ARTICLE

Direct Measure of the De Novo Mutation Rate in Autism and Schizophrenia Cohorts

Philip Awadalla,[1,2,3,15,*] Julie Gauthier,[3,15] Rachel A. Myers,[1,7,15] Ferran Casals,[1] Fadi F. Hamdan,[2,3] Alexander R. Griffing,[7] Mélanie Côté,[3] Edouard Henrion,[3] Dan Spiegelman,[3] Julien Tarabeux,[3] Amélie Piton,[3] Yan Yang,[3] Adam Boyko,[8] Carlos Bustamante,[8] Lan Xiong,[3] Judith L. Rapoport,[9] Anjené M. Addington,[9] J. Lynn E. DeLisi,[10] Marie-Odile Krebs,[11] Ridha Joober,[12] Bruno Millet,[11] Éric Fombonne,[13] Laurent Mottron,[4] Martine Zilversmit,[1] Jon Keebler,[1,7] Hussein Daoud,[3] Claude Marineau,[3] Marie-Hélène Roy-Gagnon,[2] Marie-Pierre Dubé,[5] Adam Eyre-Walker,[14] Pierre Drapeau,[6] Eric A. Stone,[7] Ronald G. Lafrenière,[3] and Guy A. Rouleau[1,2,3,*]

The role of de novo mutations (DNMs) in common diseases remains largely unknown. Nonetheless, the rate of de novo deleterious mutations and the strength of selection against de novo mutations are critical to understanding the genetic architecture of a disease. Discovery of high-impact DNMs requires substantial high-resolution interrogation of partial or complete genomes of families via resequencing. We hypothesized that deleterious DNMs may play a role in cases of autism spectrum disorders (ASD) and schizophrenia (SCZ), two etiologically heterogeneous disorders with significantly reduced reproductive fitness. We present a direct measure of the de novo mutation rate (μ) and selective constraints from DNMs estimated from a deep resequencing data set generated from a large cohort of ASD and SCZ cases ($n = 285$) and population control individuals ($n = 285$) with available parental DNA. A survey of ~430 Mb of DNA from 401 synapse-expressed genes across all cases and 25 Mb of DNA in controls found 28 candidate DNMs, 13 of which were cell line artifacts. Our calculated direct neutral mutation rate (1.36×10^{-8}) is similar to previous indirect estimates, but we observed a significant excess of potentially deleterious DNMs in ASD and SCZ individuals. Our results emphasize the importance of DNMs as genetic mechanisms in ASD and SCZ and the limitations of using DNA from archived cell lines to identify functional variants.

Introduction

The rate at which human genomes mutate is critical to understanding every aspect of medical, statistical, and evolutionary genomics. To date, human mutation rate estimates have been indirectly inferred from a human-chimpanzee divergence approach,[1] from the analysis of mutations causing human Mendelian diseases,[2,3] or, more recently, from next-generation sequencing in one nuclear family.[4] These data suggest that there will be ~2 de novo mutations (DNMs) in the genome-wide coding regions per zygote, so that such mutations may contribute to some common diseases. Indeed, DNMs in individuals with complex disorders could explain genetic factors that are not detectable through genome-wide association studies. Deep resequencing of trios or families (e.g., patients and their parents) suffering from various diseases holds the promise of discovering DNMs that potentially could have a significant impact on disease prevalence and severity. Disease-causing mutations are more likely to involve selectively constrained positions in which mutations are likely to be less tolerated or may have a substantial impact on fitness. If DNMs contribute significantly to a disorder, then there should be more functional and potentially deleterious mutations (1) in cases versus control samples and (2) in functionally constrained sites versus nonfunctional, and thus unconstrained (neutral), sites within the same cohort.

The disruption of gene function by rare deleterious penetrant mutations could represent an important cause of neurodevelopmental disorders such as schizophrenia (SCZ, MIM 181500) and autism spectrum disorders (ASD, MIM 209850). In fact, deleterious DNMs may explain observations such as the high global incidences of ASD (~0.45%)[5] and SCZ (~0.4%)[6] despite extremely variable environmental factors and reduced reproductive fitness,[7] as well as increased risk with increasing parental age.[8,9] Indeed, recent studies report an excess of de novo copy

[1]Department of Pediatrics, Université de Montréal, Montréal, Quebec H3T 1C5, Canada; [2]Centre Hospital Université Sainte-Justine Research Centre, Université de Montréal, Montréal, Quebec H3C 1G7, Canada; [3]Centre of Excellence in Neuromics of Université de Montréal, Centre Hospitalier de l'Université de Montréal and Department of Medicine, Université de Montréal, Montréal, Quebec H2L 2W5, Canada; [4]Department of Psychiatry, Hôpital Rivière-des-Prairies, Université de Montréal, Montréal, Quebec H1E 1A4, Canada; [5]Centre de Recherche Institut de Cardiologie de Montréal, Department of Pharmacology, Université de Montréal, Montréal, Quebec H1T 1C8, Canada; [6]Groupe de Recherche sur le Système Nerveux Central, Department of Pathology and Cell Biology, Université de Montréal, Montréal, Quebec H3C 3J7, Canada; [7]Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27606, USA; [8]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA; [9]Child Psychiatry Branch, National Institute of Mental Health, Bethesda, MD 20892, USA; [10]Center for Advanced Brain Imaging, Nathan S. Kline Institute, Orangeburg, NY 10962, USA; [11]University Paris Descartes, INSERM, Laboratoire de Physiopathologie des Maladies Psychiatriques, Centre de Psychiatrie et Neurosciences, U894 , Sainte-Anne Hospital, Paris 75014, France; [12]Department of Psychiatry, McGill University and Douglas Hospital, Montréal, Quebec H3A 1A1, Canada; [13]Department of Psychiatry, McGill University and Montreal Children's Hospital, Montréal, Quebec H3Z 1P2, Canada; [14]Centre for the Study of Evolution, School of Life Sciences, University of Sussex, Brighton BN1 9QG, UK
[15]These authors contributed equally to this work
*Correspondence: philip.awadalla@umontreal.ca (P.A.), guy.rouleau@umontreal.ca (G.A.R.)
DOI 10.1016/j.ajhg.2010.07.019.

