Supplemental Information

Loss-of-Function Variants in Schizophrenia Risk and *SETD1A* as a Candidate Susceptibility Gene

Atsushi Takata, Bin Xu, Iuliana Ionita-Laza, J. Louw Roos, Joseph A. Gogos, and Maria Karayiorgou

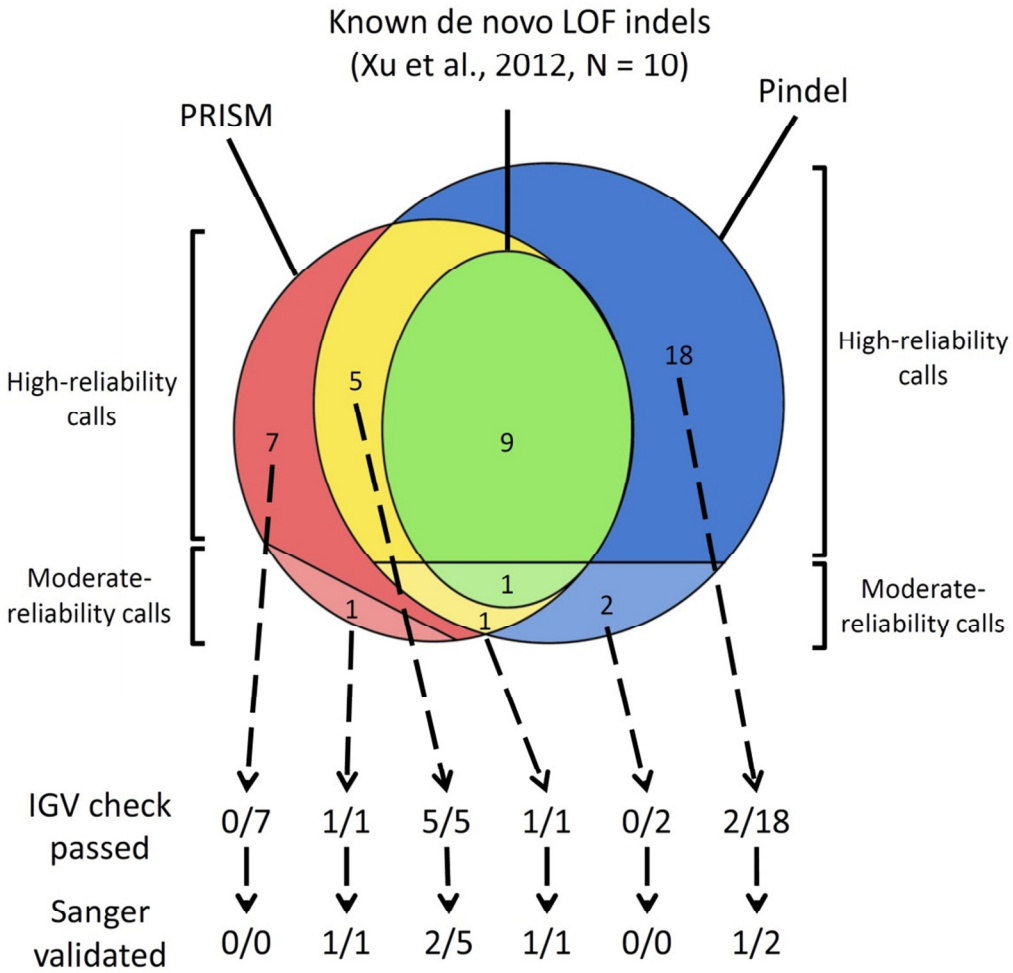


Figure S1. Venn diagram depicting the procedure for identification of *de novo* LOF indels, related to Figure 1 and Table 1

Analysis using PRISM (Jiang et al., 2012) generated a total of 24 variant calls (left circle ; 22 highreliability candidates with five or more valiant supporting reads and 1 moderate-reliability candidate with three or four supporting reads in cases, and 1 high-reliability candidate in controls). Analysis using Pindel (Ye et al., 2009) generated a total of 36 variant calls (right circle; 26 high-reliability and 4 moderate-reliability candidates in cases, and 6 high-reliability candidates in controls). 16 variant calls overlapped (yellow and light green), and all of the ten LOF indels validated in our previous study (Xu et al., 2012) were included among them. These candidates were first manually checked by Integrative Genomics Viewer (IGV) (Robinson et al., 2011), and then subjected to validation experiments by Sanger sequencing. Overall validation rates (including previously identified variants) were: PRISM: 58% (14/24); Pindel: 39% (14/36) and overlapping calls: 81% (13/16).

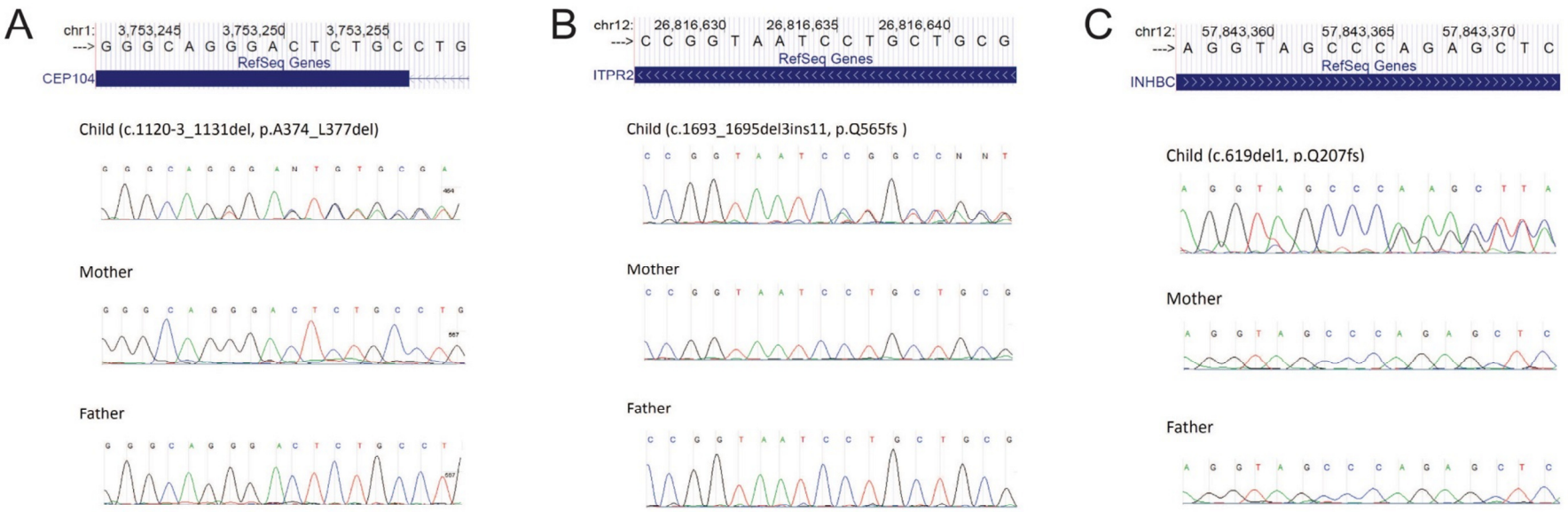
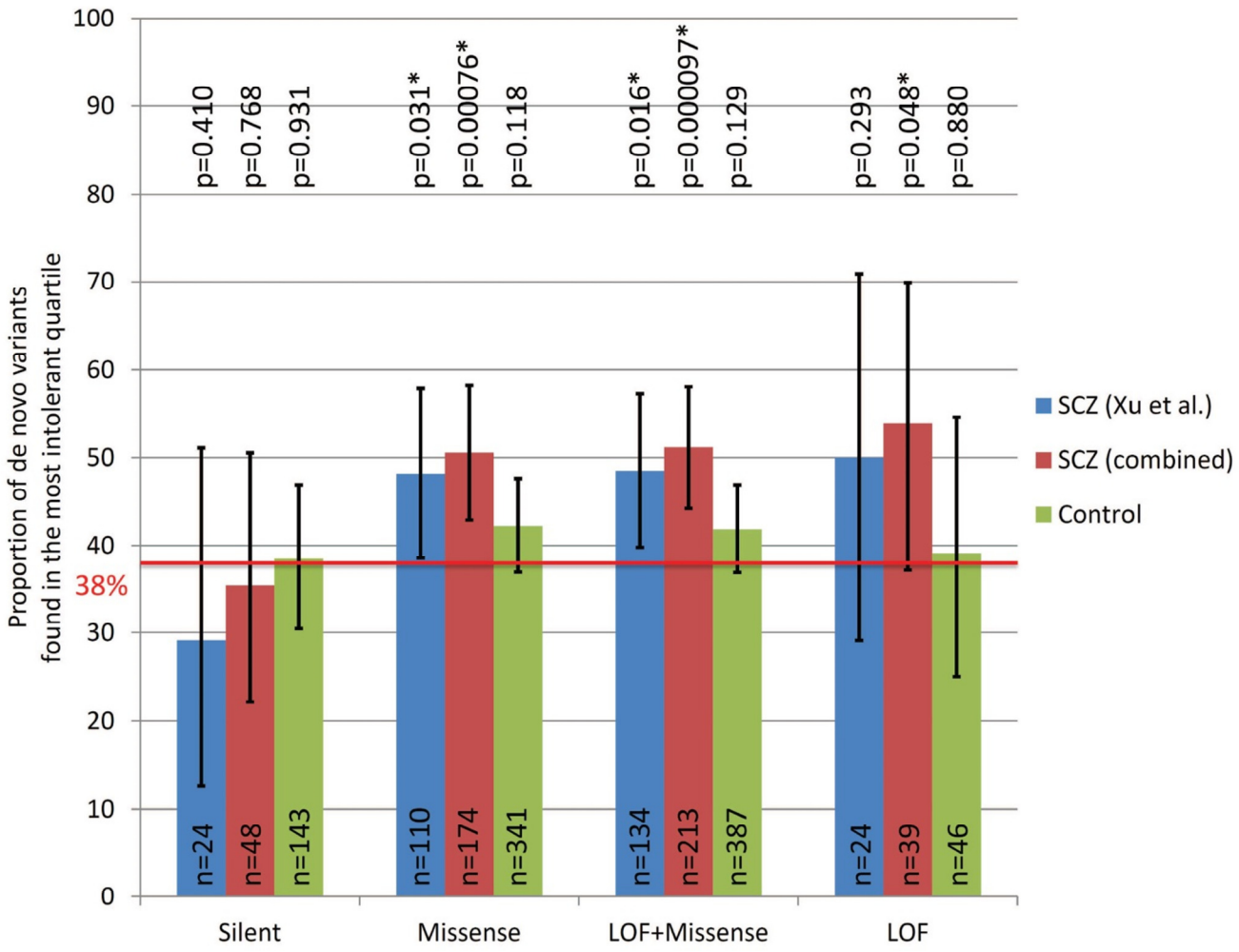


