De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia

Bin Xu1, Juliana Ionita-Laza2, J Louw Roos3,4, Braden Boone5, Scarlet Woodrick1,6, Yan Sun1, Shawn Levy5, Joseph A Gogos6,7 & Maria Karayiorgou1

To evaluate evidence for de novo etiologies in schizophrenia, we sequenced at high coverage the exomes of families recruited from two populations with distinct demographic structures and history. We sequenced a total of 795 exomes from 231 parent-proband trios enriched for sporadic schizophrenia cases, as well as 34 unaffected trios. We observed in cases an excess of de novo nonsynonymous single-nucleotide variants as well as a higher prevalence of gene-disruptive de novo mutations relative to controls. We found four genes (LAMA2, DPYD, TRRAP and VPS39) affected by recurrent de novo events within or across the two populations, which is unlikely to have occurred by chance. We show that de novo mutations affect genes with diverse functions and developmental profiles, but we also find a substantial contribution of mutations in genes with higher expression in early fetal life. Our results help define the genomic and neural architecture of schizophrenia.

Schizophrenia is a severe psychiatric disorder with a strong genetic component1. Whereas the contribution of rare de novo copy-number variants (CNVs) to schizophrenia risk is well established2–5, the contribution of de novo nucleotide-level variants has not yet been probed extensively6,7. We completed exome sequencing of 146 Africanan proband-parent family trios of subjects with a diagnosis of schizophrenia or schizoaffective disorder (SCZAFF)3,8 and 34 unaffected control trios (53 and 22 trios, respectively, have been described previously7). We also sequenced the exomes of 85 US family trios of subjects with schizophrenia or SCZAFF (Online Methods and Supplementary Table 1). None of the subjects carry rare de novo CNVs (of ≥30 kb in size), as determined in previous CNV scans of these cohorts3,4. We used a previously described analytical pipeline2 and a series of filters, including final validation by Sanger sequencing of all family members (Online Methods and Supplementary Fig. 1). More than 90% of single-nucleotide variants (SNVs) and 20% of insertions and/or deletions (indels) were validated (Supplementary Table 2).

In our control cohort, we identified 16 exonic de novo SNVs and 1 protein-truncating indel in 34 subjects (0.50 events per sample) (Table 1). The point mutation rate in the captured coding sequence was 1.28 × 10−8 mutations per base per generation. Among the 16 de novo SNVs, 11 were predicted to be nonsynonymous, missense changes, and 5 were predicted to be synonymous changes. The observed nonsynonymous-to-synonymous ratio in our control cohort of 2.20 is consistent with neutral expectation (2.23)9,10 and with those reported for control samples (unaffected siblings) in three studies of independent subsets of the Simons Simplex Collection (SSC) (2.23, n = 200 (ref. 11); 2.11, n = 31 (ref. 12); 2.99, n = 343 (ref. 13); average nonsynonymous-to-synonymous ratio = 2.65, n = 574 (combined SSC control group)).

In the 146 Africanan probands, we observed 93 confirmed de novo exonic point mutations (92 SNVs and 1 dinucleotide substitution) and 9 confirmed de novo indel events (Table 1). Six of the indels resulted in protein truncation, and three resulted in deletion of a single amino acid. Additional query of de novo SNVs located within the flanking intronic regions identified three SNVs that altered canonical splice sites (Table 1) and seven that altered the consensus sequences flanking canonical splice sites (Online Methods). Overall, 73 of 146 cases (50%) carried at least 1 likely functional de novo event (nonsynonymous, indel or splice-site mutation). The point mutation rate in the captured coding sequence was 1.73 × 10−8 mutations per base per generation, not significantly different than the one observed in our control sample. Moreover, we found no differences in the distribution or frequency of multiple de novo point mutations in cases versus controls (Supplementary Fig. 2). The 93 identified de novo point mutations included 80 nonsynonymous and 13 synonymous changes. The nonsynonymous-to-synonymous ratio of 6.15 was higher than neutral expectation (2.23)9,10 (P = 1.92 × 10−5, two-sided exact binomial test). To further assess the statistical significance of the observed enrichment of nonsynonymous variants in cases, we performed permutation testing by randomly permuting the case-control labels of the trios in our data set (Supplementary Note). On the basis of 100,000 such permuted data sets, we obtained permutation one-sided P = 0.033. In contrast, analysis of the nonsynonymous-to-synonymous ratios among private inherited variants (variants present only in one family and serving as proxies for evolutionarily young mutation events) did not show any significant difference in cases versus controls (Supplementary Table 3). Indeed, de novo variants were
npg