Table 1. Clinical Information for ASD and SCZ Individuals in which DNMs Were Confirmed

Sample	Final Diagnosis	Sex	IQ	Clinical Information	Comorbidity	Familial History of Psychiatric Illness	Age of Father at Birth (yrs)
S00004	Autism disorder	M	NA	ASQ[1] score = 23	None	None	30
S00015	Asperger syndrome	F	NA	No physical dimorphism. ADI-R[2] scores: social = 17, communication = 13, behavior = 7	Moderate scoliosis, hypopigmented skin patch	None	27
S00036	Autism disorder	M	NA	ADI-R scores: social = 23, communication = 14, behavior = 4	None	None	31
S00044	Autism disorder	M	NA	ADI-R scores: social = 24, communication = 10, behavior = 6	Minor anomaly: skull broad and flat on posterior aspect	None	40
S00096	Autism disorder	M	NA	ASQ score = 19	None	None	38
S00161	Schizoaffective	F	67	Childhood onset, age of onset 11 yrs. Patient with normal growth, no dysmorphic feature, speech impairment, and poor academic and social performance. ASQ score = 1	None	Father has lifetime depression and compulsive behavior	NA
S00285	Schizoaffective	M	NA	Schizoaffective disorder with age of onset of 19 yrs	Mild mental retardation	Parents are unaffected, two brothers are diagnosed with atypical chronic psychosis	NA
S00215	Schizophrenia	M	NA	No dysmorphic feature, moderate to severe emotional withdrawal	None	None	41

The following abbreviations are used: M, male; F, female; IQ, intelligence quotient; NA, not available.
[1] ASQ: Autism Screening Questionnaire (score > 15 = ASD).
[2] ADI-R: Autism Diagnostic Interview-Revised (total cutoff score for the communication and language domain is 8 for verbal subjects and 7 for nonverbal subjects. For all subjects, the cut off for the social interaction domain is 10 and the cut off for restricted and repetitive behaviors is 3).

number variants (CNVs) in ASD and SCZ compared to controls.[10–12] We hypothesized that sequencing of families with affected individuals will identify an excess of missense relative to silent de novo mutations and that these mutations are candidate causal mutations for ASD and SCZ. As part of the Synapse-to-Disease Project (S2D), we resequenced synaptic genes in ASD and SCZ cases and resequenced a subset of these genes in a group of population controls. Such a resequencing project will capture DNMs at greater resolution, with the potential to unambiguously identify missense or frameshift mutations not detectable via linkage, association, or CNV methods. To test our hypothesis, we examined variants identified by resequencing 401 genes in a cohort of 285 ASD and SCZ individuals and for a subset of 39 of these genes in 285 population control individuals. Our analyses demonstrate a neutral mutation rate similar to that already reported and an excess of de novo deleterious mutations associated with the disease cohorts.

Subjects and Methods

Diagnostic Screening and Selection of Patients

The cohort of patients used for the sequencing of candidate genes for discovery of DNMs included 142 unrelated ASD patients (122 males and 20 females) as previously described,[13] 65% of which had no family history of ASD or related neurological disorders. All patients were diagnosed by using the Diagnostic and Statistical Manual of Mental Disorders criteria (Table 1). Depending on the recruitment site, the Autism Diagnostic Interview-Revised or the Autism Diagnostic Observation Schedule was used. In addition, the Autism Screening Questionnaire was completed for all the subjects. We excluded patients with an estimated mental age of <18 months, a diagnosis of Rett syndrome, or Childhood Disintegrative Disorder, as well as patients with evidence of any other psychiatric and neurological conditions, including birth anoxia, rubella during pregnancy, fragile-X syndrome, encephalitis, phenylketonuria, tuberous sclerosis, Tourette syndrome, or West syndrome. The 143 SCZ subjects (95 males and 48 females) were collected from five different centers and included 28 cases of childhood-onset schizophrenia (COS) and 115 sporadic or familial cases (with unaffected parents) of adult-onset schizophrenia or schizoaffective disorder.[14–17] Sixty percent of the SCZ subjects had no family history of schizophrenia or other related neurological disorders. They were evaluated by experienced investigators who used the Diagnostic Interview for Genetic Studies (DIGS 3.0)[18] or Kiddie Schedule for Affective Disorders and Schizophrenia, as well as multidimensional neurological, psychological, psychiatric, and pharmacological assessments. Family history for psychiatric disorders was also collected by using the Family Interview for Genetic Studies (FIGS). All DIGS and FIGS results were reviewed by two or more psychiatrists for a final consensus diagnosis based on DSM-IIIR or DSM-IV at each center. Exclusion criteria included patients with psychotic symptoms mainly caused by alcohol, drug abuse, or other clinical diagnosis including major

cytogenetic abnormalities. We selected patients for which blood DNA was available so that DNMs could be validated and for which DNA was available from both parents to test for inheritance of the variants. The population control cohort (150 males and 135 females) consisted of unrelated individuals collected for the Quebec Newborn Twin Study (QNTS),[19] in which DNA samples were available from both parents and both twins (either monozygotic or dizygotic); however, only one sibling was chosen randomly for sequencing. All samples were collected through informed consent following approval of each of the studies by the respective institutional ethics review committees. Ethnic origins of the grandparents were self-reported by the parents of probands and population control subjects. The ASD cohort is composed of French Canadians (85), other European Caucasians (54), and non-Caucasians (3). The SCZ cohort is composed of European Caucasians (136) and Asians (7). The control population is composed of French Canadians (204), other European Caucasians (55), non-Caucasians (18), and individuals of mixed origin (8).