Figure S2. Sanger sequencing traces for the newly identified *de novo* LOF indels in *CEP104, ITPR2* and *INHBC*, related to Figure 1 and Table 1

Sanger sequencing traces for (A) c.1120-3_1131del, p.A374_L377del variant in *CEP104*, (B) c.1693_1695del3ins11, p.Q565fs variant in *ITPR2* and (C) c.619del1, p.Q207fs variant in *INHBC*.

A


Proportion of de novo variants found in the most intolerant quartile
100
90
80
70
60
50
40
38%
30
20
10
0
p=0.410
p=0.768
p=0.931
p=0.031*
p=0.00076*
p=0.118
p=0.016*
p=0.000097*
p=0.129
p=0.293
p=0.048*
p=0.880
n=24
n=48
n=143
n=110
n=174
n=341
n=134
n=213
n=387
n=24
n=39
n=46
Silent
Missense
LOF+Missense
LOF
SCZ (Xu et al.)
SCZ (combined)
Control

B

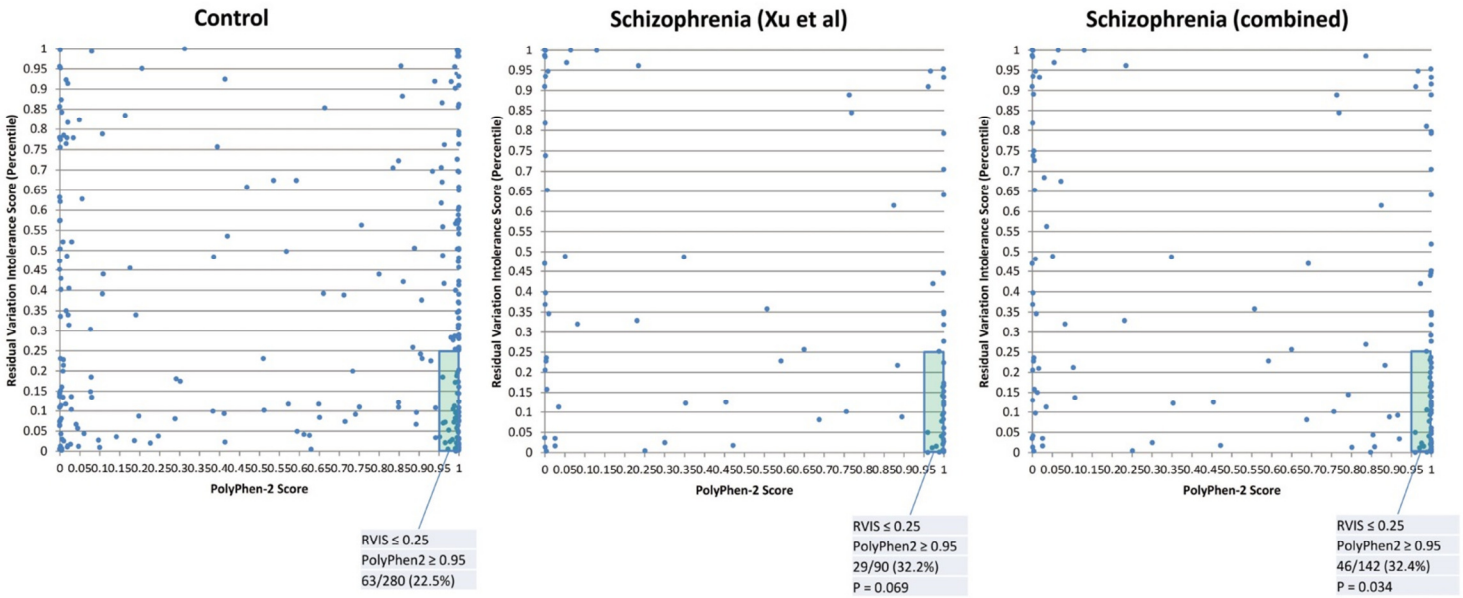

Control
Schizophrenia (Xu et al)
Schizophrenia (combined)
Residual Variation Intolerance Score (Percentile)
PolyPhen-2 Score
RVIS ≤ 0.25
PolyPhen2 ≥ 0.95
63/280 (22.5%)
RVIS ≤ 0.25
PolyPhen2 ≥ 0.95
29/90 (32.2%)
P = 0.069
RVIS ≤ 0.25
PolyPhen2 ≥ 0.95
46/142 (32.4%)
P = 0.034

Figure S3. Application of residual variation intolerance scores (RVIS) to *de novo* variants in SCZ, related to Table 1 and Figure 2

