Increased exonic *de novo* mutation rate in individuals with schizophrenia

Simon L Girard[1], Julie Gauthier[1], Anne Noreau[1], Lan Xiong[1], Sirui Zhou[1], Loubna Jouan[1], Alexandre Dionne-Laporte[1], Dan Spiegelman[1], Edouard Henrion[1], Ousmane Diallo[1], Pascale Thibodeau[1], Isabelle Bachand[2], Jessie Y J Bao[3], Amy Hin Yan Tong[3], Chi-Ho Lin[3], Bruno Millet[4,5], Nematollah Jaafari[4,6], Ridha Joober[7], Patrick A Dion[1,8], Si Lok[3], Marie-Odile Krebs[4,9,11] & Guy A Rouleau[1,2,10,11]

© 2011 Nature America, Inc. All rights reserved.

Schizophrenia is a severe psychiatric disorder that profoundly affects cognitive, behavioral and emotional processes. The wide spectrum of symptoms and clinical variability in schizophrenia suggest a complex genetic etiology, which is consistent with the numerous loci thus far identified by linkage, copy number variation and association studies[1–4]. Although schizophrenia heritability may be as high as ~80%, the genes responsible for much of this heritability remain to be identified[5]. Here we sequenced the exomes of 14 schizophrenia probands and their parents. We identified 15 *de novo* mutations (DNMs) in eight probands, which is significantly more than expected considering the previously reported DNM rate[6–8]. In addition, 4 of the 15 identified DNMs are nonsense mutations, which is more than what is expected by chance[9]. Our study supports the notion that DNMs may account for some of the heritability reported for schizophrenia while providing a list of genes possibly involved in disease pathogenesis.

We sequenced the captured product of targeted exomes prepared from 14 trios, with each trio consisting of an individual with schizophrenia and his or her parents. To avoid identifying genetic variations resulting from cell culture artifacts, we used blood DNA. We interviewed each proband separately in person using standardized criteria for diagnosing schizophrenia (DSM-IV) (Online Methods). In order to focus on non-familial cases of schizophrenia, none of the selected probands had a first- or second- degree family history of psychotic disorders (including bipolar disorder with psychotic features), and neither the probands nor their parents had a history of substance dependence. To exclude potentially causative structural genetic variations, we examined genomic DNA from every selected trio using a copy number variant–targeted array (Cytochip 2.0, Affymetrix Inc.). We found no previously reported *de novo* or other potentially causative copy number variants.

We performed exome capture on each individual using SureSelect Human All Exome Kit v.1 (Agilent Technologies Inc.) and performed sequencing using a Genome Analyzer IIx (Illumina Inc.) (**Supplementary Table 1**). We mapped all sequence reads to the reference genome using two different mapping algorithms: the Burrows-Wheeler Aligner Tool (BWA)[10] and the CLC Genomics Workbench (CLC Bio). An average of ~56 million reads were available for each individual, of which ~50% aligned to the targeted regions (sequencing statistics are presented in **Supplementary Table 1**). The exome capture was relatively effective, with an average of ~72% of targeted regions covered with a read depth >20× (**Supplementary Fig. 1** and **Supplementary Table 2**). For each of the 14 schizophrenia probands sequenced, we examined any variation from the reference genome in his or her unaffected parents and in a pool of 26 control individuals (comprising the parents of the other trios) and placed every variant found to be unique to the proband on a putative DNM list (**Supplementary Tables 3**,**4**). We kept unique variants meeting the quality thresholds for subsequent validation. We considered all mutations within coding sequences and identified 73 putative variants (**Supplementary Table 4**). For each of the putative 73 variants, we performed direct PCR amplification and Sanger sequencing in the probands and their respective parents. As a result, we validated a total of 15 variations as genuine DNMs. We found the remaining 58 variants to be either false positives or to be transmitted by one of the two parents. None of those DNMs was reported in dbSNP build 131 or in the latest release of the 1000 Genomes Project. We found no insertion or deletion mutations during this analysis.