Selection of Candidate Genes

The list of genes screened in the S2D project was generated from a repertoire of approximately 5000 potential synaptic genes compiled from several synaptic lists from different synapse databases, including the Genes-to-Cognition (G2C) database (which includes an extensive list of postsynaptic genes)[20] and the Synapse database (SynDB, which uses Synapse Ontology algorithms to mine potential synaptic genes),[21] and from an extensive list of synaptic vesicle genes.[22] It also comprises genes identified through manual searches of PubMed that are either localized at the synapse or affect synapse-related functions (i.e., plasticity, axon or dendrite outgrowth, dendritic spine morphology, learning, and memory).

The sequencing data reported here were generated, using Sanger technology, from the screening of the coding regions and splice site junctions of 122 X-linked and 279 autosomal potentially synapse-related genes (see Table S1 available online) in 142 ASD and 143 SCZ subjects. The excess of X-linked genes is due to the S2D selection procedure, which sought to include all potentially synaptic X-linked genes because of their importance in neurodevelopmental diseases[23] and because ASD is more common in males than females.[24] The genes on the autosomes were largely ones that encode glutamate receptors (including NMDA receptors, AMPA receptors, kainate receptors, and metabotropic glutamate receptors), as well as genes that encode proteins that interact with them, in particular those complexed with the NMDAR, a majority of which were identified by large proteomic studies[25] and reported in the G2C database.[20] The autosomal gene list is composed of 203 genes from the glutamate receptor complex (23 known glutamate receptors and 180 of their synaptic interactors) and 73 genes implicated in synapse function and/or cognition, interaction with ASD genes, or because of their disruption in the context of small CNVs or balanced translocations in patients with ASD, SCZ, or mental retardation. An additional three genes were included for which mutations were reported to cause ASD, SCZ, or the related neurodevelopmental disease, nonsyndromic mental retardation, which is known to coexist with ASD. Mutations in which multiple reports have previously found associations with diseases that are not related to ASD or SCZ were eliminated. In addition, data generated from the resequencing of 39 of these 401 genes in 285 population control samples were used. The 39 genes were resequenced in controls after discovering a de novo or deleterious mutation in ASD or SCZ samples.

DNA Preparation, Sequencing, and Variant Identification

Genomic DNA was extracted from peripheral blood lymphocytes and/or lymphoblastoid cell lines with Puregene extraction kits (Gentra System). A panel of seven microsatellite markers was used to confirm parentage for all samples.[26] To overcome the issue of limited DNA material, we performed the gene screening on DNA isolated from an Epstein-Barr virus transformed lymphoblastoid cell line for most cases (n = 224). The rest were done on blood DNA (n = 61). The ASD cell line samples had been frozen or regrown a maximum of two times. For the SCZ cell lines, 59 DNA samples were acquired from Coriell and are available through a request to J.L.E.D. All unique variants (heterozygous in a single individual) detected during the screen were tested in the parents' DNA. All potential de novo variants (not seen in the parents) were reconfirmed by reamplifying the fragment and resequencing the proband blood DNA and both his/her parents with reverse and forward primers in order to eliminate PCR or sequencing artifact. All de novo variants identified originally from cell line DNA were retested in the subject DNA extracted from blood to rule out variations that could have occurred during production or growth of the lymphoblastoid cell line. An identity DNA test was performed to eliminate any cell lines and/or blood inconsistencies caused by sample identification errors or nonpaternity. Primers were designed with the Exon Primer program from the UCSC Genome Browser. PCR products were sequenced on one strand with Sanger technology, done at the Genome Quebec Innovation Centre on a 3730XL DNA Analyzer System. PolyPhred (v. 5.04), PolyScan (v. 3.0), and Mutation Surveyor (v. 3.10, SoftGenetics) were used for variant detection.

Estimation of Base Pairs Screened

To estimate the number of base pairs (bp) screened, we determined the amount of coding and noncoding (intronic, UTR) sequence screened for each gene based on our PCR amplicon designs (Table S2). Briefly, genomic intervals were calculated based on forward and reverse PCR primer sequences for each amplicon. Because the DNA sequence overlapping each PCR primer is not surveyed, these genomic intervals did not include those corresponding to each PCR primer. Furthermore, because the first ~30 bp of sequence from each sequence read are of low quality, we trimmed 30 bp from each genomic interval (usually at the forward primer end). Overlapping amplicons were merged to form a single genomic interval. Then each genomic interval was annotated as to coding and noncoding sequence for each gene with tools available on the refGene table from the UCSC Genome Browser.

We defined functional and nonfunctional sites sequenced as those in which a nucleotide change would or would not lead to an altered protein sequence, respectively. We estimated that 71.2% of coding sites were functional, whereas nonfunctional sites were estimated at 28.8% of the coding and 100% of the intronic sites.[27] The number of CpG sites was estimated as 2.8% for functional sites and 1% for nonfunctional sites.[28] The corrections for increased mutation rates in CpG regions was 10× the number of CpG sites, yielding an effective bases sequenced of 10 × CpG sites + non-CpG sites for both functional and nonfunctional sites.

Evaluation of False-Negative Mutation Calls

A subset of 71 of our ASD samples was also genotyped by Affymetrix 500K arrays. We tested our ability to detect heterozygous variants by comparing our heterozygous calls from the resequencing

Table 2. De Novo Mutations Discovered by Resequencing 458,877,850 Nucleotides of DNA in ASD, SCZ, and QTNS Control Individuals