(A) The proportion of *de novo* variants found in genes in the most intolerant quartile. RVIS (Petrovski et al., 2013), which reflects gene intolerance against damaging mutations, was used to define the most intolerant quartile. P values were calculated by binomial exact tests with hypothesized probability of success = 0.38, which is the theoretical likelihood of finding *de novo* variants in the 25th percentile of most intolerant genes, considering the gene size [red line, (Petrovski et al., 2013). *$p < 0.05$. Error bars indicate 95% confidence intervals. SCZ (Xu et al.): SCZ cases from (Xu et al., 2012); SCZ (combined): SCZ cases from the following studies: (Girard et al., 2011; Gulsuner et al., 2013; Xu et al., 2012); Control: healthy subjects or non-affected siblings from the following studies: (Gulsuner et al., 2013; Iossifov et al., 2012; O'Roak et al., 2012b; Rauch et al., 2012; Sanders et al., 2012; Xu et al., 2012). *De novo* damaging (LOF and missense) variants were more frequently observed in intolerant genes in SCZ, but not in controls. (B) Two-dimensional plotting of RVIS and PolyPhen-2 scores for *de novo* missense variants. Y-axis indicates RVIS. X-axis indicates quantitative scores reflecting functional effects of missense variants predicted by PolyPhen-2 (Adzhubei et al., 2010). Left: healthy subjects or non-affected siblings from the following studies: (Gulsuner et al., 2013; Iossifov et al., 2012; O'Roak et al., 2012b; Rauch et al., 2012; Sanders et al., 2012; Xu et al., 2012), Center: SCZ cases from (Xu et al., 2012), Right: SCZ cases from the following studies: (Girard et al., 2011; Gulsuner et al., 2013; Xu et al., 2012). Missense variants predicted to be damaging (PolyPhen-2 scores ≥ 0.95) and located within the intolerant genes (RVIS < 0.25) were significantly enriched in the combined group of SCZ cases ($p = 0.034$, OR = 1.65).

A

Term	Category	Raw p value	BH-corrected p value	Fold Enrichment
chromatin regulator	SP_PIR_KEYWORDS	0.000057	0.0052	10.2
GO:0005626~insoluble fraction	GOTERM_CC_FAT	0.000218	0.0175	3.7
GO:0005624~membrane fraction	GOTERM_CC_FAT	0.000159	0.0254	3.9
calcium	SP_PIR_KEYWORDS	0.000910	0.0535	3.9
GO:0005509~calcium ion binding	GOTERM_MF_FAT	0.000420	0.0739	3.5

B

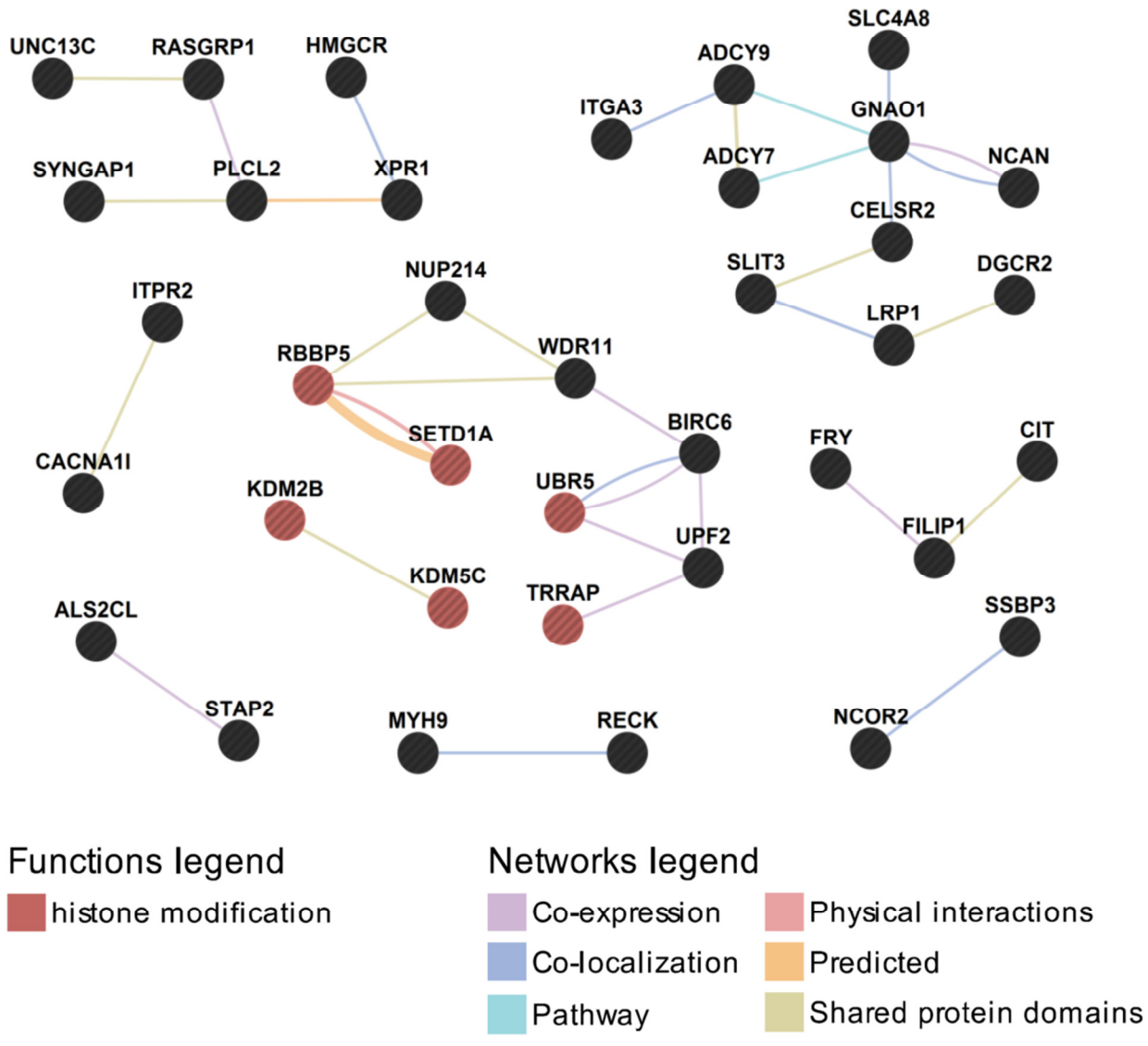


Figure S4. Damaging *de novo* variants highlight a role of chromatin regulators in SCZ, related to Figure 1 and Table 1

(A) Result of the DAVID (Huang da et al., 2009) enrichment analysis using 62 genes categorized as "intolerant" and hit by *de novo* LOF or damaging missense (PolyPhen-2 scores ≥ 0.95) variants in SCZ as an input. Terms with Benjamini Hochberg (BH)-corrected $p < 0.1$ are shown. (B) Network analysis using the same 62 genes as input. For construction of a network figure, GeneMANIA (Zuberi et al., 2013) was used. The thickness of edges represents the weight on each edge, reflecting the degree of confidence of the relationships between a gene pair. The nodes with red color indicate genes related to chromatin modification. *UBR5* was included in the chromatin modification genes as this gene was reported to play a role in the control of histones ubiquitination following DNA breakage (Gudjonsson et al., 2012).