Through this exome capture high-throughput sequencing approach, we identified 15 DNMs in 8 of the 14 schizophrenia probands (**Table 1**);

[1]Centre of Excellence in Neuromics of Université de Montréal, Centre Hospitalier de l'Université de Montréal Research Center, Montréal, Québec, Canada. [2]Centre of Excellence in Neuromics of Université de Montréal, Centre Hospitalier Universitaire Sainte-Justine Research Center, Montréal, Québec, Canada. [3]Genome Research Centre, The Li Ka Shing Faculty of Medicine, University of Hong Kong, Hong Kong, China. [4]INSERM, Université Paris Descartes, Laboratoire de Physiopathologie des Maladies Psychiatriques, Centre de Psychiatrie et Neurosciences (UMR 894), Paris, France. [5]Université Rennes 1, CH Guillaume Régnier Service Hospitalo-Universitaire de Psychiatrie, Rennes, France. [6]Université de Poitiers, CHU, Centre Hospitalier Henri Laborit, Unité de recherche clinique intersectorielle en psychiatrie, INSERM CIC-P 0802, Poitiers, France. [7]Douglas Mental Health University Institute, Department of Psychiatry, McGill University, Montréal, Québec, Canada. [8]Department of Pathology and Cellular Biology, Université de Montréal, Montréal, Québec, Canada. [9]Université Paris Descartes, Faculté de Médecine Paris Descartes, Service Hospitalo Universitaire, Centre Hospitalier Sainte-Anne, Paris, France. [10]Faculty of Medicine, Department of Medicine, Université de Montréal, Montréal, Québec, Canada. [11]These authors contributed equally to this work. Correspondence should be addressed to G.A.R. (guy.rouleau@umontreal.ca).

Received 25 February; accepted 15 June; published online 10 July 2011; doi:10.1038/ng.886

Table 1 Summary of *de novo* mutations found using deep sequencing and validated using Sanger sequencing

Individual	Chr.	Position (hg19)	Gene	Reference allele	Mutant allele	Mutation	Nucleotide substitution	Amino acid substitution	PhyloP[a]	Grantham score[b]	PolyPhen	SIFT prediction
SCZ0101	19	36,673,714	*ZNF565*	T	C	Missense	Transition	p.His385Arg	4.28987	Conservative	Possibly damaging	Damaging
SCZ0101	21	16,338,349	*NRIP1*	T	G	Missense	Transversion	p.Lys722Thr	1.81401	Moderately conservative	Probably damaging	Damaging
SCZ0201	12	57,579,450	*LRP1*	C	A	Nonsense	Transversion	p.Tyr2200X	1.52978	–	–	–
SCZ0201	17	79,637,360	*CCDC137*	A	G	Missense	Transition	p.Tyr125Cys	3.55235	Radical	Probably damaging	Damaging
SCZ0401	3	122,146,472	*KPNA1*	C	A	Nonsense	Transversion	p.Glu448X	5.78062	–	–	–
SCZ0401	19	52,826,001	*ZNF480*	C	T	Nonsense	Transition	p.Arg480X	–2.01835	–	–	–
SCZ0601	14	103,806,120	*EIF5*	A	G	Missense	Transition	p.Iso351Val	2.10831	Conservative	Benign	Tolerated
SCZ0901	3	46,717,166	*ALS2CL*	G	A	Nonsense	Transition	p.Arg733X	0.727551	–	–	–
SCZ1001	12	6,707,226	*CHD4*	G	A	Missense	Transition	p.Arg576Trp	2.53121	Moderately radical	Probably damaging	Damaging
SCZ1001	12	121,882,033	*KDM2B*	C	T	Missense	Transition	p.Gly745Ser	6.313	Moderately conservative	Probably damaging	Damaging
SCZ1001	18	6,974,966	*LAMA1*	T	C	Missense	Transition	p.Thr2187Ala	–0.76152	Moderately conservative	Benign	Tolerated
SCZ1101	11	104,825,480	*CASP4*	T	C	Missense	Transition	p.Glu39Lys	–1.1914	Moderately conservative	Possibly damaging	Tolerated
SCZ1401	1	1,153,867	*SDF4*	C	T	Missense	Transition	p.Asp295Asn	1.88812	Conservative	Benign	Damaging
SCZ1401	3	138,461,543	*PIK3CB*	T	C	Missense	Transition	p.Iso160Val	1.18761	Conservative	Benign	Tolerated
SCZ1401	12	123,804,988	*SBNO1*	G	C	Missense	Transversion	p.Asn886Lys	–0.120827	Moderately conservative	Probably damaging	Tolerated

Chr., chromosome.
[a]PhyloP score based on 46 vertebrate species. [b]We used the Grantham matrix[21] to evaluate the effect of each mutation and the classification from a previous study[22] to classify the effects.