Gene	Sample and Ref.	Diagnosis	Mutation Type	Mutation Location	Chr.	Position	Nucleotide Change and Genomic Context	Amino Acid and/or Structural Change	MAPP p Value
SHANK3	S00004	Autism disorder	INDEL	CODING	22	49500342	CGAGATTAGC(G/-)TAAGGGCCAC	Splice site delG	–
IL1RAPL1	S00015	Asperger syndrome	INDEL	CODING	X	29869731	CTTGGTGCTA(TACTCTT/-)GCTGCTTGTA	I367SfsX6	–
GSN	S00099	Asperger syndrome	INTRONIC	INTRONIC	9	123104277	GTGAGGCTGG(C/G)CCTGCCCAGC	Within intron	–
KLC2	S00036	Autism disorder	MISSENSE	CODING	11	65788196	TACTATCGGC(G/C)GGCACTGGAG	R349P	0.001
KIF5C	S00044	Autism disorder	MISSENSE	CODING	2	149575030	GGACCGTAAG(C/T)GCTACCAGCA	R802C, R872C	0.001
FLJ16237	S00096	Autism disorder	MISSENSE	CODING	7	15393678	CCATCACTTA(T/C)TTTCCATATG	F279L	0.472
NRXN1	S02959	Schizophrenia	INDEL	CODING	2	50002821	CAGCACACGG(-/ACGG)GTATGGTCGT	G1402DfsX29	–
MAP2K1	S00237	Schizophrenia	INTRONIC	INTRONIC	15	64561310	CTTCTTGTAC(G/T)GTCAGGGAGA	Within intron	–
SHANK3	S00161	Childhood-onset Schizophrenia	MISSENSE	CODING	22	49484091	GCATGACACA(C/T)GGCCTGGTGA	R536W	0.051
GRIN2B	S05650	Paranoid Schizophrenia	MISSENSE	CODING	12	13611351	CTTCTACATG(T/G)TGGGGGCGGC	L825V	<0.001
SHANK3	S00285	Schizoaffective, mild mental retardation	NONSENSE	CODING	22	49506476	TGCCCGAGAG(C/T)GAGCTCTGGC	R1117X	–
KIF17	S00215	Schizophrenia	NONSENSE	CODING	1	20886681	GGAGCAGATA(C/A)TTCCTGGATG	Y575X	–
BSN	S00237	Schizophrenia	SILENT	CODING	3	49666988	GCACTGCAGT(G/C)GTAGACCTCC	V1665V	–
ATP2B4	S00182	Disorganized Schizophrenia	SILENT	CODING	1	201935404	TCATCCGAAA(C/T)GGTCAACTCA	N195N	–
SHANK3	S04261	QNTS, unknown	MISSENSE	CODING	22	49507364	GCCACCAGTG(C/T)CTCCCAAGCC	P1429S	0.107

screen of these 71 samples with calls made for overlapping genotyped SNPs on the array (Table S3). Of the 1649 heterozygous calls made in 90 autosomal SNPs on the Affymetrix 500K chip overlapping our screened amplicons, we failed to detect 41 of those heterozygous calls in our resequencing survey, suggesting that our false-negative rate is ~2%.

Prediction of Missense Severity

The potential consequence of each missense variant was evaluated with the MAPP,[29] PolyPhen,[30] SIFT,[31] and PANTHER[32] programs. Orthologous protein sequence alignments were obtained with tools available on the Galaxy Browser website for the generation of MAPP scores.

Statistical Analysis

Excess of functional relative to nonfunctional DNMs in each category (initial, CpG, non-CpG, and effective bases sequenced) was measured via (1) binomial test ($P[X \geq$ number of functional DNMs $| q$, functional bases sequenced], where q = neutral mutation rate) and (2) Fisher's exact test (FET). Functional DNMs are defined as missense and nonsense DNMs, whereas nonfunctional DNMs include silent and intronic DNMs.

Results

Identification of De Novo Mutations

By resequencing the coding and splice junction regions of 401 genes in 142 ASD and 143 SCZ samples (and 19 of these genes in 285 Quebec Newborn Twin Study [QNTS] samples), we identified 6184 DNA variants. Of these, 2437 were unique (i.e., heterozygous in 1 of 285 unrelated individuals tested). Each of these 2437 unique variants was resequenced in the proband and both parental samples to determine inheritance mode (transmitted versus de novo). A total of 15 unique variants was confirmed in the proband blood DNA sample but not detected in either parents' blood DNA sample (Table 2). These 15 validated DNMs are either germline-derived mutations or arose as somatic mutations in blood tissue. A further 13 variants were not detected in parents' DNA, nor were they detected in the proband blood DNA sample, and they were assumed to be generated during lymphoblastoid cell line development (Table 3).

Of the 15 confirmed DNMs, 14 were detected in the ASD and SCZ cohorts (2 nonsense, 5 missense, 3 frameshifting

Table 3. Cell Line Mutations Not Observed in the Blood Sample of Patients

Gene	Sample	Status	Mutation Type	Mutation Location	Chr.	Position	Nucleotide Change and Genomic Context	Amino Acid Change	Cell Line Origin
WWC1	S00056	AUT	SILENT	INTRONIC	5	167788404	CAGAAGGAAC(G/A)GTCTGTGTGG	–	UMontreal
PLCB1	S00068	AUT	MISSENSE	CODING	20	8585862	GATTTCACTC(C/T)AGAAGTGTAC	P209L	UMontreal
PLXNB3	S00009	AUT	MISSENSE	CODING	X	152694010	GGTGACCTGG(C/T)GGCCCATTAC	A1431V	UMontreal
DRP2	S00093	AUT	MISSENSE	CODING	X	100383370	AAGCAGGCGA(C/T)GGTGGCCAGT	T203M	UMontreal
PSMD10	S00016	AUT	SILENT	CODING	X	107217953	TTGCGGCTTC(T/A)GCTGGCCGGG	S82S	UMontreal
MCF2	S00204	SCZ	SILENT	INTRONIC	X	138512219	TACAGTAATT(A/C)TTCAAGTATT	–	Krebs
SLC7A3	S00218	SCZ	INDEL	INTRONIC	X	70064365	CAGGTCAGTAT(-/A)CAAATGTTTG	InsA 3′ of exon	UMontreal
CAMK2A	S00191	SCZ	MISSENSE	CODING	5	149598465	CGAGGATGAA(G/A)ACACCAAAGG	D342N, D353N	UMontreal
GRPR	S00264	SCZ	MISSENSE	CODING	X	16080316	TCCCGGAAGC(G/T)ACTTGCCAAG	R261L	DeLisi/Coriell
ADAM22	S00193	SCZ	MISSENSE	CODING	7	87660477	AAAGTGAACC(G/A)ACAAAGTGCC	R860Q, R889Q, R896Q	UMontreal
ARHGAP6	S00261	SCZ	MISSENSE	CODING	X	11592643	GAGAGTCTCG(G/A)CCCTCGCTTG	G76D	UMontreal
ADD2	S00067	SCZ	NONSENSE	CODING	2	70744118	AAAGAAATTC(C/T)GAACCCCCTC	R404X, R710X	UMontreal
RPS6KA6	S00274	SCZ	SILENT	CODING	X	83206731	ATCAGCGGTA(T/C)ACTGCTGAAC	Y672Y	DeLisi/Coriell