TableS1. Full list of the de novo loss-of-function variants in 265 trios

Position (hg19)	Reference allele	Variant allele	Type	Effect	Property	Gene	Population	Diagnosis
Newly identified								
chr1:3753245-3753259	CAGGGACTCTGCTG	-	Deletion	Canonical splice site + Inframe	c.1120-3_1131del, p.A374_L377del	*CEP104*	US	SCZAFF
chr12: 26816636-26816638	CTG	AGGTCAGTGTC	Deletion + Insertion	Frameshift	c.1693_1695del3ins11, p.Q565fs	*ITPR2*	Afrikaner	SCZ
chr12: 57843366	A	-	Deletion	Frameshift	c.619del1, p.Q207fs	*INHBC*	Afrikaner	SCZAFF
chr16: 30976335	C	-	Deletion	Frameshift	c.1272del1, p.D424fs	*SETD1A*	Afrikaner	SCZ
chr16: 30992058-30992059	AG	-	Deletion	Canonical splice site	c.4582-2_4582-1del2	*SETD1A*	US	SCZ
Identified in the previous analysis								
chr1: 43891777-43891778	-	CA	Insertion	Frameshift	c.2999_3000insCA, p.Q1001fs	*SZT2*	Afrikaner	SCZ
chr1: 54870560	G	A	SNV	Nonsense	c.100C>T, p.Q34*	*SSBP3*	US	SCZAFF
chr1: 97915657	C	T	SNV	Nonsense	c.1863G>A, p.W621*	*DPYD*	US	SCZAFF
chr1: 173450463	A	T	SNV	Canonical splice site	c.96-2A>T	*PRDX6*	US	SCZ
chr1: 180843041-180843042	-	TTGCTTTGTTGCC	Insertion	Frameshift	c.1772_1773insTTGCTTTGTTGCC, p.I592fs	*XPR1*	US	SCZAFF
chr1: 229783350	G	T	SNV	Nonsense	c.4000G>T, p.E1334*	*URB2*	Afrikaner	SCZ
chr3: 9786691	G	C	SNV	Canonical splice site	c.2921-1G>C	*BRPF1*	Afrikaner	SCZ
chr3: 180334392-180334393	TG	-	Deletion	Frameshift	c.2497_2498del, p.Q833fs	*CCDC39*	US	SCZAFF
chr4: 77038831	G	A	SNV	Nonsense	c.1381C>T, p.R461*	*NUP54*	US	SCZ
chr6: 26157174-26157175	AA	-	Deletion	Frameshift	c.556_557del, p.K186fs	*HIST1H1E*	Afrikaner	SCZ
chr6: 33414351	G	A	SNV	Canonical splice site	c.3583-1G>A	*SYNGAP1*	Afrikaner	SCZ
chr6: 129835669-129835675	AAGCCCA	-	Deletion	Frameshift	c.9140_9146del, p.S3050fs	*LAMA2*	Afrikaner	SCZ
chr7: 87839326	T	-	Deletion	Frameshift	c.369delA, p.Q123fs	*SRI*	Afrikaner	Control
chr8: 38107313-38107317	AACTC	-	Deletion	Frameshift	c.1336_1340del, p.G447fs	*DDHD2*	Afrikaner	SCZ
chr8: 53568706-53568707	TC	-	Deletion	Frameshift	c.3682_3683del, p.E1228fs	*RB1CC1*	Afrikaner	SCZ
chr10: 11356223	T	C	SNV	Canonical splice site	c.1003+2T>C	*CELF2*	Afrikaner	SCZ
chr11: 124626163-124626164	-	AGCG	Insertion	Frameshift	c.546_547insCGCT, p.V183fs	*ESAM*	Afrikaner	SCZ
chr12: 53608241	T	A	SNV	Nonsense	c.625A>T, p.K209*	*RARG*	Afrikaner	SCZ
chr16: 57095444	G	A	SNV	Canonical splice site	c.4070+1G>A	*NLRC5*	US	SCZ
chr19: 4338685	A	T	SNV	Nonsense	c.66T>A, p.Y22*	*STAP2*	US	SCZAFF
chrX: 53245326	C	-	Deletion	Frameshift	c.510delG, p.I171fs	*KDM5C*	US	SCZ

SNV; single nucleotide variant, SCZ ; schizophrenia, SCZAFF; schizoaffective disorder

Table S2. Detailed results of functional enrichment analysis in genes with de novo variants

Term	Category	Raw p value	BH-corrected p value	Fold Enrichment	Genes
Intorelant genes with de novo LOF or damaging missense variants in schizophrenia (N = 62)					
chromatin regulator	SP_PIR_KEYWORDS	0.000057	0.005185	10.2	BRPF1, RBBP5, KDM2B, BCORL1, SETD1A, TRRAP, KDM5C
GO:0005626~insoluble fraction	GOTERM_CC_FAT	0.000218	0.017534	3.7	RECK, PITPNM1, UGT1A10, DGCR2, GNAO1, LRP1, HMGCR, RASGRP1, BIRC6, LCT, ITPR2, SLIT3
GO:0005624~membrane fraction	GOTERM_CC_FAT	0.000159	0.025365	3.9	RECK, PITPNM1, UGT1A10, DGCR2, GNAO1, LRP1, HMGCR, RASGRP1, BIRC6, LCT, ITPR2, SLIT3
calcium	SP_PIR_KEYWORDS	0.000910	0.053460	3.9	PITPNM1, LRP1, MACF1, RASGRP1, CACNA1I, CELSR2, ITGA3, NCAN, MBTPS1, ITPR2
GO:0005509~calcium ion binding	GOTERM_MF_FAT	0.000420	0.073911	3.5	PLCL2, PITPNM1, LRP1, MACF1, RASGRP1, CACNA1I, CELSR2, ITGA3, NCAN, MBTPS1, ITPR2, SLIT3
Intorelant genes with de novo LOF or damaging missense variants in controls (N = 77)					
GO:0005083~small GTPase regulator activity	GOTERM_MF_FAT	0.000348	0.083699	5.9	TBC1D15, DOCK2, MADD, IPO7, NF1, CIT, DOCK10, DOCK4
helicase	SP_PIR_KEYWORDS	0.001961	0.072225	9.3	ATRX, SHPRH, DHX37, DHX15, CHD4
SM00487:DEXDc	SMART	0.002548	0.083099	8.5	ATRX, SHPRH, DHX37, DHX15, CHD4
Intorelant genes with de novo less damaging missense or silent variants in schizophrenia (N = 77)					
104.Insulin_signaling	BBID	0.039106	0.088879	25.6	MTOR, IRS1

BH; Benjamini Hochberg, LOF; loss of function

Table S3. Analysis of transmission patterns in the Afrikaner case-control cohort

Cohort	Transmitted	Untransmitted	T:U ratio	RTR (case/control)	p value
Loss-of-function variants (nonsense, canonical splice site and frameshift variants)					
private (hit once in parental population)					
cases	811	1003	0.81	1.39	0.025
controls	159	274	0.58		
rare (non-private and frequency <= 0.05)					
cases	3838	4844	0.79	1.25	0.0053
controls	902	1425	0.63		
common (frequency > 0.05)					
cases	16311	17303	0.94	1.07	0.019
controls	3732	4231	0.88		
MODERATE effect variants					
private					
cases	12287	11589	1.06	1.00	0.563
controls	2286	2155	1.06		
rare					
cases	92066	102054	0.90	1.08	0.023
controls	20603	24663	0.84		
common					
cases	566783	593019	0.96	1.02	0.035
controls	127551	135706	0.94		
LOW effect variants					
private					
cases	7890	7155	1.10	1.01	0.25
controls	1464	1335	1.10		
rare					
cases	79821	85699	0.93	1.03	0.23
controls	17610	19562	0.90		
common					
cases	707981	727649	0.97	1.01	0.11
controls	159769	165749	0.96		

T:U ratio; transmitted to untransmitted ratio, RTR; relative transmission ratio (T:U ratio in cases / T:U ratio in controls). Detailed definition for MODERATE and LOW effect variants are described in the Experimental Procedures. P values were calculated by one-sided permutations

Table S4 Functional enrichment analysis of intolerant genes with private LOF variants in schizophrenia probands