© 2011 Nature America, Inc. All rights reserved.

we detected no DNMs in the remaining six trios. As we did not cover all exons, nor did we look at noncoding sequences or splice junctions, we cannot exclude the possibility that there might be undetected, potentially functional DNMs in these trios. The DNMs identified can be categorized as either transitions (n = 11) or transversions (n = 4). We searched the genes containing DNMs against the SZGene database[11] and found that none was previously reported to be associated with schizophrenia, which might be expected as DNMs are rare events that cannot be detected by genome-wide association studies tagging common variants. Out of the 15 validated DNMs, 4 may warrant particular interest as they are nonsense mutations that are predicted to lead to a premature stop codon, whereas the rest are missense mutations. Notably, a 4:15 ratio of nonsense to missense mutations is significantly higher than the expected ratio of 1:20, as calculated by researchers in a previous study[9] (P = 0.005467 using a binomial test, 95% confidence interval (CI) 0.077–0.55); furthermore, among all mutations reported to cause Mendelian diseases (as reported in the Human Gene Mutation Database (HGMD)), the ratio of nonsense to missense mutations is roughly 1:4, which is not significantly different from the 4:15 ratio observed in our study ($P > 0.05$).

The first nonsense mutation (c.52826001C>T), observed in individual SCZ0401, produces a stop codon (p.Arg480X) in *ZNF480* that is predicted to truncate the last 55 amino acids from ZNF480. A known polymorphism (rs113675780) that creates a stop codon at position p.Arg474X has been reported in *ZNF480*, but no population frequency data are available for that SNP. A second nonsense mutation (c.122146472C>A), also observed in individual SCZ0401, introduces a stop codon (p.Glu448X) in *KPNA1* (encoding karyopherin alpha 1) that is predicted to result in loss of the last 58 amino acids of KPNA1. No nonsense mutations have been reported in SNP databases for this gene. Notably, KPNA1 regulates and mediates V(D)J recombination through RAG1 and RAG2 (ref. 12). V(D)J recombination is an important immune system process that generates immunoglobulins and T-cell receptors[13]. Given an emerging hypothesis that autoimmunity may play a role in schizophrenia[1,2,14], further study of *KPNA1* in schizophrenia may be warranted. It is noteworthy that we found two nonsense DNMs within the same individual (SCZ0401). We speculate that only one of the two nonsense mutations might be pathogenic with respect to schizophrenia, whereas the other may or may not have a clinical effect. The third nonsense DNM (c.57579450C>A), observed in individual SCZ0201, creates a stop codon (p.Tyr2200X) in *LRP1* (encoding low-density lipoprotein receptor-related protein 1). LRP1 is a very large protein of 4,544 amino acids, and this DNM is predicted to truncate the protein to half of its normal size. Notably, LRP1 is known to be regulated by the amyloid precursor protein (APP)[15] and is believed to play a functional role in the pathogenesis of Alzheimer's disease[15,16], consistent with the notion that it may also play a role in other neuropsychiatric disorders. Three nonsense mutations in *LRP1* were previously reported in SNP databases (rs75873762, rs79339212 and rs113087094, which create stop codons at positions p.Glu177X,

Table 2 Total coding sequence with at least 20-fold coverage

Proband	Total coding sequence (bp)	Total DNMs
SCZ0101	22,031,032	2
SCZ0201	22,116,414	2
SCZ0401	16,768,992	2
SCZ0501	17,244,815	0
SCZ0601	18,757,524	1
SCZ0701	21,755,421	0
SCZ0801	22,175,382	0
SCZ0901	22,355,863	1
SCZ1001	22,111,354	3
SCZ1101	21,085,529	1
SCZ1201	21,474,141	0
SCZ1301	21,334,975	0
SCZ1401	22,870,855	3
SCZ1501	17,542,536	0
Total	289,624,833	15

Table 3 *De novo* mutation rate comparison between our schizophrenia cohort and previously reported studies of normal individuals

Study	Observed rate[a]	Current study rate	Binomial P[b]	Comments
Roach *et al.*[7]	1.10×10^{-8}	2.59×10^{-8}	0.002462	Genome sequencing of a family quartet
Lynch *et al.*[23]	1.28×10^{-8}	2.59×10^{-8}	0.0144	Survey of genes carrying disease-causing mutations
Durbin *et al.*[6] (CEU)	1.20×10^{-8}	2.59×10^{-8}	0.006336	Deep sequencing of a CEU trio
Durbin *et al.*[6] (YRI)	1.00×10^{-8}	2.59×10^{-8}	0.0009975	Deep sequencing of a YRI trio