indel, 2 silent, and 2 intronic DNMs) and 1 was a missense DNM found in the population control group (Table 2). Five of the 11 point mutations in cases were transitions, four were CpG mutations, and six were transversions. Eight of the 14 DNMs detected among ASD and SCZ samples were disruptive to translation or protein structure and/or function, including three frameshift and two nonsense DNMs, and three missense DNMs were predicted to significantly disrupt protein structure via the computational prediction method MAPP.[33] Significant MAPP scores reflect a potentially damaging amino acid change and predict deleterious consequences. Reassuringly, MAPP scores relate to allele frequencies as would be predicted by population genetics, with increasingly deleterious alleles tending toward lower frequencies (Figure S1). Three MAPP p values of DNMs in SCZ and ASD samples are significantly low (Table 2); two were found in kinesin-encoding genes (R349P in *KLC2* [MIM 611729] and R802C in *KIF5C* [MIM 604593]). One nonsense mutation (Y575X) was also found in a kinesin-encoding gene (*KIF17* [MIM 605037]). One nonsense and one splice site deletion were found within *SHANK3* (MIM 606230), and a frameshift mutation was found in each of *IL1RAPL1* (MIM 300206) and *NRXN1* (MIM 600565). We have also confirmed the damaging predicted functional impact of the five de novo missense mutations via three other prediction programs (PANTHER, SIFT, and PolyPhen) (data not shown). Only one DNM was discovered in an X-linked gene (*IL1RAPL1*).[12] More detailed description of these genes and their potential role in ASD and SCZ will be presented elsewhere (unpublished data).

Estimates of the Neutral Human Mutation Rate

To calculate human mutation rates, we estimated the total initial count of bases we resequenced in all cohorts as 458.8 Mb (Table 4; Subjects and Methods). This includes 230.6 Mb of protein-coding and 228.3 Mb of intronic sequence. In the ASD and SCZ cohorts, exonic material sequenced is 215,186,702 bases and intronic is 218,145,769 bases. The total number of replacement sites in the cases is 153,704,787 and the number of silent (synonymous and intronic) sites is 279,627,684. In the QNTS cohort, 15,422,960 bases were exonic and 10,122,419 bases were intronic. The number of replacement sites is 11,016,400 and the number of silent sites is 14,528,979.

We distinguished functional from nonfunctional sites based on the effect of a mutation on transcription or translation of the protein at a given position (see Subjects and Methods). When addressing whether there was an excess of functional DNMs[34] relative to nonfunctional or silent base pairs, we calculated the "effective bp count."[35] Because mutations are 10 times more likely to occur at CpG sites than at non-CpG sites, we calculated the total number of CpG and non-CpG sites for both functional and nonfunctional sites[28] and calculated the effective bp count (non-CpG sites + 10 × CpG sites) to account for increased mutation rates in CpG sites (Table 4).

We calculated the neutral mutation rate by examining the number of DNMs found in nonfunctional sites in our ASD, SCZ, and QNTS samples. In total, we sequenced ~294 Mb of nonfunctional DNA, in which we observed four DNMs (Table 2; in genes *GSN* [MIM 137350], *MAP2K1* [MIM 176872], *BSN* [MIM 604020], and *ATP2B4* [MIM 604020]). Because these mutations are unlikely to be pathogenic (referred to here as "neutral"), they allowed us to directly estimate the rate of neutral point mutations. We estimated this to be 1.36×10^{-8} mutations per site per generation (95% Poisson confidence interval: 0.34×10^{-8}, 2.7×10^{-8}). These estimates of neutral mutation rates are

Table 4. Base Pairs and DNMs Surveyed among ASD and SCZ Trios with No Family History

Site and Mutation Class		Initial Count[a]	CpG[b]	Non-CpG	Effective Count[c]
Functional	Nonsynonymous bases	96,065,492	2,689,834	93,375,658	120,273,996
	Nonsynonymous DNMs	6	3	3	6
Neutral	Synonymous bases	61,481,915	1,721,494	59,760,421	76,975,357
	Intronic bases	218,145,769	2,181,458	215,964,311	237,778,888
	Silent (synonymous and intronic) DNMs	4	2	2	4
p value	One-tail binomial test[d]	0.003	0.161	0.031	0.008
	Fisher's exact test	0.022	0.4041	0.1067	0.032

[a] All statistics are calculated with base pairs and DNMs calculated for only trios with unaffected families.
[b] Estimated 2.8% of initial coding sites and 1% of initial intronic sites are CpG sites.
[c] To account for increased mutation rates: (non-CpG sites) + (10 × CpG sites).
[d] P(× ≥ number of functional DNMs | nonfunctional rate).

similar to, and not significantly different from, the estimate of 2.5×10^{-8} derived from phylogenetic analyses[1,2] and the intergeneration estimate of 1.1×10^{-8} derived from next-generation sequencing data.[4]

Excess of Functional DNMs in ASD and SCZ Cohorts

If DNMs cause sporadic cases of ASD and SCZ, then DNMs will be more common in functional than in nonfunctional sites in our disease cohorts. Based on an observation of four DNMs in 294 Mb of neutral (silent and intronic) DNA, we expect 1.3 DNMs in the 96 Mb of nonsynonymous DNA sites to be resequenced in the cases with no family history of disease (65% of ASD cases and 60% of SCZ cases). However, among trios without family histories (Table 1), we observed six nonsynonymous DNMs surveyed in individuals, representing a significant enrichment of nonsynonymous DNMs ($p = 0.003$ in one-tail binomial test; $p = 0.022$ FET; Table 4). This excess remains significant even when we take into account that CpG dinucleotides mutate faster than other sites and are more common in exons than introns ($p = 0.008$ in one-tail binomial test; $p = 0.032$ FET; see Table 4 and Subjects and Methods).