Term	Category	Raw p value	BH-corrected p value	Fold Enrichment	Probability of observing significant enrichment in control intolerant genes*	Genes
nucleotide phosphate-binding region:ATP	UP_SEQ_FEATURE	0.000001	0.000515	2.4	15.4 %	COASY, MYO7A, MAP4K2, DNAH3, PINK1, DNAH2, ACSF3, DHX38, MYO15A, DHX36, GK5, ERCC3, MATK, ERCC2, EHD4, MSH6, TRPM6, SHPRH, PFKL, MYH1, MET, WNK1, ATAD2, PFKM, ACACB, MYH8, DAPK1, RECQL, MAST1, KCNT2, RFC2, CDC42BPA, TNK2, LRRK2, ABL2, LRRK1, MYH7B
GO:0016887~ATPase activity	GOTERM_MF_FAT	0.000009	0.000663	3.3	0.1 %	MSH6, ABCA7, ATP4A, IDE, DNAH3, KATNB1, ATAD2, CFTR, DNAH2, ATP12A, RECQL, ATP13A1, DHX38, ATP9B, RFC2, ATP2A1, ATP8B2, DHX36, ERCC3, ERCC2
ion transport	SP_PIR_KEYWORDS	0.000117	0.004500	2.5	5.8 %	SLC22A17, KCNH1, TRPM4, TRPM3, TRPM6, TRPM8, ATP4A, SLCO4A1, CACNB2, CFTR, ATP12A, KCNU1, SLC4A11, KCNT2, ATP2A1, SLC24A1, RYR1, KCNH6, ANO3, CHRNA6, CACNA1D, SLC22A2, SLC4A5
magnesium	SP_PIR_KEYWORDS	0.000216	0.007476	2.7	1.2 %	ADCY4, ATP4A, PFKL, ITGA11, PINK1, PFKM, ATP12A, ADPRH, MAST1, ATP13A1, ATP9B, ATP2A1, CDC42BPA, ATP8B2, TNK2, ABL2, FAHD2A, LRRK1, ERCC2
GO:0000287~magnesium ion binding	GOTERM_MF_FAT	0.000178	0.009331	2.6	0.2 %	MSH6, ADCY4, OPA1, ATP4A, PFKL, ITGA11, PINK1, PFKM, ATP12A, ADPRH, MAST1, ATP13A1, ATP9B, ATP2A1, CDC42BPA, ATP8B2, TNK2, ABL2, LRRK1, FAHD2A, ERCC2
ionic channel	SP_PIR_KEYWORDS	0.001541	0.043746	2.8	9.3 %	TRPM4, KCNH1, TRPM3, TRPM6, TRPM8, CACNB2, CFTR, KCNU1, KCNT2, RYR1, KCNH6, ANO3, CHRNA6, CACNA1D
GO:0042623~ATPase activity, coupled	GOTERM_MF_FAT	0.001227	0.045086	2.9	3.4 %	ATP4A, KATNB1, CFTR, ATP12A, RECQL, ATP13A1, DHX38, ATP9B, RFC2, ATP2A1, ATP8B2, DHX36, ERCC3, ERCC2
sh3 domain	SP_PIR_KEYWORDS	0.001755	0.045922	3.3	10 %	PLCG1, MYO15A, MYO7A, PLCG2, MPP4, CACNB2, MPP7, UBASH3A, TNK2, ABL2, MATK
calcium transport	SP_PIR_KEYWORDS	0.002135	0.051736	5.3	7.3 %	TRPM3, TRPM6, ATP2A1, SLC24A1, RYR1, CACNB2, CACNA1D
thick filament	SP_PIR_KEYWORDS	0.002576	0.054562	14.2	0.5 %	MYH1, MYOM2, MYH8, MYH7B
cytoskeleton	SP_PIR_KEYWORDS	0.002458	0.055501	2.1	6.6 %	PDLIM7, DNAH3, KATNB1, FTCD, TTLL5, RDX, ANLN, DNAH2, ARHGAP24, SYNPO2L, LLGL1, MAST1, CEP250, FAAH, MYO15A, AVIL, TCHP, FLII, CLIP1, CDK5RAP2, ABL2
GO:0015662~ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism	GOTERM_MF_FAT	0.002492	0.083776	6.3	1.2 %	ATP13A1, ATP4A, ATP9B, ATP2A1, ATP8B2, ATP12A
GO:0005262~calcium channel activity	GOTERM_MF_FAT	0.002863	0.089939	4.9	2.4 %	TRPM4, TRPM3, TRPM6, TRPM8, RYR1, CACNB2, CACNA1D

BH; Benjamini Hochberg, LOF; loss of function. *Probability of observing significant enrichment in iterations in which 309 genes were randomly selected from 3,694 intolerant genes with at least one private LOW effect variants in schizophrenia. Probability > 5 % indicates that the enrichment was likely to be explained by general properties of intolerant genes.

Table S5 Cross-comparison of genes with private LOF variants transmitted to schizophrenia probands

Gene	Position of the LOF variant	Effect of the LOF variant	Reference allele	Varant allele	Associated disease	The LOF varaiants found in EVS?	Any LOF SNVs in the gene found in EVS?	Any LOF SNVs or indels in the gene found in EVS?
Overlapping with genes included in CNVs associated with SCZ, ASD, ID (Malhotra and Sebat, 352 genes, 46 intolerant genes)								
LZTR1	chr16: 28913639	STOP_GAINED	G	T	SCZ, ASD, ID	No	Yes	Yes
RFC2	chr7: 73649897	STOP_GAINED	G	A	ASD, ID	No	Yes	Yes
Overlapping with genes hit by de novo LOF SNVs or Indels in SCZ, ASD and/or ID (170 genes, 85 intolerant genes)								
ALS2CL	chr3: 46729697	STOP_GAINED	C	A	SCZ	Yes	Yes	Yes
KCNU1	chr8: 36664952	STOP_GAINED	C	T	SCZ	Yes	Yes	Yes
SCP2	chr1: 53480591	STOP_GAINED	C	T	ASD	Yes	Yes	Yes
FAM91A1	chr8: 124796762	FRAME_SHIFT	AACTCT	A	ASD	No	No	No
ACACB	chr12: 109696170	SPLICE_SITE_DONOR	G	A	ASD	No	Yes	Yes
SYNCRIP	chr6: 86324503	STOP_GAINED	G	A	ID	No	Yes	Yes
Overlapping with candidate genes for SCZ, ASD and/or ID (Szgene, SFARI gene and Neale et al., 1659 genes, 557 intolerant genes)								
ARHGEF10	chr8: 1851668	FRAME_SHIFT	C	CT	SCZ (SZgene)	No	Yes	Yes
BRD1	chr22: 50170734	STOP_LOST	C	G	SCZ (SZgene)	No	Yes	Yes
CABIN1	chr22: 24530341	FRAME_SHIFT	CAG	C	SCZ (SZgene)	No	Yes	Yes
DGKH	chr13: 42772730	SPLICE_SITE_DONOR	G	A	SCZ (SZgene)	No	Yes	Yes
FAAH	chr1: 46871747	FRAME_SHIFT	GC	G	SCZ (SZgene)	No	No	Yes
KCNH1	chr1: 210856651	STOP_GAINED	G	C	SCZ (SZgene)	Yes	Yes	Yes
MAGI2	chr7: 77975249	FRAME_SHIFT	TG	T	SCZ (SZgene)	No	Yes	Yes
PER2	chr2: 239164301	STOP_GAINED	G	A	SCZ (SZgene)	No	Yes	Yes
PLCG1	chr20: 39794927	FRAME_SHIFT	CCTCT	C	SCZ (SZgene)	No	Yes	Yes
PSEN2	chr1: 227075813	START_LOST	A	G	SCZ (SZgene)	No	Yes	Yes
SLC6A9	chr1: 44463228	FRAME_SHIFT	AG	A	SCZ (SZgene)	No	Yes	Yes
VIPR1	chr3: 42577608	STOP_GAINED	G	A	SCZ (SZgene)	No	Yes	Yes
APC	chr5: 112174750	FRAME_SHIFT	TGAA	T	SCZ (SZgene), ASD (SFARI)	No	Yes	Yes
MET	chr7: 116415034	FRAME_SHIFT	CCAGTCCATTACTG	C	SCZ (SZgene), ASD (SFARI)	No	No	Yes
VLDLR	chr9: 2643158	SPLICE_SITE_ACCEPTOR	A	T	SCZ (Szgene), ID (Neale et al.)	No	Yes	Yes
CNTNAP2	chr7: 147092775	STOP_GAINED	C	T	ASD (SFARI and Neale et al.)	No	Yes	Yes
ARHGAP24	chr4: 86898749	FRAME_SHIFT	CG	C	ASD (SFARI)	No	Yes	Yes
CACNA1D	chr3: 53843987	SPLICE_SITE_ACCEPTOR	A	C	ASD (SFARI)	No	Yes	Yes
CNTNAP5	chr2: 125405475	STOP_GAINED	G	T	ASD (SFARI)	No	Yes	Yes
DAPK1	chr9: 90261374	SPLICE_SITE_ACCEPTOR	A	C	ASD (SFARI)	No	No	Yes
THRA	chr17: 38249276	STOP_GAINED	A	T	ASD (SFARI)	No	Yes	Yes
CDH15	chr16: 89256654	FRAME_SHIFT	CT	C	ID (Neale et al.)	No	Yes	Yes
CDK5RAP2	chr9: 123291052	SPLICE_SITE_ACCEPTOR	C	T	ID (Neale et al.)	No	Yes	Yes
ERCC2	chr19: 45855804	FRAME_SHIFT	CT	C	ID (Neale et al.)	No	Yes	Yes
ERCC3	chr2: 128047338	FRAME_SHIFT	CGGATCACG	C	ID (Neale et al.)	No	Yes	Yes
RDH12	chr14: 68196053	FRAME_SHIFT	CGCCCT	C	ID (Neale et al.)	Yes	Yes	Yes
SLC2A1	chr1: 43395157	STOP_GAINED	T	A	ID (Neale et al.)	No	No	No

