were participants in an NIMH clinical trial, Treatment of Early Onset Schizophrenia and Schizoaffective Disorder (TEOSS) (S1) or were under treatment at the University of Washington-affiliated psychiatry programs at Children’s Hospital and Regional Medical Center, Seattle. Most participating youth were outpatients. Subject consent or assent and parent or guardian assent for minors was obtained for each participant. Youth with an estimated premorbid IQ<65 were excluded.

For all cases, diagnostic status was confirmed with medical records and diagnostic interviews. The Structured Clinical Interview for DSM-IV (SCID) (S2) was used for subjects age >18 years, and the KID-SCID (S3) was administered to subjects ≤18 years. Extensive medical records were available for the adult patients, but family members were not available. For youth, information was obtained from medical records review and from interviews of parents or guardians (S1).

By self-description, the case cohort (adults and children combined) was 78% Caucasian, 18% African-American, and 4% Asian, Pacific Islander, or American Indian. Multiple individuals reported mixed ancestries. The case cohort was 74% male. Clinical information is provided in Table S1.

The adult patients were severely and chronically ill. The median length of hospitalization was 1.8 years, with some patients having been hospitalized for decades. This in part represents long-term hospitalization of forensic patients. Age of onset was defined as overt evidence of psychotic symptoms associated with functional impairment; that is hallucinations, delusions, thought disorder and bizarre disorganized behavior per DSM-IV. Because few adult patients had undergone standardized intelligence testing, the estimated number of patients with IQ<80 reflects diagnoses of suspected borderline or mild mental retardation in medical records. Most youth had a neurocognitive assessment performed upon enrollment in TEOSS (S1), although a few left treatment before intellectual testing could be completed. Information on family history of mental illness was drawn from interviews with the patient, from interviews with parents or guardians, and from medical records. For adult patients, family history data is almost entirely drawn solely from patient report and medical records. Relatives were not systematically interviewed with respect to their own mental health status. Structural variants involving genes appeared more frequently among patients with IQ<80 (7/15 patients) than among patients with IQ≥80 (15/135 patients), a significant association (P=0.002 by 2-tailed Fisher's Exact Test). However, IQ data are quite limited, and the association between structural variants and lower IQ needs to be interpreted with caution. There was no association between detection of a structural variant in a patient and gender, diagnosis, forensic involvement, or family history of psychosis.

From each case, blood was drawn into two 10ml ACD vacutainer tubes. DNA was extracted directly from one tube by a standard salting out procedure. The second tube was used for EBV immortalization of lymphoblasts in the King lab.

Controls for subjects of Caucasian ancestry were participants in the NINDS Neurogenetics Repository (S4). We used repository panels NDPT002, NDPT006, and NDPT009, with DNA samples from 278 individuals age 55 years or older. DNA had been extracted previously at Coriell Repository from immortalized lymphoblast cell lines. Absence of neurological symptoms and of relevant family history had been previously assessed by interview. Controls for subjects of African-American ancestry were unaffected individuals, age 35 and older, who participated in genomic analysis projects at Cold Spring Harbor Laboratory (S5, S6) and who consented to the use of their DNA samples anonymously but with ancestry, age, and health status retained. Inclusion of some controls was through the kind intervention of Peter Gregersen, Annette Lee, and the Academic Medical Development Company (AMDeC) (S7) and the expert technical assistance of Lisa Hufnagel and Kevin Pavon. Because frequencies of structural variants differ across populations (S8), we matched controls and cases to obtain the same distributions of ancestries in each series. To select appropriate numbers of African American and Caucasian controls, we stratified all control samples by population, assigned a random number to each sample, ranked the random numbers, and selected the appropriate number of samples from each population to obtain 49/268 (18%) African-American controls and 219/268 (82%) Caucasian controls.

Cases with childhood onset schizophrenia (COS) were recruited nationwide and assessed as previously described (S9). To summarize briefly, all 92 patients met DSM-IIIR/DSM-IV criteria for schizophrenia or psychosis not other specified (NOS), had premorbid full-scale IQ scores of 70 or above and onset of psychotic symptoms by age 12 years (S10,S11,S12,S13). As described in the text, nine COS patients with previously identified chromosomal abnormalities were excluded from this analysis, leaving 83 COS cases in this study. Also as described in the text, because the ancestries of these patients were highly heterogeneous, we evaluated the non-transmitted chromosomes of the 154 available parents as controls. Parental relationships were validated using multiple polymorphic markers. Because each parent contributed one haploid non-transmitted genome to analysis, the effective diploid sample size of our control group was 154/2 or 77.

## MUTATION DISCOVERY AND VALIDATION IN THE ORIGINAL SERIES

### Mutation discovery by ROMA

Genome scans for structural variants (SV) were performed using ROMA as previously described (S14,S15). Scans involved two-color assays performed by co-hybridizing each sample to an oligonucleotide array, using a standard reference genome for comparison. Assays were performed in duplicate with dye-swap. The array consisted of 85,000 probes, providing a mean resolution of one probe every 35kb. Log intensity ratios from duplicate scans were averaged, and normalized ratio data was segmented by a Hidden Markov Model to define copy number variants relative to the reference (S14). The normalized ratio for diploid genomes was set at 1.0, so that heterozygous deletion would be represented as intensity of 0.5 and heterozygous duplication as intensity of 1.5. Case and control samples were screened on the same batches of arrays, in the same lab, by technicians who were unaware whether a particular experiment included cases, controls, or both.