overlap between either all genes with carrying mutations in genes with no fetal brain expression bias; 1.35-fold enrichment, found among carriers of variants include missense and nonsense SNVs; functional variants include nonsynonymous, indel and splice-site mutations. Indel data were not available
deo novo previously described global expression switch occurring before birth. We determined the relative enrichment of nonsynonymous or, more generally, functional de novo variants over synonymous ones. The relative enrichment in differentially regulated mutated genes is shown for de novo mutations. Nonsynonymous variants include missense and nonsense SNVs; functional variants include nonsynonymous, indel and splice-site mutations. Indel data were not available for transmitted variants. (b) Left, an elevated frequency of cases who had multiple (≥3) childhood behavioral abnormalities (Supplementary Note) was found among carriers of de novo functional mutations in prenatally biased genes (10/29 compared to 4/51 cases carrying mutations in genes with no bias toward fetal brain expression; 4.4-fold enrichment, \( P = 0.0047 \), Fisher’s exact test). Right, an elevated frequency of cases with severe disease functional outcome (Supplementary Note) was found among carriers of de novo functional mutations in prenatally biased genes (35/38 compared to 45/66 cases carrying mutations in genes with no fetal brain expression bias; 1.35-fold enrichment, \( P = 0.007 \), Fisher’s exact test). (c) Venn diagrams depicting the overlap between either all genes with de novo mutations in schizophrenia (\( n = 145 \)) or only genes with prenatal expression bias harboring functional de novo mutations identified in the Afrikaner and US probands (\( n = 59 \)) and genes with de novo mutations identified in ASD exome scans\(^{15,13} \) (\( n = 675 \)).

Analysis of the US cohort revealed a point mutation rate of 1.73 × 10\(^{-9}\) mutations per base per generation, similar to that observed in the Afrikaner sample. The nonsynonymous-to-synonymous ratio of 3.42 (Table 1) is higher than neutral expectation (2.23\(^{10,19}\), but the difference is not statistically significant. In addition to the smaller size of the US cohort, this is likely due to greater uncertainty about family history of schizophrenia in that cohort, meaning that it may not be entirely depleted of familial cases, consistent with a lower de novo CNV rate\(^4\) compared to the Afrikaner cohort\(^2\). By contrast, when comparing the rate of loss-of-function de novo events per case (8/85; 9.4%), we observed a 3.2-fold enrichment relative to controls, similar to the enrichment observed in the Afrikaner sample.

In the combined sample of 231 affected families, the nonsynonymous-to-synonymous ratio for de novo variants (4.84; Table 1) remained higher compared to neutral expectation\(^9\) (\( P = 2.1 \times 10^{-4} \), two-sided exact binomial test). There was a differential enrichment of nonsynonymous events between de novo and rare, private inherited events (\( P < 0.0001 \)). Notably, as in both individual cohorts, the number of loss-of-function events per affected trio (20/231; 8.7%) was 2.8 times higher than in control trios (1/34, 2.9%), strongly supporting a role for gene-disrupting mutations in disease (Tables 1 and 2). Permutation testing showed a significant enrichment of likely functional de novo variants over synonymous ones (permutation one-sided \( P = 0.026 \) (Supplementary Note). We estimate that 46% of all likely functional de novo mutations identified represent genuine risk variants (Supplementary Note).

Considering phenotypic correlates, we observed a correlation between paternal age at the proband’s birth and the number of de novo events per offspring (Supplementary Fig. 3) but did not find any other significant differences between proband carriers and non-carriers of de novo functional mutations (Supplementary Tables 4 and 5 and Supplementary Note).

Genes affected by de novo variants in our study were not significantly over-represented in two previously established comprehensive lists of synaptic genes that encode proteins localized to the presynaptic terminus or postsynaptic density of neurons (Supplementary Note). In addition, pathway analyses using the Database for Annotation, Visualization and Integrated Discovery (DAVID) Annotation Tool did not identify any significantly enriched functional clusters. Evaluation of protein-protein
interactions showed a significantly greater connectivity among mutational targets than would be expected by chance (P < 0.05). Multigene protein-protein interaction clusters included one centered on MTOR and one centered on CANX, which includes genes encoding extracellular matrix and cell adhesion proteins (Supplementary Fig. 4), suggesting that diverse genes associated with schizophrenia risk may converge on a shorter list of functional modules.