[a]The observed *de novo* mutation rate is expressed in mutation per position in a haploid genome. [b]The 95% confidence intervals (CI) for the current comparison are 1.45×10^{-8} to 4.27×10^{-8}.

p.Glu2105X and p.Tyr2428X, respectively), but again, no population frequencies are yet available for these variants. The last nonsense DNM (c.46717166G>A) creates a stop codon (p.Arg733X) in *ALS2CL* (encoding ALS2-like protein) that is predicted to result in the loss of the last 190 amino acids of the longest isoform of ALS2CL. ALS2CL is a putative modulator of alsin (ALS2), which is involved in a juvenile form of amyotrophic lateral sclerosis[17]. No nonsense mutations have been reported for *ALS2CL* in SNP databases. Among the 11 missense DNMs identified, 7 were predicted to be deleterious when analyzed using the bioinformatic algorithms SIFT[18], PolyPhen[19], the PhyloP conservation score[20] or the Grantham matrix[21,22]. Notably, two missense variants, in *CCDC137* and *CHD4*, are predicted to be damaging by all four prediction algorithms. However, until the functional consequences of these 11 missense variants are investigated, it remains difficult to further predict which of these might actually contribute to schizophrenia pathogenesis.

The data generated by the 1000 Genomes Project[6] have established the normal DNM rate to be $\sim 1.1 \times 10^{-8}$ bp for a human haploid genome, consistent with previous estimates[7,8,23]. With this normal DNM rate in mind, we sought to test whether the number of exonic DNMs identified in the 14 schizophrenia probands was higher than expected. We therefore calculated the expected number of DNMs for our study considering both the number of probands and the amount of DNA sequenced from our probands. Using the SureSelect Human All Exome Kit v.1 probes overlapping with regions from the change for Consensus CDS (CCDS), we determined the number of sequence reads that were on target for each of the 14 schizophrenia probands (**Table 2**). We estimated the total number of coding base pairs screened to be 289.62 Mb, and with this sequence coverage, the expected number of DNMs in our study was ~6.37. However, we report here 15 DNMs, which represents a DNM rate of 2.59×10^{-8} bp rather than a rate of $\sim 1.1 \times 10^{-8}$ bp for a human haploid genome; a binomial test indicates that the number of DNMs observed in our study differs significantly (95% CI 1.4493×10^{-8} bp to 4.2710×10^{-8} bp) from what should have been observed considering different studies that have measured the *de novo* mutation rate in humans (**Table 3**). The DNM rate we observed is specific to an analysis which only takes into consideration exonic coding sequences; on a genome-wide scale, where the vast majority of the DNA is not constrained, the DNM rate would likely be similar in each individual. Our study strongly suggests that the enrichment of DNMs within the coding sequence of individuals with schizophrenia may underlie the pathogenesis of a substantial number of these individuals. It is important to stress that our observed DNM rate is conservative. Given the huge amount of sequence information, our *de novo* identification algorithms may have missed some DNMs in certain samples. Any missed DNMs would only increase the importance of the findings.

An analogous study was recently reported for mental retardation[24], which supported the hypothesis that DNMs might explain a substantial fraction of the heritability of this disorder. That study identified nine DNMs in seven mental retardation trios, which is similar to our present findings for schizophrenia (15 DNMs in eight trios). Notably, our approach was very similar to theirs, and we used the same capture kit. The contribution of DNMs to mental retardation has been recurrently hypothesized over the years, yet it is only now with the advent of powerful sequencing methods that the hypothesis could be directly tested and confirmed[25,26]. It is somewhat surprising that we find similar *de novo* mutation rates in schizophrenia and mental retardation, as this mechanism was well established for mental retardation but is only recently being explored in schizophrenia. In a previous smaller scale study screening synaptic genes by direct sequencing, our group reported that the number of functional DNMs appeared to be increased in individuals with schizophrenia[8], a result consistent with the present report, in which we used a more powerful approach with no inherent bias for the genes to be screened. In fact, when reanalyzing data from the previous synaptic gene report, we find that the observed DNM rate in schizophrenia (six functional DNMs over 91 Mb) is in agreement with the one reported here ($P > 0.05$)[8]. In conclusion, we confirmed our initial hypothesis that the deleterious DNM rate is higher in individuals with schizophrenia, and we enlarged this observation to encompass the exonic schizophrenia DNM rate. Our sequencing of trios in which the proband was affected with sporadic schizophrenia indicates that DNMs likely contribute to the development of schizophrenia, an observation which may explain part of schizophrenia's missing genetic causes. These results are also in accordance with the original twin studies used to estimate the heritability attributed to the disease. Furthermore, our results provide a list of new candidate genes to be screened in schizophrenia cohorts of various origins. Although sporadic cases are not necessarily representative of all individuals with schizophrenia, these new candidates should be further investigated as potentially implicated in this disease.