We followed a multi-step process to ensure that cases and controls were evaluated identically and that events detected were real. Log intensity ratios from all 418 samples were filtered to retain only events meeting four criteria: median intensity of probes in the HMM-defined segment <0.8 (deletion) or >1.2 (duplication); HMM-derived likelihood measure (LHM) (S14) >0.95; region defined by at least three adjacent probes (to preclude artifacts due to RFLPs in binding sites); and event at least 100kb in size. These criteria are conservative and no doubt led to exclusion of some true deletions and duplications. Our primary concern was to minimize false positive events in cases and in controls, even at the cost of missing events in both groups.

In order to interpret differences in CNV frequencies between cases and controls, it was crucial that ascertainment of CNVs be equivalent for cases and controls. Specifically, false negative rates should be the same, and there should be no false positives. We employed two methods to assess the ascertainment of CNVs in cases and controls. First, we determined the frequencies of all ROMA-detected CNVs in cases and controls. We detected 115 different CNVs of size greater than 100kb in the set of 418 individuals, an average of six events per individual in cases and in controls. There were no significant differences in frequencies of common CNVs between cases and controls (Figure S1). Second, we introduced simulated deletions into the ROMA reference sample and examined the sensitivity of detection in the experimental samples, as follows. A coefficient of hemizygosity for each hybridization was calculated based on the average intensity difference between autosomal probes and the X chromosome from our male reference genome. Deletions ranging in size from 1-100 probes (10 replicates of each size class) were introduced into the data at 170 randomly selected sites, excluding sites where a CNV had been detected in the same individual. The modified intensity data was then reanalyzed using the HMM algorithm, and the proportion of events detected were recorded for each size class of CNV. Similar relationships between CNV size and likelihood of detection for cases and for controls reflect similar sensitivity of CNV detection for the two groups of samples (Figure S2). Similar CNV frequencies in cases and controls and similar detection sensitivity for simulated CNVs in cases and controls suggest that ROMA was equally sensitive in detecting true CNVs in cases and controls. The validation steps described in the next sections were carried out to eliminate false positive events.

### Validation of rare variants by Illumina and NimbleGen HD2 arrays

We validated potential events of interest using Illumina and NimbleGen arrays. The focus of our interest was individually rare duplications and deletions. We defined rare events as those with median intensity <0.80 or ≥1.20, LHM>0.95, at least 100kb in size and not previously described. We compared genomic coordinates from our events to those present on the Database of Genomic Variants (DGV), version 3, November 29, 2007 update (S16). Any previously described event that had at least a 60% overlap with a newly discovered case event was considered ‘not rare’ and excluded from further evaluation. Since CNVs from the NINDS controls (S4) (NDPT002, NDPT006, and NDPT009) are reported in the DGV (S17), events in these individuals were only excluded if they were also found in other studies. The DGV is dynamic and not equally representative of all populations. Therefore it was important to define all potential events against the same version of the DGV and to evaluate cases and controls matched for ancestral populations.

We recognized that rare events were more likely than common events to be false positives, because most rare events appear in only one sample. To exclude false positives among these potential rare variants, we tested every potential rare variant in cases and controls by two independent platforms.

First, DNA samples from 24 cases potentially harboring such variants were tested using an independent method: Illumina 550K SNP arrays (S18), screened at the Nickerson Laboratory, Dept of Genome Sciences, University of Washington. DNA samples from Caucasian controls had been tested independently at the NIH using Illumina HapMap 300K SNP chips and subsequently with supplemental S240K chips, which combined comprise essentially identical coverage to the 550K arrays.

A quantitative measurement for the number of copies of each allele at each SNP was generated during the genotyping process. These data can be exploited to infer copy-number status by combining information across many probes in a genomic interval (S19). We considered both the logR ratio, a normalized intensity value that measures the total amount of DNA hybridized to a given probe and combines information from both alleles, and the B-allele frequency, which represents the total fraction of the intensity for a given site that can be attributed to the presence of the B allele (i.e. 0.0, 0.5, or 1.0 for AA, AB, and BB genotypes, respectively). Deletions appear as groups of probes with significantly depressed logR ratio values that are also heavily enriched for hemizygous SNPs (i.e. B-allele frequencies near 1.0 or 0.0). Duplications appear as elevated logR ratio values with B allele frequencies at heterozygous SNPs that deviate from 1:1 allelic ratios, since one of the alleles is present at elevated copy-number (S19). For our CNV discovery procedure, we built a simple Hidden Markov Model (HMM) that simultaneously analyzes both logR ratios and B-allele frequencies and identifies possible regions of deletion or duplication, either heterozygous or homozygous. This HMM was optimized through manual analyses on samples with well-characterized deletion and duplication events and implemented using standard available tools (S20)**.**

