Exome sequencing supports a *de novo* mutational paradigm for schizophrenia

Bin Xu[1,2], J Louw Roos[3], Phillip Dexheimer[4], Braden Boone[4], Brooks Plummer[4], Shawn Levy[4], Joseph A Gogos[2,5] & Maria Karayiorgou[1]

Despite its high heritability, a large fraction of individuals with schizophrenia do not have a family history of the disease (sporadic cases). Here we examined the possibility that rare *de novo* protein-altering mutations contribute to the genetic component of schizophrenia by sequencing the exomes of 53 sporadic cases, 22 unaffected controls and their parents. We identified 40 *de novo* mutations in 27 cases affecting 40 genes, including a potentially disruptive mutation in *DGCR2*, a gene located in the schizophrenia-predisposing 22q11.2 microdeletion region. A comparison to rare inherited variants indicated that the identified *de novo* mutations show a large excess of non-synonymous changes in schizophrenia cases, as well as a greater potential to affect protein structure and function. Our analyses suggest a major role for *de novo* mutations in schizophrenia as well as a large mutational target, which together provide a plausible explanation for the high global incidence and persistence of the disease.

Schizophrenia has a strong genetic component[1,2]. However, despite its high heritability, a large fraction of individuals with schizophrenia do not have a family history of the disease[3]. Although largely ignored in earlier efforts to model disease risk, *de novo* germline mutations may account for a substantial fraction of sporadic schizophrenia cases. In agreement with this hypothesis, rare *de novo* copy number variants (CNVs)[4] are emerging as an important genomic cause of schizophrenia and other psychiatric diseases, and the variant with the strongest statistical support for association with schizophrenia, namely the 22q11.2 microdeletion, is a *de novo* and recurrent mutation[5,6].

Availability of next-generation whole-genome or whole-exome sequencing[7] now permits the study of *de novo* mutations (point substitutions or single nucleotide variants (SNVs) and small insertions or deletions (indels)) in a systematic genome-wide manner[8,9]. Pilot studies focusing on specific synaptic genes identified a small number of putative *de novo* mutations in individuals with schizophrenia[10]. However, the full contribution of rare *de novo* SNVs and indels to schizophrenia remains unknown.

To test the hypothesis that *de novo* protein-altering mutations contribute substantially to the genetic component of schizophrenia, we sequenced the exomes of 53 family trios of subjects diagnosed with schizophrenia or schizoaffective disorder with no history of the disease in a first- or second-degree relative ('sporadic cases cohort') as well as family trios of 22 unrelated healthy controls, all recruited from the genetically homogeneous Afrikaner population of European descent in South Africa[11,12]. Presence or absence of family history in cases was not a screening criterion during recruitment but could be reliably determined because of the close-knit family structure of the families and the availability of detailed psychiatric records over several generations because of the large catchment area and long-term care provided by the local recruiting hospital[6,12,13]. Control families completed a detailed self-report questionnaire that inquired about several psychiatric conditions, including phobias, anxiety, depression and history of treatment for any of these conditions. Also, mental illness in first- or second-degree relatives was excluded. Based on previous results[6,13], we excluded carriers of rare *de novo* CNVs. Identities were coded and analyses were performed blind to affected status while maintaining knowledge of the parent-child relationships. From all 225 individuals, we extracted DNA samples from whole blood.

We enriched exonic sequences using the Agilent SureSelect technology for targeted exon capture and performed Illumina paired-end sequencing (one lane of flow cell per sample; see Online Methods). On average, we obtained 7.3 Gb of mappable sequence data per individual after exome enrichment, targeting 37 Mb from exons and their flanking regions. Overall, we covered 1.22% of the genome, a fraction corresponding to the NCBI Consensus Coding Sequences database (CCDS). We cross matched the paired-end reads to the reference genome (hg19 build) using the Burrows-Wheeler Aligner (BWA v0.5.81536)[14]. Ninety-seven point nine percent of the reads were properly aligned to the reference genome. Our median read depth was 65.2×, which is higher than the estimated average depth (33×) required for highly accurate downstream heterozygous variant detection. In addition, 92.4% of the captured target exons were covered by high quality genotype calls at least eight times to ensure good detection sensitivity[15] (**Table 1**).

Our *de novo* mutation detection pipeline is shown in **Supplementary Figure 1**. We implemented a series of filters, including final validation by standard Sanger sequencing (**Supplementary Figs. 2**,**3**), to eliminate

[1]Department of Psychiatry, Columbia University, New York, New York, USA. [2]Department of Physiology & Cellular Biophysics, Columbia University, New York, New York, USA. [3]Weskoppies Hospital & Department of Psychiatry, University of Pretoria, Pretoria, South Africa. [4]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, USA. [5]Department of Neuroscience, Columbia University, New York, New York, USA. Correspondence should be addressed to M.K. (mk2758@columbia.edu) or J.A.G. (jag90@columbia.edu).