We then examined to what extent the enrichment in functional de novo events is determined by the developmental pattern of brain expression of the mutated genes

| Variant type                  | Afrikaner cases n = 146 | US cases n = 85 | Total cases n = 231 | Afrikaner controls n = 34 |
|--------------------------------|-------------------------|-----------------|---------------------|--------------------------|
| SNVs                           | 93 (0.64)               | 53 (0.62)       | 146 (0.63)          | 16 (0.47)                |
| Nonsynonymous                  | 80 (0.55)               | 41 (0.48)       | 121 (0.52)          | 11 (0.32)                |
| Synonymous                     | 13 (0.09)               | 12 (0.14)       | 25 (0.11)           | 5 (0.15)                 |
| Nonsynonymous-to-synonymous ratio | 6.15                   | 3.42            | 4.84                | 2.20                     |
| Nonsense                       | 2 (0.01)                | 4 (0.05)        | 6 (0.03)            | 0 (0.00)                 |
| Canonical splice site          | 3 (0.02)                | 2 (0.02)        | 5 (0.02)            | 0 (0.00)                 |
| Splice consensus               | 7 (0.05)                | 4 (0.05)        | 11 (0.05)           | 1 (0.03)                 |
| All de novo indels             | 9 (0.06)                | 4 (0.05)        | 13 (0.06)           | 1 (0.03)                 |
| No frameshift                  | 3 (0.02)                | 1 (0.01)        | 4 (0.02)            | 0 (0.00)                 |
| Frameshift                     | 6 (0.04)                | 3 (0.04)        | 9 (0.04)            | 1 (0.03)                 |
| All loss of function           | 11 (0.08)               | 9 (0.11)        | 20 (0.09)           | 1 (0.03)                 |
| All likely functional          | 99 (0.68)               | 51 (0.6)        | 150 (0.65)          | 13 (0.38)                |
| Functional-to-synonymous ratio | 7.62                    | 4.25            | 6                   | 2.6                      |

The majority of mutated genes with prenatal expression bias are highly expressed during the first and second trimesters of pregnancy (Supplementary Fig. 7) and show an over-representation of genes encoding nuclear factors involved in chromatin remodeling, nuclear transport and transcriptional control, as well as in protein translation and degradation (Supplementary Tables 6 and 8). They also include genes that code for factors involved in cell-cell and cell-matrix interactions, a subset of which interact with two key adhesive proteins, THBS1 (ref. 18) and ITGA6 (ref. 19), involved in synaptogenesis, axonal growth and cortical layering (Supplementary Fig. 4). We confirmed that genes with prenatal expression bias were highly enriched for genes that are targeted by one or more microRNAs (miRNAs) (Supplementary Note) and also found a nominally significant enrichment of hsa-mir-367 and hsa-mir-1244 targets (Supplementary Tables 9 and 10). By comparison, the risk conferred by mutated genes with bias toward postnatal expression was related to

Table 2 Loss-of-function mutations in schizophrenia probands

| ID     | Schizophrenia cohort | Mutation type | Gene symbol | Loss-of-function mutations in SA/SSC controls |
|--------|----------------------|---------------|-------------|---------------------------------------------|
| trio_090 | US                   | Frameshift    | XPR1        | −                                           |
| trio_107 | US                   | Frameshift    | CCDC39      | +a                                         |
| trio_121 | US                   | Frameshift    | KDM5C       | +a                                         |
| trio_005 | Afrikaner            | Frameshift    | KIAA0467    | −                                           |
| trio_100 | Afrikaner            | Frameshift    | HIST1H1E    | −                                           |
| trio_026 | Afrikaner            | Frameshift    | RB1CC1      | −                                           |
| trio_042 | Afrikaner            | Frameshift    | ESAM        | −                                           |
| trio_092 | Afrikaner            | Frameshift    | LAMA2       | −                                           |
| trio_027 | Afrikaner            | Frameshift    | DDHD2       | −                                           |
| trio_101 | US                   | Nonsense      | SSBP3       | −                                           |
| trio_118 | US                   | Nonsense      | NUP54       | −                                           |
| trio_124 | US                   | Nonsense      | DPYD        | −                                           |
| trio_128 | US                   | Nonsense      | STAP2       | −                                           |
| trio_053 | Afrikaner            | Nonsense      | URB2        | −                                           |
| trio_085 | Afrikaner            | Nonsense      | RARG        | −                                           |
| trio_018 | Afrikaner            | Splice sitea  | SYNGAP1      | −                                           |
| trio_072 | Afrikaner            | Splice siteb  | BRRP1       | −                                           |
| trio_111 | US                   | Splice siteb  | PROX6       | −                                           |
| trio_103 | US                   | Splice siteb  | NLRC5       | −                                           |
| trio_016 | Afrikaner            | Splice siteb  | CUGBP2      | −                                           |