Secondly, all possible gene-impacting rare variants in cases and controls were analyzed by microarray comparative genomic hybridization, using the same reference genome that was used for the ROMA analysis. Microarrays consisting of 2.1 million probes per array were designed and manufactured by NimbleGen (HD2 070713_HG18_WG_CGH_HX1 design), with probes selected to achieve a uniform distribution throughout the genome, approximately one probe every 1200 bp, maximizing uniqueness within each 1200bp interval, permitting up to five exact matches to the genome. Hybridizations were performed as follows. One microgram of genomic DNA was klenow-amplified in duplicate with Cy-3 labeled random 9-mer primers in parallel with duplicate klenow amplification of the reference DNA with Cy-5 labeled 9-mers. Following reaction termination and ethanol purification, Cy-labeled DNA was dried and re-hydrated in de-ionized water. Thirty micrograms of test and reference Cy-labeled DNA were combined and co-hybridized on the HD2 array in the presence of Cy-3 and Cy-5 CPK6 48-bp oligomers. The hybridization solution was circulated across the HD2 array while maintained at 42°C for 60 hours on a MAUI hybridization System (BioMicro Systems). Slides were washed in buffers of decreasing salt concentration containing DTT and spin dried before scanning. Each slide was scanned at a 5um resolution and images were imported into NimbleScan, a software package provided by NimbleGen to identify copy number variants from HD2 image and intensity data. SignalMap (NimbleGen) was used to visualize the normalized-segmented data. Representative examples of the SignalMap GFF output are shown in Figure S3 (below). Whole genomes were scanned by Illumina and NimbleGen arrays, as described above, but only variants previously discovered by ROMA were included in this study.

## MUTATION DISCOVERY AND VALIDATION IN COS SAMPLES

Cases with childhood onset schizophrenia (COS) and their parents were evaluated independently and identically, using different platforms than the original series.

### Affymetrix 500K SNP arrays

COS patients and their parents were assessed by hybridizing genomic DNA samples to the Affymetrix Mapping 250K NspI and StyI Assay kits. Assays were carried out according to the manufacturer's protocol, beginning with 250ng DNA. To obtain estimates of copy number, we designed the algorithm described below. First, for each array, for each SNP we sum together all the available A and B allelic probe signals, to get a net A+B signal for the entire SNP, which we consider as the raw intensity signal for that marker. Then the entire array is normalized by dividing by the average of the raw signals, with the average computed across the autosomes to avoid biases from sex chromosome copy number differences. The resulting normalized signals are then converted to a Z score relative to the mean and deviation across all available arrays, using a robust estimate of the mean and deviation based on excluding the top and bottom 20% of all signals, to allow for possible common copy number variants. The resulting Z scores are not reliably normally distributed, so we use them in an empirical fashion rather than presuming them to be normal. To scan for regions of loss or gain in a given sample, we consider a window size of L SNPs ($2 \leq L \leq 8000$), scan across the genome considering all L-SNP windows, sum the Z scores for the interval for each L-SNP interval, and form the empirical distribution of these L-SNP scores across the genome for the sample in question. This distribution is very nearly normal, but the lower and upper tails fit different normal distributions, as the upper tail is generally longer, reflecting the fact that hybridization probes can have a broader range of high signals than low signals. Any L-SNP interval that scores more than 5 standard deviations (upper or lower, respectively) from the mean is selected as a hit. For each hit, we estimate the endpoints of the interval by maximizing the local odds ratio.

### Agilent arrayCGH

For Agilent arrayCGH, proband DNA and Invitrogen pooled reference DNA of the opposite sex were prepared according to Agilent CGH protocol, and hybridized to Agilent 185K or 244K oligonucleotide DNA microarrays (Agilent Technologies, Palo Alto, CA). Further, custom 8x15K targeted arrays were designed to ascertain novel events identified in the parents. The arrays were washed and scanned, and intensity data was extracted from the scanned images with Agilent Feature Extraction 8.0. Nexus 2.0 and its built in Rank Segmentation Algorithm were used to segment each interval and estimate copy number. This method is a variant of the Circular

Binary Segmentation (S21), but uses a parametric model rather than permutation testing to determine significance of segmentation clusters. It divides the genome into segments such that the probe log-ratio values in each segment are deemed drawn from a different distribution than those of adjacent segments and all values in the segment are deemed to be from the same distribution. The significance threshold was set to $5x10^{-4}$. Once the segmentation was performed, any segment with a log-ratio value > 0.25 or < -0.40 was defined as having a gain or loss, respectively.

**Targeted segmental duplication array**

COS DNA samples were hybridized to a custom BAC microarray consisting of 2007 large insert BAC clones (S22). The microarray targets regions of the genome that are flanked by segmental duplications >10 kb in length and >95% sequence identity. This includes most regions associated with known genomic disorders and an additional 105 regions with similar genomic architecture. Array comparative genomic hybridization experiments were performed in replicate with the fluorescent label swapped between the test and reference sample (GM15724, Coriell Institute). Regions were scored as microdeleted or microduplicated if the log2 ratio of two or more consecutive clones exceeded twice the standard deviation of the autosomal clones in dye-swap replicate experiments (S22). Novel copy number variants were defined as described in the text; that is, not previously detected in population controls assessed either by these arrays or by other detection platforms. Previously unreported structural variants that did not impact any known genes are indicated in Table S2. Novel structural variants that deleted or duplicated genes in the original series of cases and controls are indicated in text Table 2. Novel structural variants that deleted or duplicated genes in the COS cases and controls are indicated in Table S3.

## TESTING FOR CELL LINE ARTIFACTS

Transformation of lymphoblasts can introduce rearrangements that are detected as copy number differences at immunoglobulin gene clusters, reflecting normal VDJ-type recombination at these sites. Previous comparisons of CNV patterns in transformed lymphoblasts and blood revealed no CNVs at other sites introduced by the transformation procedure (S14). Nonetheless, in all cases in the original series, we tested genomic DNA extracted directly from peripheral blood to confirm rare variants. All rare variants were present in DNA from blood and sizes defined by ROMA were the same. DNA from peripheral blood was not available from controls. Although unlikely, it is possible in principle that one or more rare events in controls could be artifacts of transformation, which would introduce a conservative bias.