ASD; autism spectrum disorder, CNV; copy number variant, EVS; exome variant server, ID; intellectual disability, LOF; loss-of-function, SCZ; schizophrenia

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Cohorts

Detailed information for the cohorts was described in our previous study (Xu et al., 2012). Briefly, studied samples comprise trios collected from two distinct populations: the Afrikaner population from South Africa (European, mostly Dutch descent) (146 trios with SCZ probands and 34 control trios) and the US population (Northern European descent) (85 trios with SCZ probands). Of the 146 Afrikaner case probands, 122 (83.6%) had a diagnosis of SCZ, and 24 (16.4%) a diagnosis of schizoaffective disorder. Of the 85 US probands, 46 (54.1%) had a diagnosis of SCZ, and 39 (45.9%) of schizoaffective disorder. Subjects in control families were screened against presence and history of treatment for any psychiatric condition, as well as history of mental illness in first- or second-degree relatives. These subjects were recruited and characterized in the context of our ongoing large-scale genetic studies of SCZ (Xu et al., 2012; Xu et al., 2011; Xu et al., 2008; Xu et al., 2009). In the Afrikaner cohort, absence of SCZ in first- or second-degree relatives was confirmed by detailed medical records over several generations in the local recruiting hospital. In the US cohort, we were able to determine absence of disease in first-degree relatives. Paternity and maternity were confirmed before sequencing via the Affymetrix Genome-Wide Human SNP Array 5.0 as well as via a panel of microsatellite markers. DNA for all study subjects was extracted from whole blood (not cell lines), and analysis was performed blind to affected status while maintaining knowledge of the parent-child relationships. Informed consent was obtained from all participants, and the Institutional Review Committees of Columbia University and the University of Pretoria approved all procedures.

Exome library construction and generation of BAM files

Detailed information for the procedures used to construct the exome libraries and the analytical pipeline for exome sequencing data have been described elsewhere (Xu et al., 2012). Briefly, genomic DNA (~3 μg) was sheared to 200-300 bp in size using a Covaris Acoustic Adaptor (Covaris, Inc., Woburn, MA). Exonic DNA was captured using Agilent SureSelect v2 (Agilent Technologies, Santa Clara, CA, n = 85 trios) or NimbleGen SeqCap EZ v2 (Roche NimbleGen, Mannheim, Germany, n = 180 trios). Each library was quantified by PicoGreen (Invitrogen, Eugene, OR), and fragment size was measured with the Agilent Bioanalyzer (Agilent Technologies). 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 using the Kapa qPCR kit (Kapa Biosystems, Woburn, MA). Each library was normalized to a 10 nM concentration and sequenced using the Illumina HiSeq 2000 [(Illumina, San Diego, CA). Raw sequencing data were mapped to the human reference genome (hg19) using the Burrows-Wheeler Aligner (BWA, v0.5.81536) (Li and Durbin, 2009). The Genome Analysis Toolkit (GATK, version 5091) (McKenna et al., 2010) was used to remove duplicates, perform local realignment and map quality score recalibration to produce cleaned BAM files.

Analysis of *de novo* LOF indels

Short to middle-sized indels were called by PRISM (Jiang et al., 2012) and Pindel (Ye et al., 2009) with default parameters using cleaned BAM files (See Supplemental Experimental Procedures). Output files were annotated by ANNOVAR (Wang et al., 2010). To narrow down the list of candidates for *de novo* LOF indels (frameshift variants, variants affecting canonical splice sites and variants disrupting one or more exons), the following criteria were applied to the variant calls: 1) number of the unique supporting reads ≥ 5; 2) read-depth (average of the read-coverage in the upstream and downstream ten bases of the candidate indel position) ≥ 10; 3) supporting read number/read-depth ≥ 0.05; 4) not found in the list of the indel candidates from the

healthy subjects in our cohort (for probands in control trios, all of the other healthy subjects); 5) (Used only for the data from PRISM) number of the unique supporting reads with alignment scores greater than or equal to 350 ≥ 2, and contig match scores ≠ 999 or -999. In addition to the high-reliability candidates satisfying all of these criteria, we picked up moderate-reliability candidates (with three or four supporting reads and satisfying criteria 2–5) in genes in which LOF *de novo* variant(s) were identified in the previous WES studies for SCZ (Girard et al., 2011; Gulsuner et al., 2013; Xu et al., 2012; Xu et al., 2011) or high-reliability candidate(s) were detected in our present analysis from our case subjects, to maximize the opportunity to identify genes recurrently affected by *de novo* LOF mutations in SCZ. All candidates for *de novo* indels (22 high-reliability / one moderate-reliability candidate in cases, and one high-reliability candidate in controls from PRISM; 26 high-reliability / four moderate-reliability candidates in cases, and six high-reliability candidates in controls from Pindel) were first manually inspected using the Integrative Genomics Viewer (Robinson et al., 2011). The candidates with no read(s) supporting indel in the proband or with multiple reads supporting indel in either of the parents were excluded from further analysis. The remaining candidates were then subjected to Sanger sequencing for validation.

Assessment of the probability to observe at least two *de novo* LOFs in the same gene

We used a simulation procedure to estimate empirically the probability of observing at least two *de novo* LOFs hitting the same gene, based on gene length and GC content.