aFrom ref. 13. bCanonical splice site.
their involvement in intracellular signaling processes (GTPase, diacylglycerol (DAG) or calcium signaling; Supplementary Table 8), which regulate diverse aspects of neuronal connectivity.

We next set out to evaluate which of the individual genes that are mutated in schizophrenia are more likely to confer disease risk. We found four genes altered by two de novo events each in unrelated probands (three of these genes were affected in cases across the two different populations tested) (Table 3 and Supplementary Note). These genes were mutated twice by a combination of a nonsense and a missense de novo SNV (DPYD) or a combination of a splice-site mutation (Supplementary Tables 11 and 12) with either a de novo missense SNV (TRRAP and VPS39) or an indel (LAMA2). None of these genes were affected in the Afrikaner or the SSC control group, and no such mutational combinations were reported for any gene among the 488 with de novo events in the SSC control group. Given the number of de novo mutations in our data set, observation of four such recurrent events is associated with a P value of 0.002 (Supplementary Note), with the occurrence of one of the four combinations (in LAMA2) being individually statistically significant (P = 0.017).

LAMA2 encodes the laminin γ2 chain, which constitutes one of the subunits of laminin 2 and 4 and binds to ITGα6, which is encoded by a gene with prenatal expression bias that was mutated in our cases (Supplementary Fig. 4). The indel in LAMA2 disrupts a key C-terminal domain, and the splice-site mutation affects a highly conserved nucleotide at position –1 relative to the canonical splice acceptor AG motif and is expected to disrupt splicing (Supplementary Table 12). Homozygous mutations in LAMA2 lead, with variable penetrance, to congenital muscular dystrophy characterized by central nervous system (CNS) involvement, including white matter abnormalities, cognitive impairment, seizures and neuronal migration defects. In addition, a de novo mutation in another member of the laminin gene family, LAMA1, was described in another schizophrenia cohort.

DPYD encodes dihydropyrimidine dehydrogenase, the initial and rate-limiting factor in the pathway of pyrimidine catabolism that also modulates production of β-alanine, a neuromodulator of inhibitory transmission in the brain. We identified one missense and one nonsense SNV in the Afrikaner and US cohorts, respectively. Abnormal urinary excretion of thymine and uracil confirmed DPYD deficiency in the carrier of the missense mutation (Supplementary Fig. 8). Heterozygous deletions, either encompassing or within DPYD, as well as altered expression, have been described in ASD and intellectual disability. Neither autistic features nor intellectual disability were present in carriers of the DPYD mutation in our cohorts, suggesting variable expressivity. A genome-wide association study (GWAS) mega-analysis in schizophrenia identified the strongest association (P = 1.6 × 10^{-11}) at rs1625579, a variant located at 1p21.3 in the intron of MIR137 (ref. 26) and within a haplotype block (D’ > 0.9) that extends to the 5’ exons of DPYD (Supplementary Fig. 9). We did not find mutations in MIR137; therefore, association with rs1625579 may reflect the contribution of DPYD variants to disease.

The splice-site mutation in TRRAP affects position +6 relative to the splice junction site and is predicted to disrupt splicing and binding of the SRP55 splicing factor within a splicing enhancer. Notably, TRRAP and VPS39 are also mutated in ASD cases.

We also compared the identified functional de novo mutations to the de novo CNVs identified previously in our two cohorts (22 CNVs affecting 156 genes). Five genes (DGCR2, TOP3B, CIT, STAG1 and SMAP2) were altered by both de novo SNVs and CNVs (Table 3), two of them in affected individuals across the two different populations tested. Two of these genes are within the 22q11.2 schizophrenia susceptibility locus.