## IDENTIFICATION OF MUTATION BREAKPOINTS IN THE ORIGINAL SERIES

### Genomic Sequencing

In order to determine genomic breakpoints of validated duplications and deletions in the original series, PCR primers were designed from approximately 1kb boundaries (Illumina 550K and NimbleGen HD2 data) surrounding rare structural mutations. Long-range PCR amplification was performed on genomic DNA from the appropriate subject, using Takara LA Taq as described in the manufactures protocol. PCR products were purified and sequenced with BigDyeV3.1 chemistry on an ABI 3130XL capillary instrument. Experimentally derived sequences were aligned to the UCSC Genome Browser and exact genomic breakpoints determined. Diagnostic PCR assays were designed and performed to confirm the breakpoints.

### Genes disrupted by mutations in cases and controls

The 24 genes disrupted by structural variants in cases were significantly over-represented in pathways important for brain development (text Tables 2 and 3). The 12 genes disrupted in controls were not over-represented in any pathway. A meta-analysis of results based on CNV detection methods of lower resolution than those used in this study suggested that signaling genes as a general class were enriched in genomic regions with copy number variants (S23). Subsequent analyses suggest that this apparent over-representation of signaling genes was an artifact of over-estimation of CNV size, so that events were mistakenly defined as impacting a gene when in fact they were intergenic. Also, individuals with neurological illnesses were included in the meta-analysis. Neither of these problems appeared in this study, because higher resolution platforms were used, breakpoints were precisely defined, and controls were tested for absence of neurological illnesses.

We were concerned nonetheless that the apparent over-representation of neurodevelopmental signaling genes might be the result simply of the larger number of genes disrupted by mutations in cases (24 genes in 150 cases) compared to controls (12 genes in 268 controls). We carried out a simple simulation, as follows. For each replicate of the simulation, from the 24 genes disrupted by mutations in cases, we selected at random 12 genes, equal to the number of genes disrupted by mutations in controls. We then assessed each subset of 12 genes from the case series by undirected PANTHER analysis (S24) and undirected Ingenuity Pathway Analysis (S25) and noted which pathways and processes were over-represented. We repeated this process 50 times. Table S4 (below) indicates the proportion of simulations in which the random subsets of 12 case genes appeared significantly more frequently than expected by chance in each of the neurodevelopmental pathways and processes revealed by the entire set of 24 case genes. For each pathway, the number of replicates over-represented by subsets of 12 case genes was significantly greater than zero.

## REFERENCES

S1. J. McClellan *et al., J Am Acad Child Adolesc. Psychiatry* **46** 969-978 (2007).

S2. M. B. First, M. Gibbon, R. L. Spitzer, J. Williams, *User's guide for the structured clinical interview for DSM-IV Axis I Disorders: Research version.* American Psychiatric Press, Washington DC (1996).

S3. Matzner F *et al.*, *Preliminary test-retest reliability of the KID-SCID.* American Psychiatric Press, Washington DC (1997).

S4. H. C. Fung *et al.*, *Lancet Neurol.* **5,** 911 (2006).

S5. M. K. Mitchell *et al., J. Urban Health* **81,** 301 (2004).

S6. W. K. Chung, *Gender Med*. **4,** 248 (2007).

S7. M.K. Mitchell *et al., J Urban Health* **81**, 301 (2004).

S8. J. O. Korbel *et al., Science* **318,** 420 (2007).

S9. R. Nicolson, J. L. Rapoport, *Biol. Psychiatry* **46**, 1418 (1999).

S10. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders, 4th ed.* American Psychiatric Press, Washington DC (1994).

S11. K. McKenna *et al., J Am Acad Child Adolesc. Psychiatry* **33**, 636 (1994).

S12. J. Kaufman *et al., J Am Acad Child Adolesc. Psychiatry* **36**, 980 (1997).

S13. P. J. Ambrosini *et al., J Am Acad Child Adolesc. Psychiatry* **39**, 49 (2000).

S14. J. Sebat *et al., Science* **305**, 525 (2004).

S15. R. Lucito *et al*., *Genome Res*. **13**, 2291 (2003).

S16. A. J. Iafrate *et al*., *Nat. Genet.* **36,** 949 (2004).

S17. J. Simon-Sanchez et al., *Hum Mol Genet*. 16, 1 (2007).

S18. K. L. Gunderson *et al., Pharmacogenomics* **7**, 641 (2006).

S19. D. A. Peiffer *et al*., *Genome Res.* **16,** 1136 (2006).

S20. D. N. Day *et al*.. *Bioinformatics* **23,** 1424 (2007).

S21. A. B. Olshen *et al., Biostatistics* **5**, 557-72 (2004).

S22. A. J. Sharp *et al.*, *Am J Hum Genet*. **77**, 78-88 (2005)

S23. G. M. Cooper, D. A. Nickerson, E. E. Eichler, *Nat. Genet.* **39,** S22 (2007).