First, we simulated a set of 50,000 positions uniformly distributed in the capture regions described in Roche NimbleGen "SeqCapEZ_Exomev2.0" target region and annotation files. To simulate splice site positions, we extended the boundary of each exon 10 bp in each direction to cover the 20 bp region surrounding the exon boundary. We then selected 10,000 positions within 10 bp surrounding the exon boundary. For each of these random positions we obtained the GC content as follows: we determined

whether the base at the chosen position is G/C or A/T using the hgGcPercent program (http://genomewiki.ucsc.edu/index.php/Kent_source_utilities). The GC content was taken into account when the actual mutations were simulated. We simulated 100,000 datasets, each containing the number of LOF variants in our original data (a total of 25 variants with 18 coding LOFs and 7 splice site LOFs). Each of these mutations was generated by first randomly choosing a position from the 50,000 positions for coding LOF variants, and from the 10,000 positions for splice site LOF variants. Then a mutation was generated at the chosen position depending on the GC content at the position, as follows: if G/C then with probability 1 a mutation was generated; if A/T then the probability of a mutation was 1/1.76 = 0.57 (the mutation rate at GC bases is 1.76 fold higher than at AT bases [Sanders et al., 2012]). To assess the significance of observing at least one gene with any combination of two or more LOF variants (e.g. two coding LOFs, two splice site LOFs or each one coding and splice site LOFs in the case of a gene with two LOF hits), we counted for each simulated dataset the number of genes identified to harbor two or more LOF mutations. The p value was calculated as the proportion of datasets where at least one gene has two or more LOF mutations. Note that the resulting p value is experiment-wide (exome-wide) p value.

We have used an additional method to evaluate the significance of observing at least two *de novo* LOF mutations in *SETD1A* by estimating the probability of a LOF *de novo* mutation per chromosome specific for the *SETD1A* locus. To calculate a specific local mutation rate, we used TADA software (He et al., 2013) with default parameters, which estimates the mutation rate per gene based on its exonic length and its nucleotide content (Sanders et al., 2012). Then, we used a Poisson model for the probability of observing two or more LOF *de novo* events in this gene. Bonferroni correction for multiple testing was performed using the number of all genes (~20,000 genes).

Genotyping of inherited SNVs and indels from exome data

SNV calls were made by GATK for all trios jointly. Conventional indel calls were made by Dindel software (Albers et al., 2011) using one cleaned BAM file per run. 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 SNVs used in this study were restricted to sites that passed the standard GATK filters to eliminate SNVs with strand bias, low quality for the depth of sequencing achieved, homopolymer runs, and SNVs near indels. The indel variants were restricted to sites that passed the Dindel filters to eliminate the variants where reference homopolymer length was longer than 10, read quality below 20 and nonreference allele was not covered by at least one read on both strands. The resulting VCF files were merged into one file with VCFtools (Danecek et al., 2011) and the effect of the variants in the merged VCF files were further annotated by Anntools (Makarov et al., 2012) that was built on top of SnpEff (Cingolani et al., 2012). According to the definition in SnpEff, we defined 2,029 nonsense variants, 395 variants disrupting start and stop codons, 1,524 canonical splice site variants and 3,058 frameshift variants as "LOF" variants; 114,376 missense variants and 1,813 inframe indels were categorized as variants with "MODERATE" effect; and 83,573 silent variants, 2,811 start codon-generating variants in untranslated regions and 19 nonsynonymous variants that generate alternative start codons as variants with "LOW" effect. Additional annotation of evolutionary impact and damaging prediction were derived using ANNOVAR (Wang et al., 2010) when needed.

Transmission analysis of inherited variants

The global T:U ratio was derived for each sample group and variant type. We excluded from further analysis variants that showed Mendelian errors in one or more family to

ensure that only high quality variants are analyzed. The gene-variant relationship was determined by SnpEff annotation. To assess the statistical significance of the relative overtransmission of LOF variants in affected individuals as compared to controls, we performed permutation testing by randomly permuting the case/control labels of the trios in our datasets. To avoid potential confounding due to differences in DNA capture kits used for construction of exome libraries, labels for capture methods (Agilent or NimbleGen) were used as covariates in our permutation procedures. We generated 100,000 permuted datasets, and for each such permuted dataset we calculated the corresponding RTR ratios of LOF variants. We calculated the one-sided p value as the proportion of permutations where the statistic RTR is greater than or equal to the one observed in the original dataset.

Functional enrichment analysis

Functional gene-set enrichment analyses were performed for intolerant genes that carry *de novo* damaging variants and for intolerant genes with private inherited LOF variants in SCZ by using DAVID (The Database for Annotation, Visualization and Integrated Discovery) (Huang da et al., 2009) version 6.7 with default databases. Biological modules containing more than 1,000 genes were excluded from the result, as these modules included too broad functional categories of genes and were thus less likely to specify biological processes related to SCZ. To correct for potential biases in the analyses, we performed DAVID analyses using control gene sets, as follows: intolerant genes hit by *de novo* damaging variants in control subjects (N = 77) and intolerant genes hit by *de novo* but less damaging variants (silent variants and missense variants with PolyPhen-2 scores < 0.95) in SCZ (N = 77) for the analysis of intolerant genes with *de novo* damaging variants in SCZ; and intolerant genes hit by private LOW effect variants in SCZ for the analysis of intolerant genes with private inherited LOF variants in SCZ. Intolerant genes hit by private LOF variants in control subjects were not

used as a control gene set because the number of such genes (N = 29) was not sufficient to obtain good statistical power. These analyses allow us to identify potential modules that are not specific to SCZ, but rather related to properties of intolerant genes, or other factors such as gene size and GC content. In the control analysis for private inherited LOF variants, we kept the number of genes the same as in the LOF case (namely 309) and performed repeated random samplings of 309 genes from the total set of 3,694 intolerant genes with at least one private LOW effect variant in SCZ. For a specific biological module that we found to be enriched in the analysis of private inherited LOF variants in SCZ (e.g. for the term "nucleotide phosphate-binding region:ATP"), we estimated its enrichment among intolerant genes with private LOW effect variants empirically by counting the number of times this particular module appears among the significant modules with p values lower than that in the original analysis (e.g. p values lower than 0.000515 for the term "nucleotide phosphate-binding region:ATP") in the DAVID applications to these random gene sets (we chose 1,000 such random sets). If this number was greater than 50 (5% of the random samplings) we concluded that the module is probably not specific to SCZ, but rather related to properties of intolerant genes.

Construction of gene-gene interaction network

For gene-gene network construction, GeneMANIA (Zuberi et al., 2013) was used with default parameters with these exceptions: 1) to generate a network figure specifically consisting of input genes, no related genes and attributes were displayed; 2) for the selection of gene-expression data sources, data not derived from brain or neuronal tissues was excluded; 3) to avoid results to be dominated by a single data type and/or a single data source, network weighting was performed using the "Equal by data type" option, and data sources with extremely large number of connections (i.e. more than a million connections) were excluded.

Analysis of correlation between clinical variables and per-individual variant status

We performed non-parametric correlation analysis (Spearman's rank correlation), between four clinical variables (severity and functional outcome of the disease, age at disease onset, history of childhood learning disabilities, and comorbidity of mental retardation), and five genotypic scores for each SCZ proband (number of private LOFs in intolerant genes, number of private and rare LOFs in intolerant genes, number of private LOFs in all genes, number of private and rare LOFs in all genes, and number of *de novo* LOFs in all genes). All four clinical variables were evaluated during the in-person diagnostic assessment and the administration of the Diagnostic Interview for Genetic Studies (DIGS). Severity of the disease and functional outcome was scored as 1 or 2 for cases with an episodic shift of disease or mild deterioration, where in between periods of illness there were periods of return to near normality (1=Episodic Shift and 2=Mild Deterioration). Scores of 3, 4, or 5 represent a "downhill" course culminating in social and occupational incapacitation (3=Moderate Deterioration, 4=Severe Deterioration, and 5=Stable Level of Severe Decline). Age at disease onset was defined as the age at which full DSM-IV criteria for SCZ or schizoaffective disorder were first met. History of learning difficulties was recorded as positive if there had been a diagnosis of a learning disability, or clear history of being a "slow learner", requiring remediation at school, or placement in a special class. Finally, mental retardation was considered as comorbidly present if there was a record of a clear diagnosis made by the school or a doctor during childhood. As we performed a total of 20 pairwise correlation analyses, we considered correlations with $p < 0.0025$ (= 0.05/20) as statistically significant.