If DNMs cause disease, we also expect point mutations with larger effects to be more frequent than expected in the disease group. Among our five nonsynonymous DNMs in trios with no family history, two are nonsense mutations (ratio 1:2.5), similar to previous estimates[3] for DNMs causing Mendelian diseases (1:3.9) that are cataloged in the Human Gene Mutation Database (HGMD). Also, the ratio of synonymous to missense DNMs in the ASD and SCZ cohort is similar to that observed for HGMD.[3] Under a neutral model,[3] we would expect a ratio of 1 nonsense to 19.7 missense DNMs[3] when only point mutations are considered. In HGMD, the number of missense to nonsense DNMs was significantly higher than the neutral expectation. Using a binomial test, our observed number of missense to nonsense DNMs was also significantly higher than the neutral expectation ($p = 0.04$), suggesting that some of the mutations are predisposed to be pathogenic. All of these observations suggest an excess of potentially disease-predisposing DNMs in the SCZ and ASD cohort. Taken together, these lines of evidence suggest that mutations with functional effects are overrepresented within the synapse genes sequenced in individuals showing sporadic ASD and SCZ.

Comparing DNMs and Segregating Variant Ratios

Functional and nonfunctional segregating variants provide an expectation of the proportion of functional and nonfunctional DNMs. We compared the ratio of functional and nonfunctional DNMs to the ratios of the same classes of segregating variants in the ASD and SCZ cohorts (Table 5). Similar observations were found for the QNTS cohort. The comparison was significant when the functional and nonfunctional DNMs were compared to all segregating sites (FET, $p < 0.001$) and to unique SNP classes in the ASD and SCZ cohort (FET, $p = 0.003$). Given that the ratio of functional to nonfunctional was two times higher for DNMs relative to segregating sites, this suggests an excess of deleterious DNMs. Furthermore, given that rare SNP classes are likely enriched for slightly deleterious missense mutations,[3] this significant comparison can be considered conservative. Under the expectation that most highly deleterious mutations will be selectively removed in one generation, the unique comparisons above

Table 5. Comparisons of Constraint for Genes Expressed at the Synapse in the ASD and SCZ Cohort

	Functional	Nonfunctional	Functional: Nonfunctional	p Value
Point and indel DNMs	10	4	2.5	–
Unique SNPs	785	1652	0.48	0.003[a]
All SNPs	1306	4878	0.27	<0.001[a]

Shown are the counts of point and indel mutations or segregating SNPs for the different categories of variation.
[a] p value is the result of Fisher exact test comparisons for DNMs versus the two allele frequency classes of SNPs (unique or all).

provide insight into the proportion of deleterious mutations in humans that are selectively removed relative to segregating variation. Potentially disease-causing DNMs were more frequent than nonfunctional DNMs in our cohorts relative to expectations inferred from segregating mutations.

Discussion

In the present study, we have attempted to directly estimate the mutation rate with a large set of resequencing data generated from a common disease-based project. In addition, we have tried to validate that DNMs are a possible genetic factor of ASD and SCZ. There are three main conclusions that can be drawn from our study. First, the source of biological material (blood DNA versus cell line DNA) is crucial while doing experimental analyses with resequencing data seeking DNMs. All DNMs analyzed here were confirmed by resequencing, via standard Sanger technology, from DNA samples extracted from blood in the proband and in the parents. In so doing, we discovered that ~50% of our identified DNMs are the result of mutations that most likely occurred during the transformation and propagation of lymphoblastoid cell lines, thus creating false-positive DNMs. This observation was also recently stressed in CNV analyses by The Wellcome Trust Case Control Consortium.[36] This biological artifact, if unnoticed, would have significantly biased our results and would have contributed to a doubling of mutation rates for all classes of sites. Interestingly, 8 of 13 cell line mutations were X-linked, suggesting that this chromosome is particularly susceptible to the generation or accumulation of deleterious mutations after transformation of lymphoblasts with Epstein-Barr virus. These mutations, which are hemizygous in males, may also be positively selected because they contribute to higher fitness in cells carrying these mutations. An awareness of the high rate of mutation observed in cell lines, some of which are archived at the Coriell Institute, is critical to any large-scale whole-genome sequencing project, and potentially to those taking advantage of next-generation sequencing technologies, to capture rare and/or pathogenic mutations. Not only will the inherent error rate of the technologies be critical, but so, too, will the choice of samples and the way those samples are being maintained or cultured. Second, by using a direct calculation and classical Sanger sequencing, we validated the reported estimates of the neutral mutation rate in humans. Third, our study confirms the critical role that large-sample, high-resolution nucleotide surveys play in detecting potentially disease-causing DNMs.