Received 26 May; accepted 12 July; published online 7 August 2011; doi:10.1038/ng.902

Table 1 Overview of exome sequencing data

	Average	Percent
Total sequence (bp)	7,475,900,117	100.00
Aligned sequence (bp)	7,319,357,510	98.34
Aligned paired reads	69,747,339	97.55
Aligned singleton reads	566,300	0.35
Median read depth	65.2×	–
1× coverage	37,191,631	98.84
4× coverage	36,190,388	96.18
8× coverage	34,769,095	92.40
20× coverage	30,428,158	80.87
30× coverage	27,171,231	72.21

variants that would appear *de novo* either from undercalling in the parents or from systematic false positive calls in the subjects (Online Methods). In total, in the affected trios, we observed 34 *de novo* point mutations (33 SNVs and 1 dinucleotide substitution) and 4 *de novo* indel candidates (**Table 2**). Among the 34 *de novo* point mutations, 32 are predicted to be non-synonymous missense mutations and 2 are predicted to be synonymous. Of the 32 non-synonymous mutations, 19 affect evolutionarily conserved residues and are predicted to affect protein function by PolyPhen-2 analysis. Three of the indels result in protein truncations and one causes a single amino acid deletion. Additional analyses of *de novo* SNVs located within the flanking intronic regions identified two SNVs located within predicted donor or acceptor splice sites (**Table 2**). Notably, among the identified exonic SNVs, one synonymous and three non-synonymous SNVs were also located within predicted donor or acceptor splice sites (**Table 2**). Further analysis using the Human Splicing Finder tool (see URLs) showed that two out of the six mutations directly alter splice signals and may interfere with splicing (**Table 2**). By our filtering

Table 2 *De novo* mutations identified in 53 schizophrenia trios

Gene	Mutation type	NS or S	PolyPhen-2	Grantham score	phyloP score	Chr. (position)	Nucleotide change	Amino acid change	Diagnosis	Sex	Trio ID
PLCL2	SNV	NS	Probably damaging	112	4.99	3 (17,051,253)	TGT-aGT	p.Cys30Ser	SCZ	M	trio_002
WDR11	SNV	NS	Probably damaging	29	6.30	10 (122,664,879)	CGC-CaC	p.Arg1081His	SCZ	M	trio_011
DPYD	SNV	NS	Probably damaging	125	3.89	1 (97,981,407)	GGA-aGA	p.Gly539Arg	SCZ	M	trio_016
OR4C46	SNV	NS	Probably damaging	125	1.61	11 (51,515,885)	GGA-aGA	p.Gly202Arg	SCZ	F	trio_019
UGT1A3	SNV	NS	Probably damaging	15	0.53	2 (234,637,866)	TTG-aTG	p.Leu32Met	SCZ	F	trio_023
FAM3D	SNV	NS	Probably damaging	194	1.18	3 (58,622,886)	TAC-TgC	p.Tyr147Cys	SCZ	M	trio_024
KLF12	SNV	NS	Probably damaging	112	6.02	13 (74,289,537)	TCT-TgT	p.Ser45Cys	SCZ	M	trio_033
ADCY7	SNV	NS	Probably damaging	56	4.93	16 (50,349,011)	AGC-gGC	p.Ser1020Gly	SCZ	M	trio_038
GPR153	SNV	NS	Probably damaging	89	5.80	1 (6,314,661)	ACC-AtC	p.Thr102Ile	SCZAFF-dpr	M	trio_040
PML	SNV	NS	Probably damaging	81	2.93	15 (74,290,439)	ACG-AtG	p.Thr75Met	SCZAFF-dpr	M	trio_044
SLC26A8	SNV	NS	Probably damaging	56	1.87	6 (35,927,251)	GAG-aAG	p.Glu512Lys	SCZAFF-bp	F	trio_077
CCDC108	SNV	NS	Probably damaging	46	2.43	2 (219,900,235)	AAT-AgT	p.Asn105Ser	SCZ	F	trio_080
TRAK1	SNV	NS	Probably damaging	29	4.80	3 (42,261,055)	CAT-CgT	p.His678Arg	SCZ	F	trio_083
FASTKD5	SNV	NS	Probably damaging	60	5.61	20 (3,128,479)	GCA-GgA	p.Ala413Gly	SCZ	M	trio_089
DGCR2	SNV	NS	Probably damaging	103	6.24	22 (19,028,681)	CCT-CgT	p.Pro429Arg	SCZ	M	trio_091
ACOT6	SNV	NS	Possibly damaging	194	–0.08	14 (74,086,428)	TAT-TgT	p.Tyr170Cys	SCZ	F	trio_047
PITPNM1	SNV	NS/splice[a]	Probably damaging	101	0.30	11 (67,267,884)	CGG-tGG	p.Arg217Trp	SCZ	M	trio_039
NPRL2	SNV	NS/splice[b]	Possibly damaging	56	6.11	3 (50,385,987)	GGC-aGC	p.Gly231Ser	SCZ	F	trio_023
MAGEC1	SNV	NS	Unknown	58	0.37	X (140,993,957)	ACT-AgT	p.Thr256Ser	SCZ	M	trio_003
TRRAP	SNV	NS	Unknown	21	5.12	7 (98,498,329)	ATC-tTC	p.Ile295Phe	SCZ	M	trio_033
COL3A1	SNV	NS/splice[c]	Unknown	155	2.54	2 (189,851,792)	TCT-TtT	p.Ser152Phe	SCZ	M	trio_089
GIF	SNV	NS	Benign	29	0.83	11 (59,603,474)	GTA-aTA	p.Val294Ile	SCZAFF-dpr	F	trio_001
TEKT5	SNV	NS	Benign	89	0.25	16 (10,783,119)	ATC-AcC	p.Ile237Thr	SCZ	M	trio_011
THBS1	SNV	NS	Benign	56	5.66	15 (39,881,442)	GAG-aAG	p.Glu605Lys	SCZ	M	trio_015
PAG1	SNV	NS	Benign	29	0.15	8 (81,905,378)	GTC-aTC	p.Val29Ile	SCZ	M	trio_020
RGS12	SNV	NS	Benign	98	0.06	4 (3,429,844)	CCA-CtA	p.Pro518Leu	SCZAFF-dpr	M	trio_040
SAP30BP	SNV	NS	Benign	56	5.37	17 (73,702,542)	GGC-aGC	p.Gly274Ser	SCZ	F	trio_047
ZNF530	SNV	NS	Benign	46	–0.04	19 (58,118,122)	AGT-AaT	p.Ser410Asn	SCZAFF-bp	F	trio_077
MTOR	SNV	NS	Benign	46	2.82	1 (11,293,489)	AAT-AgT	p.Asn796Ser	SCZ	F	trio_080
INPP5A	SNV	NS	Benign	64	2.62	10 (134,463,942)	GCG-GtG	p.Ala80Val	SCZAFF-bp	F	trio_093
EDEM2	SNV	NS	Benign	83	3.39	20 (33,703,457)	TAC-cAC	p.Tyr469His	SCZ	M	trio_095
CELF2	SNV	S/splice[d]	Coding-synon	–	3.06	10 (11,356,223)	GGT-GGc	p.Gly345Gly	SCZ	M	trio_016
SLC26A7	SNV	S	Coding-synon	–	0.03	8 (92,346,630)	CAG-CAa	p.Gln250Gln	SCZ	M	trio_094
VPS35	SNV	Splice[e]	–	–	–	16 (46,705,610)	C/T	–	SCZ	M	trio_002
ADAMTS3	SNV	Splice[f]	–	–	–	4 (73,185,683)	G/A	–	SCZ	F	trio_023
GPR115	DNV	NS	Probably damaging	99	3.35	6 (47,682,855)	CTC-aaC	p.Leu625Asn	SCZ	F	trio_080
SPATA5	Indel	Amino acid deletion	Damaging	215[g]	–	4 (123,855,728)	TTCTT-caa-CAACA	–	SCZ	F	trio_023
RB1CC1	Indel	Frameshift deletion	Damaging	215	–	8 (53,568,705)	ACTGT-tc-TCTGT	–	SCZ	M	trio_026
LAMA2	Indel	Frameshift deletion	Damaging	215	–	6 (129,835,668)	GGTGG-aagccca-AAGCC	–	SCZ	M	trio_092
ESAM	Indel	Frameshift deletion	Damaging	215	–	11 (124,626,163)	tggac-AGCG-agcgg	–	SCZ	M	trio_042