S24. H. Mi, N. Guo, A. Kejariwal, P. D. Thomas, *Nucl. Acids Res.* **35,** D247 (2007)

S25. Ingenuity Systems, www.ingenuity.com (2007)

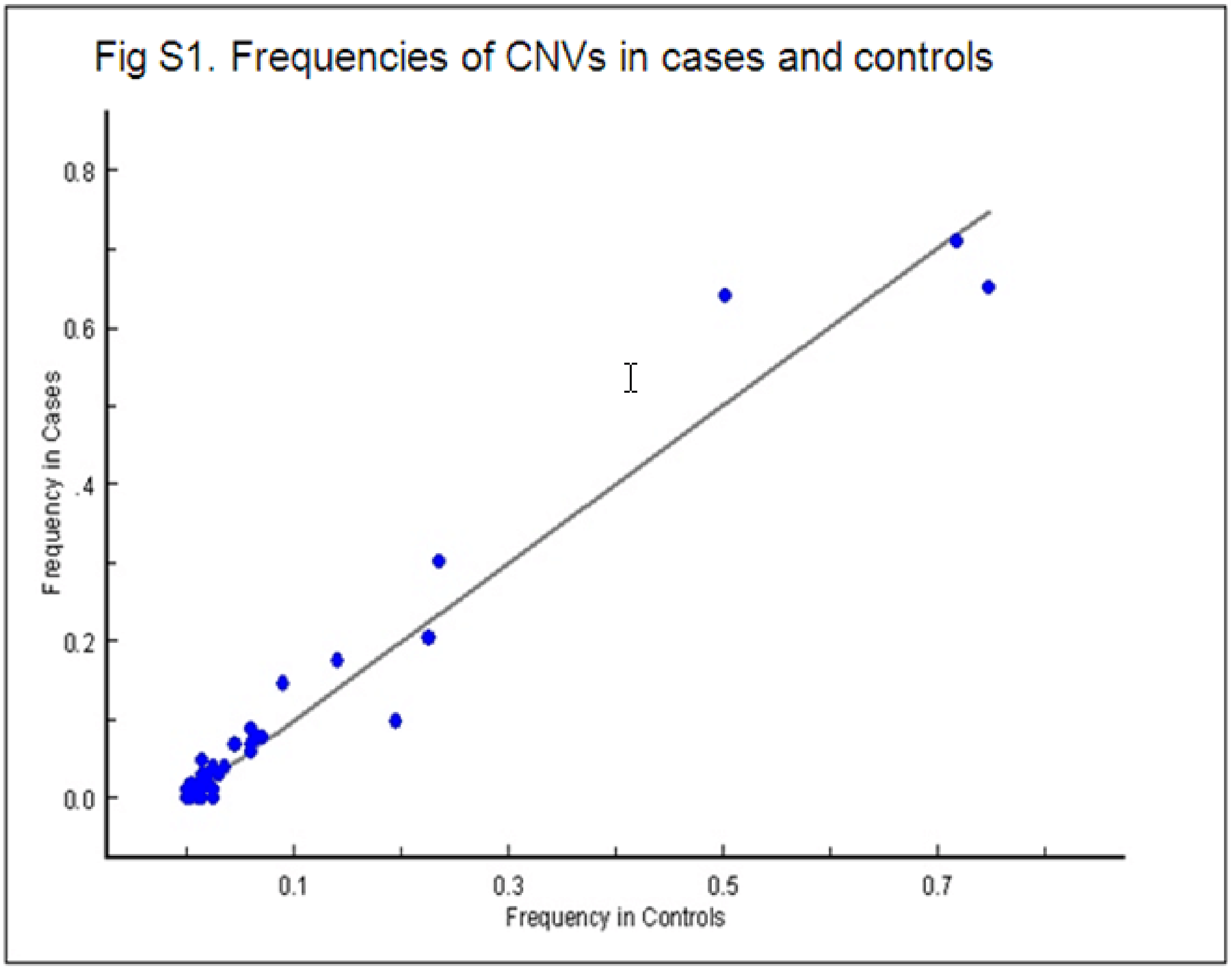


**Fig S1**. Frequencies of common CNVs in controls (X-axis) and in cases (Y-axis)

**Fig. S2. ROMA sensitivity in cases and controls**

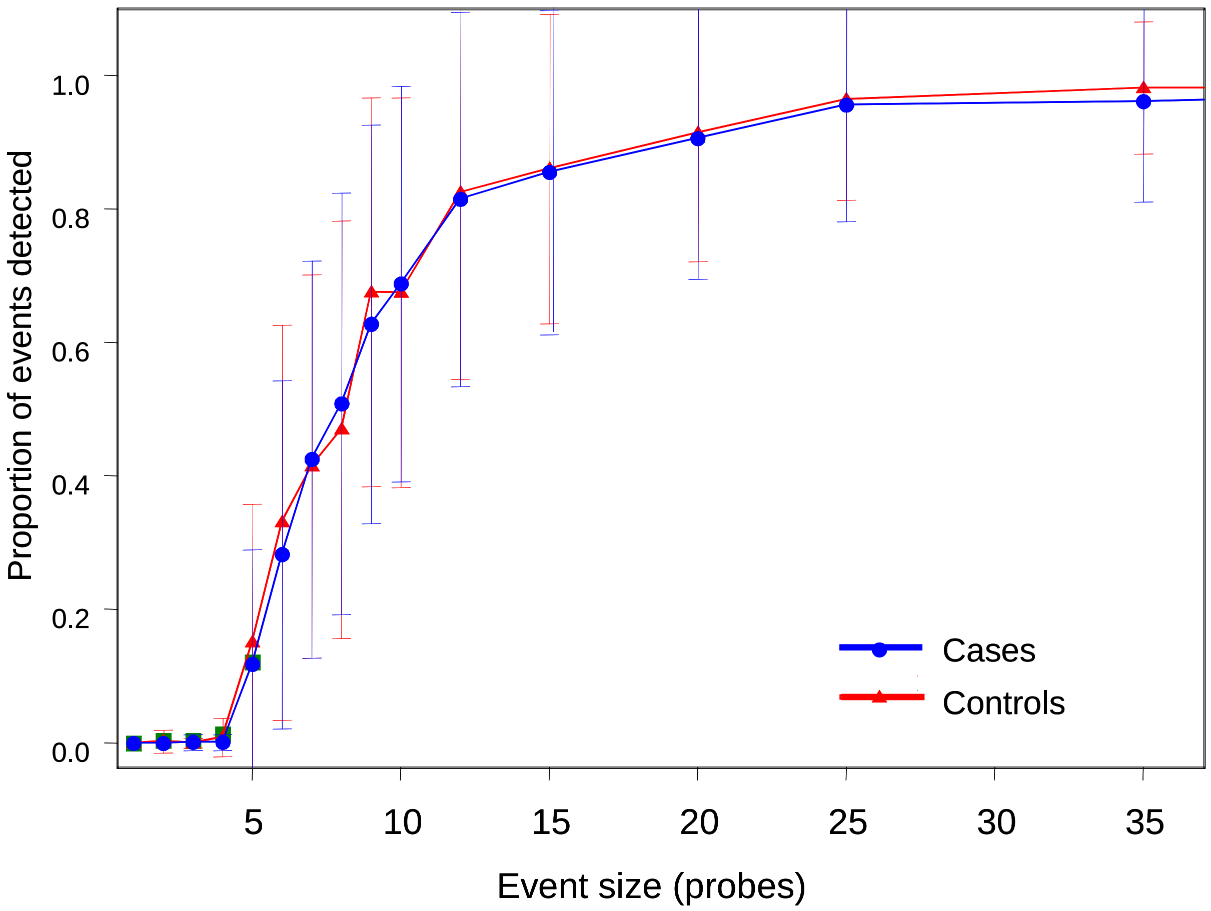