We acknowledge that there are some weaknesses in our present study. The amount of resequencing in the controls is substantially lower than the resequencing in the cases. In fact, direct comparison in terms of sequences covered between cases and controls would be the optimal way to directly estimate the rate of mutation and detect significant differences in mutation rate between the cases and controls. Nevertheless, our analyses show that the rate of potentially deleterious DNMs is significantly higher in functional compared to nonfunctional sites within the disease cohorts, suggesting a role of functional DNMs in the etiology of ASD and SCZ. Given that our estimate of the neutral human mutation rate is consistent with a recent genome-wide estimate[4] and the accumulation of more direct observations of mutation, the confidence intervals of mutation rate estimates will begin to narrow. The rate of functional mutations in this survey is almost five times that of neutral or genome-wide rates, supporting our conclusions that we have likely detected mutations that are causal with respect to ASD and SCZ. By resequencing the genes in which DNMs were discovered among QNTS (random) participants, we were able to validate that these genes, among a substantial number of random individuals, do not carry functional DNMs, with the exception of one locus (*SHANK3*, Table 2). At *SHANK3*, a nonsynonymous mutation of low predicted functional impact was discovered in the QNTS cohort. *SHANK3* has been previously implicated in ASD[37] and may be a rapidly evolving gene in humans, with substantial neurological phenotypic impact.

By demonstrating that functional DNMs are at higher relative frequencies than segregating polymorphisms (Table 5), we show that DNMs may have a substantial role in ASD and SCZ etiology. The power of the genomics approaches employed here is that DNMs are not subject to the same demographic processes that shape segregating site variation. As a result, it is not necessary to test or correct for population structure, nor are our analyses subject to population stratification or admixture issues associated with GWAS analyses. By using a simple genomics approach that compares different classes of sites, we have sufficient power to map candidate mutations that are more likely to contribute to these diseases.

From sequencing only 8% of genes expressed in the synapse, functional DNMs were found in 5% of individuals with no family history, exhibiting a wide range of clinical phenotypes (see Subjects and Methods and Table 1). Although we biased our sampling strategy toward likely candidate genes, our predictions appear to have been poor regarding the X chromosome. It is therefore possible that by sequencing all 5000 synapse-related genes, we may uncover many of the mutations responsible for sporadic cases of ASD and SCZ. Furthermore, because nonfunctional DNMs are predicted to be relatively rare in ASD and SCZ genes (1.36×10^{-8} nonsynonymous DNMs per site), it may be easy to determine the likely causative mutations.

Supplemental Data

Supplemental Data include one figure and four tables and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

We would like to thank all the families and individuals who participated in this study. We are thankful for the efforts of the members of the Genome Québec Innovation Centre Sequencing and Bioinformatic groups. This work was supported by Genome Canada and Génome Québec and received cofunding from Université de Montréal for the Synapse-to-Disease (S2D) Project, funding from the Canadian Foundation for Innovation to both G.A.R and P.A., and cofunding from the Ministré de Exploration, Innovation et Economique of Québec. G.A.R. holds the Canada Research Chair in Genetics of the Nervous System; P.A. holds career awards from the Fonds de Recherche Santé Québec and Génome Québec.

Received: January 22, 2010
Revised: July 22, 2010
Accepted: July 27, 2010
Published online: August 26, 2010

Web Resources

The URLs for data presented herein are as follows:

Coriell Institute for Medical Research, http://www.coriell.org/
Galaxy Browser, http://main.g2.bx.psu.edu/
Human Gene Mutation Database, http://www.hgmd.cf.ac.uk/ac/index.php
McGill University and Genome Québec Innovation Centre, http://www.genomequebecplatforms.com/mcgill/
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/
PANTHER Classification System, http://www.pantherdb.org/tools/csnpScoreForm.jsp
PolyPhen, http://genetics.bwh.harvard.edu/pph/
SIFT, http://sift.jcvi.org/
Synapse-to-Disease Project, http://www.synapse2disease.ca/
UCSC Genome Browser, http://genome.ucsc.edu/