URLs. dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP/; RefSeq, http://www.ncbi.nlm.nih.gov/RefSeq/; UCSC browser, http://genome.ucsc.edu/cgi-bin/hgGateway; SzGene, http://www.szgene.org/; Human Gene Mutation Database, http://www.hgmd.cf.ac.uk/ac/index.php; Consensus Coding Sequence Project, http://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi; 1000 Genomes Project, http://www.1000genomes.org/.

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: *ZNF565* mRNA, NM_152477; *NRIP1* mRNA, NM_003489; *LRP1* mRNA, NM_002332; *CCDC137* mRNA, NM_199287; *KPNA1* mRNA, NM_002264; *ZNF480* mRNA, NM_144684; *EIF5* mRNA, NM_183004; *ALS2CL* mRNA, NM_001190707; *CHD4* mRNA, NM_001273; *KDM2B* mRNA, NM_032590; *LAMA1* mRNA, NM_005559; *CASP4* mRNA, NM_033306; *SDF4* mRNA, NM_016176; *PIK3CB* mRNA, NM_006219; *SBNO1* mRNA, NM_001167856.

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

© 2011 Nature America, Inc. All rights reserved.

ACKNOWLEDGMENTS
Subjects were selected from the Psydev study, promoted by INSERM, through the Collaborative Network for Family Study in Psychiatry ('Réseau d'étude familiale en Psychiatry', REFAPSY), supported by the Fondation Pierre Deniker. We wish to thank S. Bannour, M.-J. Dos Santos, M.A. Gorsane, N. Benjemaa, M. Chayet, S. Leroy, F. Mouaffak, K. Ossian and F. Polides for participating in recruitment of subjects and for their technical help. This work was supported by Genome Canada and Génome Québec and received co-funding from Université de Montréal for the Synapse to Disease (S2D) project as well as funding from the Canadian Foundation for Innovation, Brain & Behavior Research Foundation. Bioinformatic analysis was supported by a Canadian Institutes of Health Research (CIHR) Team grant (RMF92086). G.A.R. is grateful for the support received through his positions as Canada Research Chair in Genetics of the Nervous System and Jeanne-et-J.-Louis-Levesque Chair for the Genetics of Brain Diseases. N. Jaafari was supported by an award, bourse chercheur invité, of the Région Poitou-Charente.

AUTHOR CONTRIBUTIONS
J.G., L.X., S.L.G. and G.A.R. designed the study. M.-O.K., L.X., B.M., R.J. and N.J. recruited cases and collected clinical information. I.B. and A.H.Y.T. performed exome capture and sequencing. S.L.G., D.S., J.Y.J.B. and C.-H.L. performed alignments and variant detection. A.N., S.Z., L.J., S.L.G. and P.T. performed variant validation. S.L.G., A.D.-L., D.S., E.H., O.D., A.H.Y.T., J.Y.J.B., C.-H.L. and S.L. performed bioinformatic analyses. S.L.G., P.A.D., M.-O.K., S.L. and G.A.R. wrote the paper.