Our findings implicate contribution from a diverse set of de novo mutations to the genomic architecture of schizophrenia in the context of a mutation-selection balance model and highlight the importance of using family samples where disease history has been thoroughly ascertained. Focusing on our comprehensively ascertained Afrikaner cohort, we estimate that at least 17.6% of sporadic cases carry a de novo pathogenic exonic mutation (Supplementary Note), and at least 9.9% carry a de novo CNV. Thus, such mutations may contribute to risk in approximately one-fourth to one-third of all sporadic cases. Given that results from scans of non-exonic

Table 3 Genes altered by recurrent de novo events

| Sample ID | Gene symbol | Locus | Mutation type | DNA change | RNA or amino-acid change | Cohort | Gene name |
|-----------|-------------|-------|---------------|------------|--------------------------|--------|-----------|
| trio_124  | DPYD        |       | Nonsense      | c.1863G>A  | p.Trp621*                | US     | Dihydropyrimidine dehydrogenase |
| trio_016  | DPYD        | 1p21.3| Missense      | c.1615G>A  | p.Gly539Arg              | Afrikaner |          |
| trio_092  | LAMA2       |       | Frameshift    | c.9139_9146del7 | p.Ser305Thr*27       | Afrikaner |          |
| trio_049  | LAMA2       | 6q22.13| Splice site  | c.4718−3   | Splice-site mutation    | Afrikaner |          |
| trio_033  | TRRAP       |       | Missense      | c.883A>T   | p.Ile295Phe              | Afrikaner |          |
| trio_120  | VPS39       |       | Missense      | c.7223.+6   | Splice-site mutation    | US     | Transformation/transcription domain-associated protein |
| trio_125  | VPS39       | 15q15.1| Splice site  | c.441+8    | Splice-site mutation    | US     | Vacular protein sorting 39 homolog (S. cerevisiae) |

Genes altered by both de novo SNVs and de novo CNVs

| Sample ID | Gene symbol | Locus | Mutation type | DNA change | RNA or amino-acid change | Cohort | Gene name |
|-----------|-------------|-------|---------------|------------|--------------------------|--------|-----------|
| trio_091  | DGCR2       | 22q11.2| Missense      | c.1163C>G  | p.Pro388Arg               | Afrikaner | DiGeorge syndrome critical region gene 2 |
| trio_064  | TOP3B       | 22q11.2| Missense      | c.1415G>A  | p.Arg472Gln              | Afrikaner | Topoisomerase (DNA) III α |
| trio_121  | CIT         | 12q24.23| Missense     | c.238T>C   | p.Tyr80His                | US     | Citron (rho-interacting, serine/threonine kinase 21) |
| trio_111  | STAG1       | 3q22.3 | Missense      | c.667A>T   | p.Thr223Ser              | US     | Stomal antigen 1 |
| trio_078  | SMAP2       | 1p34.2 | Missense      | c.896G>A   | p.Ser299Asn               | Afrikaner |          |

*Consensus splice-site mutation. Mutations in LAMA2 and TRRAP are predicted to be damaging (Supplementary Tables 11 and 12). From ref. 3. Intragenic duplication. Del, deletion; dup, duplication.
regions are still forthcoming, this is likely an underestimate. Our find-
ing also contribute to the understanding of the neural architecture of schizophrenia risk. Given that we estimate the number of schizophrenia risk loci at more than 850 (Supplementary Note), our findings imply an exquisite sensitivity of the neural circuits underlying susceptibility to schizophrenia to precise levels or activity of many diverse proteins and signaling modules and suggest that focusing on circuits may be more com-
mensurate with the heterogeneity of schizophrenia than other proposed mechanisms that concentrate on specific neurotransmitters or cell types37.

In addition, we show that, in determining disease risk, not only the func-
tion of the mutated gene but also the timing of the genetic insult may be of critical importance. Specifically, although de novo mutations affect genes with diverse functions and developmental profiles, we describe a substantial contribution of mutations in developmentally regulated genes with higher expression during early and mid-stage fetal life and show that such mutations are enriched among adult cases with prominent early pre-
psychotic, deviant behaviors. Notably, a bias towards prenatally expressed genes was also demonstrated for genes affected by multiple types of schizophrenia-associated genetic variation33. Our findings provide a mechanistic context to interpret epidemiological correlations among various prenatal environmental insults during the first and second trimesters of pregnancy and risk for schizophrenia34. Moreover, the fact that expression of many prenatally biased genes is under strict miRNA control may explain emerging links between miRNA dysregulation and psychiatric disorders30. The challenge remains to identify the affected biological processes and neural circuits and to determine how they are affected. Unbiased network-based approaches as well as animal and cellular models of recurrent mutations will be invaluable in reaching this goal31.