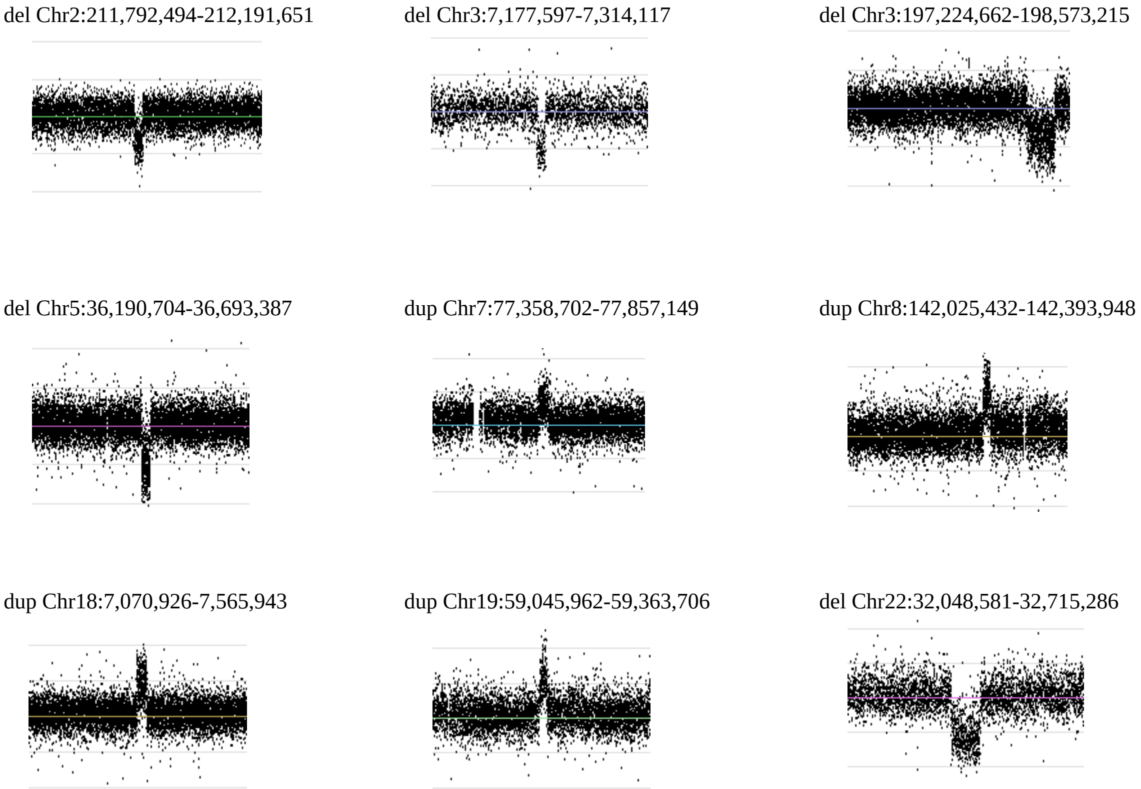


**Fig S2**. Sensitivity of ROMA in detecting simulated CNVs introduced into the ROMA data at 170 randomly selected sites

**Fig S3**. Examples of arrayCGH on the NimbleGenHD2 platform, visualized with SignalMap software (NimbleGen). Coordinates as in Table 2.