NS, non-synonymous; S, synonymous; SNV, single nucleotide variant; DNV, dinucleotide variant; chr., chromosome; SCZ, schizophrenia; SCZAFF-dpr, schizoaffective disorder depressed subtype; SCZAFF-bp, schizoaffective disorder bipolar subtype; M, male; F, female.

[a–f]The difference (%) between the mutant and reference sequence in the HSF algorithm-derived consensus values at donor or acceptor splice sites: a = –2.48; b = 56.53; c = –0.68; d = –34.92; e = 0; f = –8.14 g. We used the maximum Grantham score (215) for the indels.

Table 3 NS:S ratio comparison between *de novo* and rare inherited mutations in schizophrenia trios

	Cases					Controls				
Class	Total number	NS	S	NS:S	P^a	Total number	NS	S	NS:S	P^a
De novo mutations	34	32	2	16.0		7	4	3	1.33	
Novel inherited variants	14,378	8,867	5,511	1.61	0.0002	6,213	3,825	2,388	1.60	0.81
Private inherited variants	6,727	4,223	2,504	1.69	0.0003	3,079	1,956	1,123	1.74	0.73

S, synonymous variants; NS, non-synonymous variants.
[a]*P* value for *de novo* compared to inherited mutations.

criteria, all identified *de novo* mutations were absent in a total of 1,658 control chromosomes (the exomes of the 679 individuals from the 1000 Genomes Project[16] included in dbSNP132, as well as of all 150 unaffected parents in our two cohorts). Overall, we found 27 out of 53 cases (~51%) to carry at least one *de novo* mutational event. This rate is comparable to that reported for 20 parent-child trios with autism spectrum disorders (51%)[8] but is somewhat lower than the rate reported for 10 parent-child trios with intellectual disability (90%)[9]. Ten of the 27 cases carried more than one *de novo* mutation, and the rest each carried a single mutation or indel.