**METHODS**

Methods and any associated references are available in the online version of the paper.

**Accession codes.** Exome sequencing data are available upon request.

**Note:** Supplementary information is available in the online version of the paper.

**ACKNOWLEDGMENTS**

We are enormously grateful to all the families who participated in this research. We thank H. Pretorius and nursing sisters R. van Wyk, C. Botha and H. van Berg for their assistance with subject recruitment, family history assessments and diagnostic evaluations. We thank S.L. Lundy for valuable assistance with clinical database maintenance and L. Rodriguez-Murillo for help with Supplementary Figure 9. We also thank B. Plummer and M. Robinson and the HudsonAlpha Genomics Services Laboratory for experimental support. Finally, we thank L.I.J. Mienie for the thymine loading test. This work was partially supported by National Institute of Mental Health (NIMH) grants MH061399 (to M.K.) and MH077235 (to I.A.G.) and the Lieber Center for Schizophrenia Research at Columbia University. B.X. was partially supported by a National Alliance for Research in 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 the analysis and validation of sequence data. I.I.-L. performed statistical analysis of the sequence data. I.L.R. contributed to sample collection and clinical characterization. S.W. and Y.S. contributed to sample preparation and de novo mutation validation. B.B. performed exome library construction, capture and sequencing and initial analysis of SNV genotyping and indel variant calls. S.L. supervised the sequencing project at the HudsonAlpha Institute.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.

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

1. Rodríguez-Murillo, L., Gogos, J.A. & Karayiorgou, M. The genetic architecture of schizo-
phrenia: new mutations and emerging paradigms. Annu. Rev. Med. 63, 63–80 (2012).

2. Willemsen, M.H. et al. Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11. Proc. Natl. Acad. Sci. USA 92, 7612–7616 (1995).

3. Xu, X. et al. Strong association of de novo copy number mutations with sporadic schizophrenia. Nat. Genet. 40, 880–885 (2008).

4. Neale, B.M. et al. Genome-wide association study identifies five new schizophrenia loci in bipolar disorder and schizophre-

5. Neale, B.M. & Krueger, R.F. Analyzing genetic variation and complex disease. Nat. Rev. Genet. 11, 381–392 (2010).

6. Malhotra, D. et al. De novo mutations in schizophrenia spectrum disorders. Nat. Genet. 40, 751–760 (2008).

7. Sanders, S.J. et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485, 242–245 (2012).

8. O’Roak, B.J. et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485, 246–250 (2012).

9. Iossifov, I. et al. High frequencies of de novo CNVs in bipolar disorder and schizo-

10. Allely, E. & O’Roak, B.J. De novo copy number variation in autism. Nat. Rev. Genet. 11, 279–292 (2010).

11. Wang, S. et al. Genetic analysis and validation of sequence data. I.I.-L. performed statistical analysis of the sequence data. I.L.R. contributed to sample collection and clinical characterization. S.W. and Y.S. contributed to sample preparation and de novo mutation validation. B.B. performed exome library construction, capture and sequencing and initial analysis of SNV genotyping and indel variant calls. S.L. supervised the sequencing project at the HudsonAlpha Institute.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.

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

1. Rodríguez-Murillo, L., Gogos, J.A. & Karayiorgou, M. The genetic architecture of schizo-
phrenia: new mutations and emerging paradigms. Annu. Rev. Med. 63, 63–80 (2012).

2. Willemsen, M.H. et al. Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11. Proc. Natl. Acad. Sci. USA 92, 7612–7616 (1995).

3. Xu, X. et al. Strong association of de novo copy number mutations with sporadic schizophrenia. Nat. Genet. 40, 880–885 (2008).