del Chr2:211,792,494-212,191,651

del Chr3:7,177,597-7,314,117

del Chr3:197,224,662-198,573,215

del Chr5:36,190,704-36,693,387

dup Chr7:77,358,702-77,857,149

dup Chr8:142,025,432-142,393,948

dup Chr18:7,070,926-7,565,943

dup Chr19:59,045,962-59,363,706

del Chr22:32,048,581-32,715,286

**Table S1.** Clinical information on adult and youth with schizophrenia in the original series

| | N | Proportion | Mean (S.D.) | Range |
|---|---|---|---|---|
| Youth with schizophrenia (N = 30) | | | | |
| Diagnosis | | | | |
| Schizophrena | 23 | 0.77 | | |
| Schizoaffective disorder | 7 | 0.23 | | |
| Age of onset* | | | 13.2 (2.1) years | 9 - 17 years |
| Age enrolled | | | 14.2 (2.2) years | 9 - 18 years |
| Full Scale IQ (N = 26) | | | 95.4 (18.5) | 60 - 127 |
| Estimated IQ < 80* | 6 | 0.20 | | |
| Family history of mental illness ($1^o$ relatives)* | | | | |
| Schizophrenia/psychosis | 4 | 0.13 | | |
| Depression | 8 | 0.27 | | |
| Bipolar disorder | 3 | 0.10 | | |
| Substance Abuse | 6 | 0.20 | | |
| Western State Hospital recruitment (N = 120) | | | | |
| Diagnosis | | | | |
| Schizophrena | 119 | 0.99 | | |
| Schizoaffective disorder | 1 | 0.01 | | |
| Age of onset* | | | 20.9 (5.1) years | 13 - 40 years |
| Age enrolled | | | 40.0 (10.4) years | 16 - 64 years |
| Median length of index hospitalization | | | 1.8 years | 1 month - 24 years |
| Number of hospitalizations | | | 6.1 (4.5) | 1 - 22 hospitalizations |
| Full Scale IQ (N = 12) | | | 92.5 (17.7) | 64 - 123 |
| Estimated IQ < 80* | 9 | 0.08 | | |
| Forensic involvement* | 77 | 0.64 | | |
| Family history of mental illness ($1^o$ relatives)* | | | | |
| Schizophrenia/psychosis | 22 | 0.18 | | |
| Depression | 17 | 0.14 | | |
| Bipolar disorder | 9 | 0.08 | | |
| Substance Abuse | 45 | 0.38 | | |

*see text of supplementary online materials for details

**Table S2**. Novel structural variants >100kb detected in genomic DNA in schizophrenia cases and controls, that do not impact genes

| Chr | Start (hg18) | End (hg18) | Size (bp) | Type |
|---|---|---|---|---|
| Original cases and controls | | | | |
| Cases (N = 150) | | | | |
| 1 | 186,167,694 | 186,861,470 | 693,776 | dup |
| 2 | 76,385,979 | 76,506,055 | 120,076 | del |
| 6 | 86,811,196 | 87,625,627 | 814,431 | dup |
| 8 | 77,184,932 | 77,381,835 | 196,903 | del |
| 9 | 29,458,218 | 29,744,786 | 286,568 | dup |
| 11 | 81,575,797 | 82,039,711 | 463,914 | del |
| 13 | 87,658,310 | 88,466,330 | 808,020 | del |
| 16 | 78,344,332 | 78,458,818 | 114,486 | del |
| 21 | 23,032,927 | 23,481,793 | 448,866 | del |
| Controls (N = 268) | | | | |
| 1 | 191,577,701 | 191,818,828 | 241,127 | del |
| 3 | 22,559,694 | 22,931,263 | 371,569 | del |
| 5 | 159,030,472 | 159,253,103 | 222,631 | dup |
| 5 | 173,629,482 | 174,010,133 | 380,651 | dup |
| 13 | 80,097,169 | 82,440,119 | 2,342,950 | del |
| 14 | 39,923,379 | 40,336,749 | 413,370 | del |
| 14 | 61,792,989 | 61,931,611 | 138,622 | del |
| 21 | 23,032,928 | 23,209,873 | 176,945 | del |
| Childhood onset schizophrenia (COS) cases and controls | | | | |
| Cases (N = 83) | | | | |
| 3 | 67,837,385 | 67,967,590 | 130,205 | del |
| 7 | 83,231,037 | 83,341,369 | 110,332 | del |
| 7 | 144,394,475 | 144,768,973 | 374,498 | del |
| 10 | 128,335,145 | 128,527,314 | 192,169 | del |
| 11 | 90,794,964 | 90,957,655 | 162,691 | dup |
| 14 | 40,276,253 | 40,605,077 | 328,824 | del |
| Controls (N = 77) | | | | |
| 2 | 40,858,263 | 41,152,415 | 294,152 | dup |
| 4 | 75,540,844 | 75,774,745 | 233,901 | dup |
| 10 | 6,758,245 | 6,878,948 | 120,703 | dup |
| 16 | 61,953,576 | 62,127,751 | 174,175 | del |
| 20 | 4,469,289 | 4,573,558 | 104,269 | del |