Using the same pipeline and filtering criteria, we identified seven exonic *de novo* SNVs but no indel candidates in 7 out of 22 control subjects. Among these seven *de novo* point mutations, four are non-synonymous missense mutations and three are synonymous mutations. In addition, we identified one *de novo* mutation within a predicted intronic splice site (**Supplementary Table 1**). Overall, 7 out of 22 controls carry at least one *de novo* event, with one control having more than one *de novo* mutation. The fraction is lower when compared to cases, but the difference between the fraction of cases and controls carrying at least one *de novo* event is not statistically significant (Fisher's exact test, $P = 0.2$). There was no difference in the coverage between cases and controls or between trios with and without *de novo* events (**Supplementary Fig. 4**).

The overall *de novo* rate in affected families (0.75 events per family) is comparable to several empirical estimates of the background *de novo* mutation rate[8,16], suggesting that we identified most of the *de novo* events in these trios. Several lines of evidence suggest that the identified mutations have a high likelihood of causation with respect to schizophrenia. First, our screen yielded a ratio of non-synonymous missense ($n = 32$) to synonymous ($n = 2$) *de novo* changes (NS:S ratio) of 16:1, which is considerably higher than the 2.85:1 ratio expected based on the probability of causing an amino acid change under a random model (71.25% non-synonymous missense substitutions and 25% synonymous substitutions[17]). By contrast, the ratio of non-synonymous missense ($n = 4$) to synonymous ($n = 3$) *de novo* changes in the control cohort (NS:S ratio of 1.33:1) is consistent with neutral expectation and is very close to the NS:S ratio reported by the 1000 Genomes Project (1.14–1.45)[16]. Second, we found non-synonymous *de novo* point mutations in large excess compared to neutral ones relative to rare inherited variants, which are less likely to contribute to the pathogenesis in the sporadic cases (**Table 3**). Specifically, we first compared the relative enrichment of non-synonymous *de novo* point mutations to that observed among all novel (that is, not observed in dbSNP132) inherited variants segregating in cases. Our analysis identified a NS:S substitution ratio of 1.61:1, consistent with previous analyses of normal genetic variation[16,18]. Thus, in sporadic schizophrenia cases, rare *de novo* variants are approximately ten times more likely than inherited rare variants to harbor non-synonymous changes (χ^2 test $P = 0.0002$). A similar analysis in the control cohort did not show significant differences in the number of non-synonymous changes between the two types of variants. The NS:S ratios were 1.33 and ~1.60 for rare *de novo* and inherited variants, respectively (relative enrichment 0.83; χ^2 test $P = 0.81$). We obtained similar results when we limited the analysis to private inherited variants (that is, present only in one affected family), which serve as proxies for evolutionarily young mutation events. This analysis yielded a NS:S ratio of ~1.69:1 in cases (relative enrichment 9.5; χ^2 test $P = 0.0003$) and ~1.74 in controls (relative enrichment 0.76; χ^2 test $P = 0.73$). Consistent with expectations that disease mutations have a greater effect on protein function[19], we observed a more striking enrichment when we restricted our analysis to non-synonymous SNVs predicted by PolyPhen-2 to affect protein function. For such changes, the NS:S ratios in schizophrenia cases were 9.5 and ~0.79 for rare *de novo* and private inherited variants, respectively (relative enrichment 12.1; χ^2 test $P < 0.0001$).

We explored further the possibility that *de novo* mutations in cases have a greater potential to affect protein structure and function than private inherited variants by examining the evolutionary conservation of affected nucleotides (using the phyloP score[20]) as well as the potential of the *de novo* protein-altering mutations to affect the structure or function of the resulting proteins (using the Grantham score[21]) (**Table 2**). When we compared the cumulative distribution of these scores between *de novo* and private inherited variants in the sporadic cases cohort (**Fig. 1**), we observed that the distribution of the *de novo* variants was clearly shifted to the right (phyloP $P = 0.0005$ and Grantham $P = 0.14$). Overall, our analysis shows an enrichment of highly conserved and disruptive amino acid changes among *de novo* events and suggests a high likelihood for pathogenicity. Notably, carriers of one or more *de novo* mutations appear to be indistinguishable from other individuals with schizophrenia in terms of sex distribution, clinical presentation and developmental course (**Supplementary Note** and **Supplementary Table 2**).