4. Neale, B.M. et al. Genome-wide association study identifies five new schizophrenia loci in bipolar disorder and schizophre-

5. Neale, B.M. & Krueger, R.F. Analyzing genetic variation and complex disease. Nat. Rev. Genet. 11, 381–392 (2010).

6. Malhotra, D. et al. De novo mutations in schizophrenia spectrum disorders. Nat. Genet. 40, 751–760 (2008).

7. Sanders, S.J. et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485, 237–241 (2012).

8. O’Roak, B.J. et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485, 246–250 (2012).

9. Iossifov, I. et al. High frequencies of de novo CNVs in bipolar disorder and schizo-

10. Allely, E. & O’Roak, B.J. De novo copy number variation in autism. Nat. Rev. Genet. 11, 279–292 (2010).

11. Wang, S. et al. Genetic analysis and validation of sequence data. I.I.-L. performed statistical analysis of the sequence data. I.L.R. contributed to sample collection and clinical characterization. S.W. and Y.S. contributed to sample preparation and de novo mutation validation. B.B. performed exome library construction, capture and sequencing and initial analysis of SNV genotyping and indel variant calls. S.L. supervised the sequencing project at the HudsonAlpha Institute.
ONLINE METHODS

Cohorts. The samples analyzed here comprise trios collected from two distinct populations: the Afrikaner population from South Africa (European, mostly Dutch descent) (146 trios with schizophrenia) and the US population (Northern European descent) (85 trios with schizophrenia). Of the 146 Afrikaner probands, 122 (83.6%) had a diagnosis of schizophrenia, and 24 (16.4%) were diagnosed with SCZAFF disorder. Of the 85 US probands, 46 (54.1%) had a diagnosis of schizophrenia, and 39 (45.9%) were diagnosed with SCZAFF disorder. The control cohort consisted of 34 trios with established Afrikaner heritage. Control families included unaffected subjects screened against presence and history of treatment for any psychiatric condition, as well as history of mental illness in first- or second-degree relatives. Both affected and control trios were recruited and characterized in the context of our ongoing large-scale genetic studies of schizophrenia and have been described previously3,7,8. Because de novo mutations are more likely to account for sporadic forms of the disease, we took great care to determine reliably and in depth the family history status and generate cohorts enriched in sporadic cases (Supplementary Note). However, the presence or absence of a family history was not used as a screening criterion.

In the Afrikaner cohort, it was possible to determine absence of disease in first- or second-degree relatives because of the cohesive family structures, the large catchment area and the fact that long-term care was provided by the in first- or second-degree relatives because of the cohesive family structures, the large catchment area and the fact that long-term care was provided by the

Relatives. de novo

The large catchment area and the fact that long-term care was provided by the in first- or second-degree relatives because of the cohesive family structures, and the presence of or absence of a family history was not used as a screening criterion.

Exome library construction. For exome capture and sequencing, genomic DNA (~3 μg) was sheared to 200–300 bp in size using a Covaris Acoustic Adapter. Fragments were then end repaired, polyA tailed and ligated to sequencing adaptor oligonucleotides using reagents from New England BioLabs. Libraries were barcoded using the Illumina index read strategy, which uses 6-base sequences within the adaptor that are sequenced separately from the genomic DNA insert. Ligated products were size selected during purification steps. The DNA library was subsequently enriched for sequences with 5’ and 3’ adaptors by PCR amplification using primers complementary to the adaptor sequences (ligation-mediated PCR, LM-PCR). Exonic DNA was captured using two hybridization systems: Agilent SureSelect v2 (n = 85 trios) and NimbleGen SeqCap EZ v2 (n = 180 trios). After capture, another round of LM-PCR was performed to generate the final libraries. Each library was quantified by fluorescent methods (PicoGreen, Invitrogen), and fragment size was measured with the Agilent Bioanalyzer. Finally, the molar concentration of each library was measured using the size information from the Agilent Bioanalyzer and DNA quantitation information from an RT-PCR assay (Kapa Biosystems; performed according to the manufacturer’s protocol). Each library was normalized to a 10 nM concentration and sequenced using an Illumina HiSeq 2000.