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. Purcell, S.M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
2. Stefansson, H. *et al.* Common variants conferring risk of schizophrenia. *Nature* **460**, 744–747 (2009).
3. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).
4. McCarthy, S.E. *et al.* Microduplications of 16p11.2 are associated with schizophrenia. *Nat. Genet.* **41**, 1223–1227 (2009).
5. Cardno, A.G. & Gottesman, I.I. Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics. *Am. J. Med. Genet.* **97**, 12–17 (2000).
6. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
7. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
8. 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).
9. Kryukov, G.V., Pennacchio, L.A. & Sunyaev, S.R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80**, 727–739 (2007).
10. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
11. Allen, N.C. *et al.* Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nat. Genet.* **40**, 827–834 (2008).
12. Jones, J.M. & Simkus, C. The roles of the RAG1 and RAG2 "non-core" regions in V(D)J recombination and lymphocyte development. *Arch. Immunol. Ther. Exp. (Warsz.)* **57**, 105–116 (2009).
13. Willerford, D.M., Swat, W. & Alt, F.W. Developmental regulation of V(D)J recombination and lymphocyte differentiation. *Curr. Opin. Genet. Dev.* **6**, 603–609 (1996).
14. Shi, J. *et al.* Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* **460**, 753–757 (2009).
15. Liu, Q. *et al.* Amyloid precursor protein regulates brain apolipoprotein E and cholesterol metabolism through lipoprotein receptor LRP1. *Neuron* **56**, 66–78 (2007).
16. Hollenbach, E., Ackermann, S., Hyman, B.T. & Rebeck, G.W. Confirmation of an association between a polymorphism in exon 3 of the low-density lipoprotein receptor-related protein gene and Alzheimer's disease. *Neurology* **50**, 1905–1907 (1998).
17. Hadano, S. *et al.* A gene encoding a putative GTPase regulator is mutated in familial amyotrophic lateral sclerosis 2. *Nat. Genet.* **29**, 166–173 (2001).
18. Ng, P.C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).
19. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
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. Li, W.H., Wu, C.I. & Luo, C.C. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J. Mol. Evol.* **21**, 58–71 (1984).
23. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. USA* **107**, 961–968 (2010).
24. Vissers, L.E. *et al.* A *de novo* paradigm for mental retardation. *Nat. Genet.* **42**, 1109–1112 (2010).
25. Hamdan, F.F. *et al.* *De novo* mutations in *FOXP1* in cases with intellectual disability, autism, and language impairment. *Am. J. Hum. Genet.* **87**, 671–678 (2010).
26. Hamdan, F.F. *et al.* Mutations in *SYNGAP1* in autosomal nonsyndromic mental retardation. *N. Engl. J. Med.* **360**, 599–605 (2009).

© 2011 Nature America, Inc. All rights reserved.
npg

ONLINE METHODS

Subjects. All probands were recruited through a large clinical genetics schizophrenia group (M.-O.K.). For the current study, we recruited a cohort of 14 individuals with schizophrenia and their parents (42 individuals in total). DNA was extracted from blood for all samples using the PureGene DNA kit. Paternity or maternity was confirmed using genetic microsatellite markers.

Clinical assessment procedures. All participants, after receiving a full description of the study, gave their informed written consent. All study procedures were approved by the French ethics committees and were in accordance with the Declaration of Helsinki. We investigated unrelated French adults with schizophrenia (cases) from the Departments of Psychiatry at the Sainte-Anne Hospital in Paris, the Guillaume Regnier Hospital in Rennes and the Henri Laborit Hospital in Poitiers, all of which take part in the collaborative network ReFaPsy. All investigators were individually trained in the coordinating center in order to ensure consensual procedures and ratings.

Cases fulfilled DSM-IV criteria for schizophrenia or schizoaffective disorders. Concordant diagnosis was reached by two clinicians based on the review of the direct structured interview using the Diagnosis Interview for Genetic Studies (DIGS version 3.0) conducted by trained psychiatrists and psychologists and on information from practitioners and family and prospectively documented clinical case reports. Family histories of psychiatric disorders were collected using the Family Interview for Genetic Studies (FIGS). Additional specific questionnaires were performed to collect information on obstetrical complication and early development. Exclusion criteria for all subjects included neurologic hard signs, a history of head trauma and substance dependence.

Cases included in the present study were selected from the initial cohort using the following criteria: sporadic cases (no family history of psychosis in the first- and second- degree relatives, including bipolar disorders with psychotic features) and availability of a sufficient amount of native DNA from a blood sample.

Clinical characteristics. The total selected sample comprised 14 cases (seven males and seven females of age 29 ± 11.6 years (mean ± s.d. throughout)). Schizophrenia cases were distributed between the undifferentiated subtype ($n = 7$), the disorganized subtype ($n = 6$) and schizoaffective ($n = 1$). The number of years of education was 12.9 ± 5.8 years. The ages of the father and mother when the probands were born were 29.6 ± 9.0 years and 27.5 ± 8.3 years, respectively. Mean birth weight was in the normal range of 3,383 ± 1,575 g (range 2,750–4,000 g).