All the mutations occurred in different genes, precluding a statistical assessment for any specific locus. Identification of recurrent mutations will provide proof of disease causality. With one exception, none of the affected genes has been previously associated with genetic loci or biological pathways unequivocally associated with schizophrenia. We therefore used phyloP and Grantham scores as a guide to prioritize

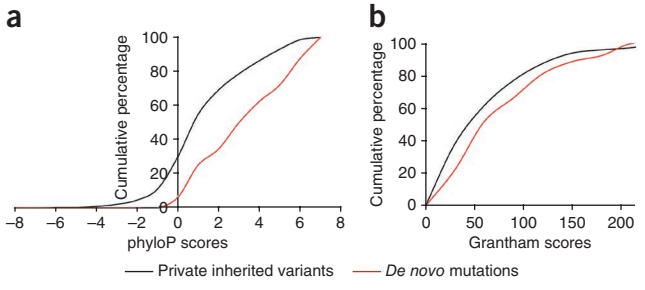

Figure 1 Assessment of the predicted pathogenicity of *de novo* SNVs identified in schizophrenia cases with respect to protein function. Comparison of the distribution of phyloP scores (which depend on the evolutionary conservation of affected nucleotides) (**a**) and Grantham scores (which depend on the properties of the changed residue) (**b**) for *de novo* mutations and private inherited variants found in schizophrenia cases.

for further discussion events that are more likely to be causal. In addition to protein truncating indels (in *LAMA2*, *SPATA5* and *RB1CC1*), there are 12 SNVs with phyloP scores ≥4 and 9 SNVs with Grantham scores ≥100, whereas 3 SNVs (in *DGCR2*, *KLF12* and *PLCL2*) show high values for both scores (**Table 2**). Most notable among the putative pathogenic events is a p.Pro429Arg substitution in *DGCR2* in a male with schizophrenia (**Supplementary Note**). *DGCR2* is located in the 22q11.2 locus and is hemizygously deleted by recurrent *de novo* microdeletions at this locus, which have high penetrance (~30%) and account for up to 2% of sporadic schizophrenia cases. The gene encodes a putative transmembrane adhesion receptor of unknown function[22]. The p.Pro429Arg substitution is located within a conserved domain of the protein (**Supplementary Fig. 2**) and shows one of the highest Grantham and phyloP scores among all the identified changes. Identification of a disruptive *de novo* SNV in *DGCR2* in a case with structurally intact 22q11.2 chromosomes suggests that disruption of this gene may contribute to the elevated schizophrenia risk associated with the 22q11.2 locus. Whether heterozygous deletions or point mutations in the *DGCR2* are sufficient to render the susceptibility to schizophrenia observed in 22q11.2 microdeletion carriers or whether additional genetic interactions are required[23] cannot be resolved until more *DGCR2* mutations are identified and their penetrance is determined. Additional putative pathogenic events were identified in three G-protein–coupled receptor (GPR) genes (*GPR153*, *GPR115* and *OR4C46*)[24,25] as well as in genes encoding proteins thought to either modulate (*RGS12*) or mediate aspects of GPR signaling, such as regulation of cAMP levels (*ADCY7*)[26]. Notably, we recently reported an association between schizophrenia and structural *de novo* mutations in another gene encoding a GPR (*VIPR2*)[27] and showed that these mutations alter cAMP levels. For other genes with high phyloP or Grantham scores, such as *WDR11*, *PLCL2*, *TRAK1*, *KLF12* and *LAMA2*, a potential causal link with schizophrenia is suggested by literature on previously described mutations, model organisms and other functional studies (**Supplementary Note**).

Our work shows that *de novo* protein-altering mutations contribute substantially to the genetic component of schizophrenia and, taken together with previous estimates of the *de novo* CNV rate in the same population[6], indicates that *de novo* mutations account for more than half of the sporadic cases of schizophrenia. Our findings are also in line with results from genome-wide scans for *de novo* CNVs[6,28], or CNVs in general[29,30], supporting the notion that multiple *de novo* genetic variants that affect many different genes contribute to the genetic risk of schizophrenia. The complexity of the neural substrates affected in schizophrenia and other psychiatric disorders offers a large mutational target comprised of many genes. We propose that this large number of targets that, when mutated, can give rise to schizophrenia, along with the relatively high rate of protein-altering mutations empirically shown in this study, provides a plausible explanation for both the high global incidence and the persistence of schizophrenia despite extremely variable environmental factors, severely reduced fecundity and increased mortality. Our findings are an important step toward understanding the pathogenesis of the disease and emphasize the challenge in determining the neural substrates that these diverse genetic risk factors converge upon to generate a common pattern of clinical dysfunction and symptoms[23,31–34].