References

1. Nachman, M.W., and Crowell, S.L. (2000). Estimate of the mutation rate per nucleotide in humans. Genetics *156*, 297–304.
2. Kondrashov, A.S. (2003). Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum. Mutat. *21*, 12–27.
3. Kryukov, G.V., Pennacchio, L.A., and Sunyaev, S.R. (2007). Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. Am. J. Hum. Genet. *80*, 727–739.
4. Roach, J.C., Glusman, G., Smit, A.F., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M., et al. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science *328*, 636–639.
5. Rutter, M. (2005). Incidence of autism spectrum disorders: Changes over time and their meaning. Acta Paediatr. *94*, 2–15.
6. Saha, S., Chant, D., Welham, J., and McGrath, J. (2005). A systematic review of the prevalence of schizophrenia. PLoS Med. *2*, e141.
7. Bassett, A.S., Bury, A., Hodgkinson, K.A., and Honer, W.G. (1996). Reproductive fitness in familial schizophrenia. Schizophr. Res. *21*, 151–160.
8. Croen, L.A., Najjar, D.V., Fireman, B., and Grether, J.K. (2007). Maternal and paternal age and risk of autism spectrum disorders. Arch. Pediatr. Adolesc. Med. *161*, 334–340.
9. Malaspina, D., Brown, A., Goetz, D., Alia-Klein, N., Harkavy-Friedman, J., Harlap, S., and Fennig, S. (2002). Schizophrenia risk and paternal age: A potential role for de novo mutations in schizophrenia vulnerability genes. CNS Spectr. *7*, 26–29.
10. Xu, B., Roos, J.L., Levy, S., van Rensburg, E.J., Gogos, J.A., and Karayiorgou, M. (2008). Strong association of de novo copy number mutations with sporadic schizophrenia. Nat. Genet. *40*, 880–885.
11. Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O.P., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J.E., et al; GROUP. (2008). Large recurrent microdeletions associated with schizophrenia. Nature *455*, 232–236.
12. Piton, A., Michaud, J.L., Peng, H., Aradhya, S., Gauthier, J., Mottron, L., Champagne, N., Lafrenière, R.G., Hamdan, F.F., Joober, R., et al; S2D Team. (2008). Mutations in the calcium-related gene IL1RAPL1 are associated with autism. Hum. Mol. Genet. *17*, 3965–3974.
13. Gauthier, J., Bonnel, A., St-Onge, J., Karemera, L., Laurent, S., Mottron, L., Fombonne, E., Joober, R., and Rouleau, G.A. (2005). NLGN3/NLGN4 gene mutations are not responsible for autism in the Quebec population. Am. J. Med. Genet. B. Neuropsychiatr. Genet. *132B*, 74–75.
14. DeLisi, L.E., Shaw, S.H., Crow, T.J., Shields, G., Smith, A.B., Larach, V.W., Wellman, N., Loftus, J., Nanthakumar, B., Razi, K., et al. (2002). A genome-wide scan for linkage to chromosomal regions in 382 sibling pairs with schizophrenia or schizoaffective disorder. Am. J. Psychiatry *159*, 803–812.
15. Gochman, P.A., Greenstein, D., Sporn, A., Gogtay, N., Nicolson, R., Keller, A., Lenane, M., Brookner, F., and Rapoport, J.L. (2004). Childhood onset schizophrenia: Familial neurocognitive measures. Schizophr. Res. *71*, 43–47.
16. Joober, R., Rouleau, G.A., Lal, S., Dixon, M., O'Driscoll, G., Palmour, R., Annable, L., Bloom, D., Lalonde, P., Labelle, A., and Benkelfat, C. (2002). Neuropsychological impairments in neuroleptic-responder vs. -nonresponder schizophrenic patients and healthy volunteers. Schizophr. Res. *53*, 229–238.
17. Mechri, A., Bourdel, M.C., Slama, H., Gourion, D., Gaha, L., and Krebs, M.O. (2009). Neurological soft signs in patients with schizophrenia and their unaffected siblings: Frequency and correlates in two ethnic and socioeconomic distinct populations. Eur. Arch. Psychiatry Clin. Neurosci. *259*, 218–226.
18. Gourion, D., Goldberger, C., Bourdel, M.C., Bayle, F.J., Millet, B., Olie, J.P., and Krebs, M.O. (2003). Neurological soft-signs and minor physical anomalies in schizophrenia: Differential transmission within families. Schizophr. Res. *63*, 181–187.
19. Lemelin, J.P., Boivin, M., Forget-Dubois, N., Dionne, G., Séguin, J.R., Brendgen, M., Vitaro, F., Tremblay, R.E., and Pérusse, D. (2007). The genetic-environmental etiology of cognitive school readiness and later academic achievement in early childhood. Child Dev. *78*, 1855–1869.
20. Croning, M.D., Marshall, M.C., McLaren, P., Armstrong, J.D., and Grant, S.G. (2009). G2Cdb: The Genes to Cognition database. Nucleic Acids Res. *37* (Database issue), D846–D851.
21. Zhang, W., Zhang, Y., Zheng, H., Zhang, C., Xiong, W., Olyarchuk, J.G., Walker, M., Xu, W., Zhao, M., Zhao, S., et al. (2007). SynDB: A Synapse protein DataBase based on synapse ontology. Nucleic Acids Res. *35* (Database issue), D737–D741.

22. Trinidad, J.C., Specht, C.G., Thalhammer, A., Schoepfer, R., and Burlingame, A.L. (2006). Comprehensive identification of phosphorylation sites in postsynaptic density preparations. Mol. Cell. Proteomics *5*, 914–922.
23. Laumonnier, F., Cuthbert, P.C., and Grant, S.G. (2007). The role of neuronal complexes in human X-linked brain diseases. Am. J. Hum. Genet. *80*, 205–220.
24. Skuse, D.H. (2000). Imprinting, the X-chromosome, and the male brain: Explaining sex differences in the liability to autism. Pediatr. Res. *47*, 9–16.
25. Collins, M.O., Yu, L., Coba, M.P., Husi, H., Campuzano, I., Blackstock, W.P., Choudhary, J.S., and Grant, S.G. (2005). Proteomic analysis of in vivo phosphorylated synaptic proteins. J. Biol. Chem. *280*, 5972–5982.
26. Gauthier, J., Champagne, N., Lafrenière, R.G., Xiong, L., Spiegelman, D., Brustein, E., Lapointe, M., Peng, H., Côté, M., Noreau, A., et al; S2D Team. (2010). De novo mutations in the gene encoding the synaptic scaffolding protein SHANK3 in patients ascertained for schizophrenia. Proc. Natl. Acad. Sci. USA *107*, 7863–7868.
27. Eyre-Walker, A., and Keightley, P.D. (1999). High genomic deleterious mutation rates in hominids. Nature *397*, 344–347.
28. Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. USA *103*, 1412–1417.
29. Stone, E.A., and Sidow, A. (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res. *15*, 978–986.
30. Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous SNPs: Server and survey. Nucleic Acids Res. *30*, 3894–3900.
31. Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814.
32. Thomas, P.D., Kejariwal, A., Campbell, M.J., Mi, H., Diemer, K., Guo, N., Ladunga, I., Ulitsky-Lazareva, B., Muruganujan, A., Rabkin, S., et al. (2003). PANTHER: A browsable database of gene products organized by biological function, using curated protein family and subfamily classification. Nucleic Acids Res. *31*, 334–341.
33. Stone, E.A., and Sidow, A. (2007). Constructing a meaningful evolutionary average at the phylogenetic center of mass. BMC Bioinformatics *8*, 222.
34. Smith, N.G., and Eyre-Walker, A. (2001). Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. Mol. Biol. Evol. *18*, 982–986.
35. Eyre-Walker, A. (1998). Problems with parsimony in sequences of biased base composition. J. Mol. Evol. *47*, 686–690.
36. Craddock, N., Hurles, M.E., Cardin, N., Pearson, R.D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D.F., Giannoulatou, E., et al; Wellcome Trust Case Control Consortium. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature *464*, 713–720.
37. Moessner, R., Marshall, C.R., Sutcliffe, J.S., Skaug, J., Pinto, D., Vincent, J., Zwaigenbaum, L., Fernandez, B., Roberts, W., Szatmari, P., and Scherer, S.W. (2007). Contribution of SHANK3 mutations to autism spectrum disorder. Am. J. Hum. Genet. *81*, 1289–1297.