The age at onset of prodomes was 12.9 ± 6.6 years, and the age at the first episode of psychosis was 19.3 ± 8.1 years. Only two cases had a lower than average educational achievement (6 and 9 years). Six cases had slight language difficulties (including stuttered speech, dyslexia and delayed reading or learning) or delayed developmental milestones (language, walking or wetting). Nine cases had minor physical anomalies as screened using an adapted version of Waldrop.

Exome capture and sequencing. Targeted enrichment was performed with Agilent SureSelect All Exome Kit v.1 optimized for Illumina sequencing with 2 μg of genomic DNA. This version of the kit is designed to cover approximately 38 Mb of genomic sequences, mainly protein coding sequences. Exon-enriched DNA libraries from 42 individuals of 14 families were sequenced individually on a single lane of the Illumina Genome Analyzer IIx platform (Illumina), which produced 76-bp end reads, in accordance with the manufacturer's instructions. Approximately 4–5 billion base calls were generated for each sample.

Read mapping. BWA[10] was used as the main aligner. The mapping was performed against the human genome (hg19) and indexed using the bwtsw algorithm included with BWA. The alignment was made using a maximum mismatch penalty of three. All other parameters from BWA were kept at the default value. The alignment was generated using a pair-end mode, and SAMTools[27] was used to store the alignment. All PCR duplicates were removed from the alignments. A second alignment was made based on the default setting of quality filtering. High-quality pair-end (76 × 2 bp) sequencing reads were obtained and mapped to the reference human genome assembly GRCh37 (NCBI Build 37.1) using CLC Genomics Workbench (CLC Bio). The nine alternate assembly loci were excluded from the reference. Only uniquely mappable reads with at least 90% similarity with the reference were considered.

Assessment of enrichment. Target specificity was defined as the percentage of reads mapped onto the target exon regions out of the number of reads that are uniquely mappable to the whole genome reference. Coverage of the mapped reads to the exon targets was tabulated by read depth at each exon nucleotide position and by the mean fold coverage within each gene.

Variant calling and annotation. VarScan[28] and CLC Genomic Workbench were used for variant detection. For practical reasons, the parameters for variant calling were chosen by comparison with NCBI dbSNP Build 131. The minimum coverage threshold was chosen such that increasing the coverage could not substantially improve the proportion of SNVs found in dbSNP v131. The DNM calls relied on the premise that the parameters for allele calls in the parents should be less stringent than that for variant calls in the probands because of the non-uniform coverage between individuals. For VarScan, the required parameters for DNM calls to meet the threshold for the proband were: (i) a minimum coverage of 20; (ii) mutation frequency (defined as the percentage of reads supporting the newly identified allele) >20% with at least three reads supporting the mutation; and (iii) a minimal PHRED score quality of 25. For the parents, the criteria were: (i) a minimum coverage of 5; (ii) mutation frequency >10% with at least two reads supporting the mutation; and (iii) a minimal PHRED score quality of 15. For the CLC Genomic Workbench software, a coverage threshold of 15× for variation calls in probands and allele calls in parents was used to detect DNMs. Annotation was made using ANNOVAR[29] against database RefSeq v46 and dbSNP v131.

Variant filtering and segregation analysis. Each variant from any proband was compared to his or her own parents and to a common pool of all other parents ($n = 26$). Each unique variant was kept. A further functional filter was applied to remove any non–protein-disruptive variant (intronic, intergenic or untranslated region).

Validation of *de novo* mutations. Primers were designed by Primer Select from DNAStar or Exon Primer to allow specific amplification of regions containing variants to validate. PCR was performed using the AmpliTaq Gold DNA Polymerase (Applied Biosystems) according to the manufacturer's instructions. To visualize DNA fragments, a small fraction of the PCR product was loaded on a 1.5% agarose gel containing ethidium bromide. PCR products were sequenced at the Genome Quebec Innovation Centre using a 3730XL DNAnalyzer (Applied Biosystems), and a Mutation Surveyor (v.3.10, SoftGenetics) was used for mutation detection analysis.

27. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
28. Koboldt, D.C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283–2285 (2009).
29. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

© 2011 Nature America, Inc. All rights reserved.