URLs. Picard, http://picard.sourceforge.net/; SAMtools, http://samtools.sourceforge.net/; PolyPhen-2, http://genetics.bwh.harvard.edu/pph2/; UCSC Table Browser, http://genome.ucsc.edu/cgi-bin/hgTables; The Human Splicing Finder (HSF, Version 2.4.1) software, http://www.umd.be/HSF/; R, http://www.r-project.org/; dbSNP v132, ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606/VCF/v4.0/00-All.vcf.gz; GATK VCF annotation file for hg19, ftp://gatk-ftp:PH5UH7Pa@ftp.broadinstitute.org/refseq/ ; Dindel, http://sites.google.com/site/keesalbers/soft/dindel.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. Reference sequences are available from NCBI under the following accession codes: *PLCL2*, NM_001144382; *WDR11*, NM_018117; *DPYD*, NM_000110; *OR4C46*, NM_001004703; *UGT1A3*, NM_019093; *FAM3D*, NM_138805; *KLF12*, NM_007249; *ADCY7*, NM_001114; *GPR153*, NM_207370; *PML*, NM_002675; *SLC26A8*, NM_052961; *CCDC108*, NM_152389; *TRAK1*, NM_001042646; *FASTKD5*, NM_021826; *DGCR2*, NM_005137; *ACOT6*, NM_001037162; *PITPNM1*, NM_001130848; *NPRL2*, NM_006545; *MAGEC1*, NM_005462; *TRRAP*, NM_003496; *COL3A1*, NM_000090; *GIF*, NM_005142; *TEKT5*, NM_144674; *THBS1*, NM_003246; *PAG1*, NM_018440; *RGS12*, NM_002926; *SAP30BP*, NM_013260; *ZNF530*, NM_020880; *MTOR*, NM_004958; *INPP5A*, NM_005539; *EDEM2*, NM_001145025; *CELF2*, NM_001083591; *SLC26A7*, NM_134266; *VPS35*, NM_018206; *ADAMTS3*, NM_014243; *GPR115*, NM_153838; *SPATA5*, NM_145207; *RB1CC1*, NM_014781; *LAMA2*, NM_000426; *ESAM*, NM_138961.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank all the families who participated in this research. We also thank H. Pretorius and nursing sisters R. van Wyk, C. Botha and H. van den Berg for their assistance with subject recruitment, family history assessments and diagnostic evaluations. We thank Y. Sun for technical assistance with DNA extractions and sample preparations and J. Grun for information technology support. We also thank E. Fledderman and S. Thomas for support of the sequencing studies and M. Robinson for critical project support. This work was supported in part by National Institute of Mental Health (NIMH) grants MH061399 (to M.K.) and MH077235 (to J.A.G.) and the Lieber Center for Schizophrenia Research at Columbia University. B.X. was partially supported by a National Alliance for Research on Schizophrenia and Depression (NARSAD) Young Investigator Award.

AUTHOR CONTRIBUTIONS

B.X., J.A.G. and M.K. designed the study, interpreted the data and prepared the manuscript. B.X. developed the analysis pipeline and had the primary role in analysis and validation of sequence data. J.L.R. collected the samples and was the primary clinician on the project. S.L. and B.P. performed exome library construction, capture and sequencing. P.D. contributed to the analysis of the data. B.B. contributed to the primary sequence data analysis. S.L. supervised the sequencing project at HudsonAlpha Institute and contributed to the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Gottesman, I.I. & Shields, J. A polygenic theory of schizophrenia. *Proc. Natl. Acad. Sci. USA* **58**, 199–205 (1967).
2. Sullivan, P.F., Kendler, K.S. & Neale, M.C. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* **60**, 1187–1192 (2003).
3. Lichtenstein, P. *et al.* Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* **373**, 234–239 (2009).
4. Lupski, J.R. Genomic rearrangements and sporadic disease. *Nat. Genet.* **39**, S43–S47 (2007).
5. Karayiorgou, M. *et al.* Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11. *Proc. Natl. Acad. Sci. USA* **92**, 7612–7616 (1995).

6. Xu, B. *et al.* Strong association of *de novo* copy number mutations with sporadic schizophrenia. *Nat. Genet.* **40**, 880–885 (2008).
7. Cirulli, E.T. & Goldstein, D.B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **11**, 415–425 (2010).
8. O'Roak, B.J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. *Nat. Genet.* **43**, 585–589 (2011).
9. Vissers, L.E. *et al.* A *de novo* paradigm for mental retardation. *Nat. Genet.* **42**, 1109–1112 (2010).
10. Awadalla, P. *et al.* Direct measure of the *de novo* mutation rate in autism and schizophrenia cohorts. *Am. J. Hum. Genet.* **87**, 316–324 (2010).
11. Abecasis, G.R. *et al.* Genomewide scan in families with schizophrenia from the founder population of Afrikaners reveals evidence for linkage and uniparental disomy on chromosome 1. *Am. J. Hum. Genet.* **74**, 403–417 (2004).
12. Karayiorgou, M. *et al.* Phenotypic characterization and genealogical tracing in an Afrikaner schizophrenia database. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **124B**, 20–28 (2004).
13. Xu, B. *et al.* Elucidating the genetic architecture of familial schizophrenia using rare copy number variant and linkage scans. *Proc. Natl. Acad. Sci. USA* **106**, 16746–16751 (2009).
14. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
15. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
16. 1000 Genomes Project Consortium. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
17. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. USA* **107**, 961–968 (2010).
18. Li, Y. *et al.* Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.* **42**, 969–972 (2010).
19. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33** (suppl.), 228–237 (2003).
20. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
21. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864 (1974).
22. Kajiwara, K. *et al.* Cloning of *SEZ-12* encoding seizure-related and membrane-bound adhesion protein. *Biochem. Biophys. Res. Commun.* **222**, 144–148 (1996).
23. Karayiorgou, M., Simon, T.J. & Gogos, J.A. 22q11.2 microdeletions: linking DNA structural variation to brain dysfunction and schizophrenia. *Nat. Rev. Neurosci.* **11**, 402–416 (2010).
24. Bjarnadóttir, T.K. *et al.* The human and mouse repertoire of the adhesion family of G-protein-coupled receptors. *Genomics* **84**, 23–33 (2004).
25. Gloriam, D.E., Schioth, H.B. & Fredriksson, R. Nine new human Rhodopsin family G-protein coupled receptors: identification, sequence characterisation and evolutionary relationship. *Biochim. Biophys. Acta* **1722**, 235–246 (2005).
26. Cruz, M.T. *et al.* Type 7 adenylyl cyclase is involved in the ethanol and CRF sensitivity of GABAergic synapses in mouse central amygdala. *Front. Neurosci.* **4**, 207 (2011).
27. Vacic, V. *et al.* Duplications of the neuropeptide receptor gene *VIPR2* confer significant risk for schizophrenia. *Nature* **471**, 499–503 (2011).
28. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).
29. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237–241 (2008).
30. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539–543 (2008).
31. Fénelon, K. *et al.* Deficiency of *Dgcr8*, a gene disrupted by the 22q11.2 microdeletion, results in altered short-term plasticity in the prefrontal cortex. *Proc. Natl. Acad. Sci. USA* **108**, 4447–4452 (2011).
32. Sigurdsson, T., Stark, K.L., Karayiorgou, M., Gogos, J.A. & Gordon, J.A. Impaired hippocampal-prefrontal synchrony in a genetic mouse model of schizophrenia. *Nature* **464**, 763–767 (2010).
33. Arguello, P.A. & Gogos, J.A. Cognition in mouse models of schizophrenia susceptibility genes. *Schizophr. Bull.* **36**, 289–300 (2010).
34. Arguello, P.A. & Gogos, J.A. Modeling madness in mice: one piece at a time. *Neuron* **52**, 179–196 (2006).