Exome data analysis for de novo SNVs and indels. The exome data analysis pipeline has been described previously7. Briefly, raw sequencing data were mapped to the human reference genome (hg19) using the Burrows-Wheeler Aligner (BWA v0.5.8.1536). The Genome Analysis Toolkit (GATK, version 5.0) was used to remove duplicates, perform local realignment and map quality score recalibration to produce a cleaned BAM file and then make genotype calls for all trios jointly. 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 were within the targeted coding regions and overlap with known and likely benign SNPs reported in dbSNP v132 (see URLs). The filtered genotype calls were further validated using the mpileup module in SAMtools (see URLs) as described previously7. Indel calls were made by Dindel software using one cleaned BAM file per run. The resulting VCF files were further revalidated using the same SAMtools procedure as for the point mutations. To determine potential mutations at splice donor or acceptor sites, GATK variant calls were made in a batch fashion (90 samples per batch) that covered each target coding region and the 50-bp flanking segments on each side of it. The variants in the resulting VCF files were annotated according to refGene-big-table-hg19.txt (see URLs). A variant was annotated as a ‘canonical splice-site mutation’ if it disrupted the largely invariable core canonical 2-bp acceptor (AG) or donor (GU) sites. De novo variants within 10 bp of the exon-intron boundary, included in the consensus sequence flanking core canonical splice sites and therefore likely to modulate splicing efficiency, were annotated as ‘consensus splice-site mutations’. Candidate de novo variants were tested using standard Sanger sequencing on an ABI 3730xl DNA Analyzer (Applied Biosystems) to validate the presence of each mutation in the proband and its absence in the parental genomes, using custom primers that were designed using Primer3 software on the basis of the ~500 bp of sequence flanking each variant. The total number of de novo SNVs found and validated in a given cohort was divided by the total number of bases analyzed to calculate the per-base rate of point mutation in the captured coding sequence.

Variant detection pipeline and quality control. Because the entire capture and sequencing procedure was conducted without knowledge of the affected status in all three cohorts, we expected no bias between cohorts. To further demonstrate that variant detection and quality control were consistent across all samples and all experimental conditions, we compared the percentage of average reads at 1×, 8×, 20× and 30× coverage under each condition (Supplementary Fig. 1). There were no differences for any of the parameters analyzed.

Statistics. The two-sided exact binomial test was conducted using R. Fisher’s exact test or the χ-squared test with Yates’ correction was used for the analysis of contingency tables, depending on the sample sizes, using R.

Annotation of the functional impact of the de novo mutations. The functional impact of the de novo mutations was annotated from several different resources. The PolyPhen-2 online batch query server (see URLs)23 was used with full annotation settings to determine the nonsynonymous or synonymous nature of the mutations and 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 using the Grantham matrix table33. The phyloP score for each coding variant was extracted from the phyloP46wayAll table in the UCSC Table Browser (see URLs). For splice-site variants, we considered mutations directly disrupting canonical splice sites as severe disruptive events without further analysis. For mutations in consensus splice sites, we used a mutation analysis module in the Human Splice Finder (HSF) program (Version 2.4.1; see URLs)34 to predict their functional impact. Briefly, 100 nucleotides of sequence surrounding the exon-intron boundary were extracted from the UCSC browser and the wild-type and mutated sequences were imported into the HSF mutation analysis module to detect potential disruption of splicing signals. The HSF-derived results for the identified consensus splice-site mutations are shown in Supplementary Tables 11 and 12.

Gene set enrichment and protein-protein interaction network analyses. The DAVID Functional Annotation Chart25 (see URLs) was used to determine whether a given gene set with de novo mutations was enriched in particular gene ontology (GO) terms or functional keywords defined in Swiss-Prot and the Protein Information Resource. Genes with mutations in schizophrenia were mapped in the database, and functional annotation chart analysis was conducted with the default settings. We used the Disease Association Protein-protein Link Evaluator (DAPPLE; see URLs)36 to determine whether there was a higher protein-protein interaction than expected by chance among the genes altered by likely functional de novo variants. A list of all genes with likely functional mutations was submitted to the DAPPLE server with default settings.

Temporal expression profile analysis of the genes carrying de novo mutations. To investigate developmental expression of genes mutated in
schizophrenia, we took advantage of the Human Brain Transcriptome (HBT) database (see URLs), a compendium of exon-level expression profiles across developmental stages from embryonic to late adulthood. Genes harboring de novo events were grouped into three classes (with prenatal brain-biased, postnatal brain-biased and unbiased expression) according to their temporal expression trajectory in reference to a global expression turning point occurring between the mid-late and late fetal stages. For each class, the ratio of nonsynonymous or likely functional variants to neutral variants was calculated.

32. Adzhubei, I.A. et al. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249 (2010).
33. Grantham, R. Amino acid difference formula to help explain protein evolution. Science 185, 862–864 (1974).
34. Desmet, F.O. et al. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. 37, e67 (2009).
35. Huang, D.W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4, 44–57 (2009).
36. Rossin, E.J. et al. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet. 7, e1001273 (2011).