**TableS3**. Novel structural variants in cases with childhood onset schizophrenia (COS) and non-transmitted chromosomes from their parents (controls)

| | | | Size | | Duplicated or | COS | Inherited or |
|---|---|---|---|---|---|---|---|
| Chr | Start (hg18) | End (hg18) | (kb) | Type of event | deleted genes | family | *de novo* |
| COS cases (N = 83) | | | | | | | |
| 1 | 151,514,380 | 151,762,871 | 248 | duplication | 2 | 885 | inherited |
| 2 | 1,618,945 | 1,835,426 | 216 | duplication | 2 | 1358 | inherited |
| 2 | 1,713,636 | 1,857,129 | 143 | duplication | 2 | 534 ^ | not known |
| 2 | 50,023,212 | 50,137,825 | 115 | deletion | 1 | 581 | not known |
| 2 | 65,637,097 | 65,879,935 | 243 | duplication | 1 | 1182 | inherited |
| 2 | 179,643,864 | 182,145,339 | 2501 | deletion | 6 | 483 | *de novo* |
| 3 | 9,100,744 | 9,220,529 | 120 | duplication | 1 | 499 | inherited |
| 3 | 45,458,901 | 45,576,135 | 117 | duplication | 1 | 481 | inherited |
| 5 | 64,795,287 | 64,937,409 | 142 | duplication | 5 | 1677 | inherited |
| 6 | 119,596,633 | 119,740,850 | 144 | deletion | 1 | 1870 | inherited |
| 7 | 44,420,900 | 44,540,491 | 120 | duplication | 1 | 1127 | not known |
| 7 | 64,126,564 | 66,883,376 | 2757 | duplication | 13 | 449 | inherited |
| 8 | 13,400,795 | 14,679,483 | 1279 | duplication | 2 | 755 ^ | inherited |
| 8 | 53,563,161 | 54,043,063 | 480 | duplication | 3 | 534 ^ | not known |
| 10 | 15,688,654 | 15,833,865 | 145 | duplication | 1 | 452 | inherited |
| 10 | 28,990,284 | 29,166,175 | 176 | duplication | 2 | 755 ^ | inherited |
| 15 | 96,246,764 | 96,933,404 | 687 | duplication | 2 | 588 ^ | not known |
| 16 | 29,652,656 | 30,085,308 | 433 | duplication | 24 | 676 | inherited |
| 16 | 29,657,405 | 30,235,818 | 578 | duplication | 24 | 2011 | inherited |
| 16 | 80,737,839 | 82,208,451 | 1471 | duplication | 2 | 691 ^ | inherited |
| 16 | 82,997,582 | 83,108,554 | 111 | deletion | 2 | 1719 | inherited |
| 18 | 7,067,237 | 7,576,777 | 510 | duplication | 2 | 552 | inherited |
| 18 | 61,907,915 | 62,675,869 | 768 | duplication | 1 | 1251 | inherited |
| 19 | 23,413,380 | 23,810,606 | 397 | deletion | 3 | 588 ^ | not known |
| 20 | 14,921,777 | 15,034,862 | 113 | deletion | 1 | 691 ^ | inherited |
| X | 8,384,117 | 8,726,291 | 342 | duplication | 3 | 1374 | not known |
| Y | 14,441,161 | 14,623,937 | 183 | duplication | 1 | 1012 | *de novo* |
| Controls (N = 77) | | | | | | | |
| 3 | 184,352,512 | 184,574,024 | 222 | duplication | 3 | 886 | na |
| 5 | 23,608,440 | 24,032,911 | 424 | duplication | 1 | 1791 | na |
| 7 | 109,838,498 | 110,129,771 | 291 | deletion | 1 | 665 | na |
| 9 | 15,337,518 | 15,497,987 | 160 | duplication | 2 | 1566 | na |
| 9 | 97,760,740 | 98,090,615 | 330 | duplication | 3 | 2072 | na |
| 9 | 118,728,654 | 118,901,548 | 173 | deletion | 1 | 1798 | na |
| 11 | 47,891,299 | 48,072,808 | 182 | duplication | 1 | 1825 | na |
| 11 | 48,161,650 | 48,386,976 | 225 | duplication | 5 | 609 | na |
| 13 | 90,860,332 | 91,099,862 | 240 | deletion | 1 | 1889 | na |
| 18 | 62,445,926 | 64,327,328 | 1881 | duplication | 1 | 1842 | na |

^Individuals with more than one event

**Table S4**. Fifty replicates of PANTHER and Ingenuity Pathway Analysis, each based on a randomly selected subset of 12 genes from the 24 genes disrupted by mutations in cases. No pathways were over-represented by the 12 genes disrupted by mutations in controls.

| Pathway or process | Proportion of replicates in which pathway was over-represented by subsets of 12 "case genes" |
|---|---|
| Signal transduction* | .84 |
| Neuronal activities* | .66 |
| Nitric oxide signaling^ | .66 |
| Synaptic long term potentiation^ | .68 |
| Glutamate receptor signaling^ | .38 |
| ERK/MAPK signaling^ | .70 |
| PTEN signaling^ | .32 |
| Neuregulin signaling^ | .42 |
| IGF-1 signaling^ | .38 |
| Axonal guidance signaling^ | .64 |
| Synaptic long term depression^ | .26 |
| G-protein coupled receptor signaling^ | .28 |
| Integrin signaling^ | .40 |
| Ephrin receptor signaling^ | .38 |
| Sonic hedgehog signaling^ | .62 |

*Undirected PANTHER analysis

^Undirected Ingenuity Pathway Analysis