ONLINE METHODS

Cohorts. All 53 schizophrenia families were recruited from the Afrikaner population in South Africa. Heritage was established by surname and by having four Afrikaans-speaking grandparents. Informed consent was obtained from all participants. The institutional review committees of Columbia University and the University of Pretoria approved all procedures. Diagnostic evaluations were done in person as previously described[6,13]. Family history was obtained from the proband, each participating parent and additional relatives as needed, by two independent raters, a nursing sister, who recorded pedigree information, and by the clinical interviewer, who inquired in detail about family history during the clinical interview[6,13]. For additional cohort characteristics, see the **Supplementary Note**. The control cohort consisted of 22 families (trios) with established Afrikaner heritage recruited from the Afrikaner community. Paternity and maternity were confirmed before sequencing for all case and control families using the Affymetrix Genome-Wide Human SNP Array 5.0 (refs. 6,13) as well as with a panel of microsatellite markers.

Exome library construction. Exome enrichment was conducted using the SureSelect Human All Exon Target Enrichment System (Agilent Technologies) as described[35]. Briefly, 3 μg of genomic DNA was fragmented by sonication using the Covaris S2 to achieve a uniform distribution of fragments with a mean size of 300 bp. The sonicated DNA was purified using Agencourt's AMPure XP Solid Phase Reversible Immobilization paramagnetic bead (SPRI) followed by polishing of the DNA ends by removing the 3′ overhangs and filling in the 5′ overhangs resulting from sonication using T4 DNA polymerase and Klenow fragment (New England Biolabs). Following end polishing, a single A base was added to the 3′ end of the DNA fragments using Klenow fragment (3′ to 5′ exo minus). This prepared the DNA fragments for ligation to specialized adaptors that have a T-base overhang at their 3′ ends. The end-repaired DNA with a single A-base overhang was ligated to the Illumina paired-end adaptors in a standard ligation reaction using T4 DNA ligase and 2–4 μM final adaptor concentration, depending on the DNA yield following purification after the addition of the A base. Following ligation, the samples were purified using SPRI beads, quality controlled by assessment on the Agilent Bioanalyzer and then amplified by six cycles of PCR to maintain complexity and avoid bias caused by amplification.

Library capture and sequencing. We prepared 500 ng of amplified, purified DNA (DNA library) for hybridization by adding the DNA library to Agilent blocking reagents, denaturing at 95 °C and incubating at 65 °C. All subsequent steps were performed at 65 °C. Hybridization buffer was added to the prepared library, and the entire mix was then added to an aliquot of the Agilent SureSelect Capture Library and mixed. The DNA library and biotin-labeled capture library were hybridized by incubation at 65 °C for 24 h. Following hybridization, streptavidin-coated magnetic beads were used to purify the RNA:DNA hybrids formed during hybridization. The RNA capture material was digested by acid hydrolysis following elution from the purification beads. The neutralized captured DNA was purified, desalted and amplified by 12 cycles of PCR using Herculase II Fusion DNA polymerase. The libraries were purified following amplification, and the library was assessed using the Agilent Bioanalyzer. A single peak between 300–400 bp indicates a properly constructed and amplified library ready for sequencing. Final quantitation of the library was performed using the Kapa Biosciences Real-time PCR assay, and appropriate amounts of the library were loaded onto the Illumina flowcell for sequencing by paired-end 50 nt sequencing on the Illumina HiSequation (2000) instrument. Sequencing was performed largely as described[16]. Following dilution to 10 nM final concentration based on the real-time PCR and bioanalyzer results, the final library stock was then used in paired-end cluster generation at a final concentration of 6–8 pM to achieve a cluster density of 600,000/mm^2 (on the Illumina HiSequation (2000) instrument). Following cluster generation, 50 nt paired-end sequencing was performed using the standard Illumina protocols.

Exome data analysis for *de novo* coding point mutations, indels and splice site mutations. Raw sequencing data for each individual were mapped to the human reference genome (build hg19) using the Burrows-Wheeler Aligner (BWA v0.5.81536)[14]. The BWA aligned sequencing reads were processed by Picard (see URLs) to label the PCR duplicates. The Genome Analysis Toolkit (GATK, version 5091) was then used to remove duplicates, perform local realignment and map quality score recalibration to produce a 'cleaned' BAM file for each individual. SNP calls were made by the Unified Genotyper module in GATK using the 'cleaned' BAM files in batch fashion (90 samples per batch). The resulting Variant Call Format (VCF, version 4.0) files were annotated using the GenomicAnnotator module in GATK to identify and label the called variants that are within the targeted coding regions and overlap with known and likely benign SNPs reported in dbSNP v132 (see URLs). The annotated VCF files were then filtered using the GATK variant filter module with a hard filter setting and a custom script for initial filtering. Variant calls that failed to pass the following filters were eliminated from the call set: (i) MQ0 > = 4 && ((MQ0 / (1.0 * DP)) > 0.1); (ii) QUAL < 30.0 || QD < 5.0 || HRun > 5 || SB > 0.00; (iii) cluster size 10; (iv) contain dbSNP id; and (v) outside the targeted regions. Combined VCF files were then split into individual files, and variants in each offspring were compared to variants present in their parents using a custom script pipeline in order to determine the inheritance pattern and annotate *de novo* mutations.

Because the GATK Unified Genotyper is set to maximize the sensitivity of variant calls, it allows for a substantial portion of false positives among candidate variants even following the initial filtering process. To address this issue and eliminate potential false positive calls in the offspring and false negative calls in the parents, we took advantage of the inheritance information provided by our family design and revalidated all variants identified using the mpileup module in the SAM tools (see URLs) according to the following rules: (i) the forward reference (fr) count (the number of forward reads that match the reference base at this locus), the reverse reference (rr) count (the number of reverse reads that match the reference base at this locus), the forward non-reference (fnr) count (the number of forward reads that do not match the reference base at this locus) and the reverse non-reference (rnf) count (the number of reverse reads that do not match the reference base at this locus) in the offspring must be two or greater; (ii) total read depth in both parents must be ten or greater; (iii) both fr and rr count in both parents must be two or greater; (iv) either fnr or rnr count in both parents must be zero; (v) the fnr and rnr count to total count ratio in the parental population (defined as all 150 parental samples sequenced) must be less than $1/2n$, where n is the population size; and (vi) if any of rules i–v was violated, the sequence information was considered insufficient to make a *de novo* call at this locus.

Indel calls were made by the Dindel software using one 'cleaned' BAM file per run. The resulting VCF files were used to determine inheritance patterns using the same procedure described above for the point mutations. To determine potential mutations at splice-donor or acceptor sites, GATK variant calls were made in a batch fashion (with 90 samples per batch) that covered each target coding region and 50-bp flanking segments in each direction. The variants in the resulting VCF files were annotated according to the GATK annotation file for hg19 (see URLs).

The PolyPhen-2 (see URLs) online server was used to determine the non-synonymous and synonymous nature of the mutations and to predict their functional impact by further classifying them as non-tolerated (damaging) or benign at a given site. The Grantham score for each coding variant was determined by the Grantham matrix table[21]. The phyloP score for each coding variant was extracted from the 'phyloP46wayAll' table in the UCSC Table Browser (see URLs). The Human Splicing Finder (HSF, Version 2.4.1) software (see URLs) was used to predict potential functional impact of the mutations at splice sites.

Statistics. The Kolmogorov-Smirnov test was used to compare the distribution of phyloP and Grantham scores among *de novo* or private inherited mutations in cases. The Kolmogorov-Smirnov test was conducted using R (see URLs). Fisher's exact test or χ^2 test with Yates' correction was used for the analysis of contingency tables, depending on the sample sizes.

***De novo* mutation validation.** Candidate *de novo* variants were tested using standard Sanger sequencing on an ABI 3730xl DNA Analyzer to validate presence of each mutation in the subjects and the absence of each in the parental genomes, by designing custom primers (Sigma) based on ~500 bp of genomic sequence flanking each variant. *De novo* occurrence of mutations was not confirmed for 6 out of 46 candidate alterations in cases and 2 out of 9 candidate alterations in controls.

35. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).