Prioritization of candidate genes with residual variation intolerance score (RVIS)

A residual variation intolerance score (RVIS) and a percentile rank for each gene were obtained from Petrovski et al., 2013. We defined the 25th percentile of the most intolerant genes assessed by RVIS as "intolerant", because this threshold was shown to be useful in enriching for genes associated with neurodevelopmental diseases. (Petrovski et al., 2013). Enrichment of intolerant genes among genes hit by each class of variants (silent, missense, LOF+missense, or LOF) was evaluated by binomial exact tests, with hypothesized probability of success = 0.38. This hypothesized probability was the theoretical likelihood of finding *de novo* variants in the 25th percentile intolerant genes, considering the gene size. Quantitative scores from PolyPhen-2 prediction (Adzhubei et al., 2010) for the two-dimensional (2D) plotting of RVIS and PolyPhen-2 scores were obtained by using SeattleSeq Variation Annotation (http://snp.gs.washington.edu/SeattleSeqAnnotation137). In instances where one missense variant caused amino acid substitution of multiple transcripts, the highest score was used.

SUPPLEMENTAL TEXT

Clinical histories of the two schizophrenia patients carrying *de novo* LOF mutations in the *SETD1A* gene

Patient 5-33 (JAS) carries the *de novo* frameshift indel variant D424fs. At the time of recruitment, he was a 38-year old divorced male staying with his parents and working as a security officer. His first psychotic break was at age 21, although his school performance started declining when he was 16. As a child, he reached his developmental milestones late. He exhibited separation anxiety, was a tense child and scared of the dark. He was admitted to a psychiatric hospital at 21 in a psychotic state. He had thought process disturbance, paranoid and other delusions. He was married for 10 years but due to his illness and the paranoid delusions interfering with his daily life his wife left. In addition to his schizophrenia he has been bothered with recurrent

thoughts that force him to perform certain behaviors and rituals meeting diagnostic criteria for OCD. Onset of OCD seems to be contemporaneous with schizophrenia at 21, although the parents report that he had certain rituals as a child. His illness, since the onset, has followed a moderately deteriorative course. The subject, as a child before the age of 10, demonstrated the following five early deviant behaviors: social isolation, excessive fears, inattentiveness, learning difficulties, and odd behavior that included OCD-like rituals.

Patient 20-162 (JN) carries the *de novo* indel variant c.4582-2_4582-1del2 that changes the canonical splice acceptor site sequence adjacent to exon 16 from AG to GG. At the time of the interview, he was a 20-year old single man who lived with his parents at home. He started experiencing strange obsessive thoughts followed by compulsive behaviors since 4th grade, which became worse by the time he reached 7th grade. He meets full diagnostic criteria for OCD. His OCD symptoms diminished but persisted. During high school, he began to become more socially withdrawn. He started saying strange things to others and was engaging in odd and disorganized behavior. Since that first episode of psychosis he has not returned to his previous level of functioning. He has had persistent difficulty with motivation, flattened affect, disorganized behavior, social isolation and delusional thinking. He meets full diagnostic criteria for SCZ with an onset age at 18. The patient was inattentive as a child and a slow learner, although he was never diagnosed with a learning disability and finished high school. He also reached developmental milestones with delays. He has vocal tics when under stress.

Detailed results for DAVID gene-set enrichment analyses

Genes with de novo variants

We performed a DAVID (The Database for Annotation, Visualization and Integrated Discovery) (Huang da et al., 2009) gene-set enrichment analysis using 62 genes that are intolerant to variation and harbor *de novo* LOF or damaging missense (PolyPhen-2

scores ≥ 0.95) variants as an input. In this analysis, “chromatin regulator” was the most significantly enriched term (Benjamini Hochberg [BH])-corrected p = 0.005, fold enrichment = 10.2, Figure 2A, detailed lists of genes in each term are shown in Table S2). Other proteins related to calcium signaling (“calcium” and “calcium ion binding [GO: 0005509]”) and localized to membrane fraction (“insoluble fraction [GO: 0005626]” and “membrane fraction [GO: 0005624]”) were nominally overrepresented. Enrichment for these terms was not observed among intolerant genes hit by *de novo* LOF or damaging missense variants in controls from various cohorts (number of genes = 77) (Gulsuner et al., 2013; Iossifov et al., 2012; O'Roak et al., 2012b; Rauch et al., 2012; Sanders et al., 2012; Xu et al., 2012), or among intolerant genes hit by less damaging missense (PolyPhen score < 0.95) or silent variants in cases (number of genes = 77, Table S2, terms with BH-corrected p < 0.1 are shown; none of the terms is statistically significant).

Genes with inherited variants

We performed a DAVID analysis on a total of 309 unique intolerant genes that carry at least one private LOF variant that was transmitted to the cases. After excluding terms found to be significantly enriched among intolerant genes with at least one private LOW effect variant (see Experimental Procedures and Table S4), which are likely to reflect a general property of intolerance to genetic variation, we observed significant enrichment of terms related to ATPase activity (“ATPase activity [GO: 0016887]”, BH-corrected p = 0.0007, fold enrichment = 3.3 and “ATPase activity, coupled [GO:0042623]”, BH-corrected p = 0.045, fold enrichment = 2.9), which include transmembrane ATPase (e.g. *ATP2A1*, *ATP4A* and *ATP12A*) and ATP dependent DNA helicase genes (e.g. *ERCC2* and *ERCC3*), as well as of terms related to magnesium ion (“magnesium”, BH-corrected p = 0.0074, fold enrichment = 2.7 and “magnesium ion binding [GO:0000287]”, BH-corrected p = 0.0093, fold enrichment = 2.6), including various genes such as *ADCY4* encoding adenylate cyclase type 4, *PINK1* encoding a

gene linked to Parkinson's disease, and several transmembrane ATPase genes described above (Table S4).

Transmission analysis using geographically matched cases and controls

When we restricted the transmission analysis to the geographically matched Afrikaner case trios (n = 146) and control trios (n = 34), excluding all European American case trios (n = 85), the results for RTRs of LOF, MODERATE and LOW effect variants remained very similar to those observe for the combined group (i.e. Afrikaner and European American) (Table S3). This result indicates that our observation was not influenced by the combination of two population groups. This is consistent with a recent report that demographic history has likely had little impact on the importance of rare variants for most complex traits (Simons et al., 2013).

Cross-comparison between intolerant genes harboring private LOF variants in SCZ and lists of genes implicated in psychiatric diseases

To explore individual promising candidate genes further, we cross-compared the list of intolerant genes harboring at least one private LOF variant in SCZ with the following lists of genes from previously published literature: 1) genes included in CNVs with replicated evidence of association with SCZ, ASD and/or ID (Malhotra and Sebat, 2012); 2) genes hit by *de novo* LOF SNVs or indels in SCZ, ASD and/or ID (de Ligt et al., 2012; Girard et al., 2011; Gulsuner et al., 2013; Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012b; Rauch et al., 2012; Sanders et al., 2012; Vissers et al., 2010; Xu et al., 2012); and 3) curated candidate gene lists for SCZ (SZgene, [Allen et al., 2008]), ASD (the list in Neale et al., 2012 and SFARI gene [Abrahams et al., 2013]) and ID (Neale et al., 2012) (Table S5).

The result of cross-comparison with genes included in CNVs with replicated evidence of association with SCZ, ASD and/or ID is described in the main text.

We found six genes hit by *de novo* LOF variants in previous studies of SCZ, ASD and/or ID harboring private LOF variants in our sample. *De novo* LOF variants in *KCNU1* and *ALS2CL* were reported in SCZ (Girard et al., 2011; Gulsuner et al., 2013). *De novo* LOF variants in *SCP2*, *FAM91A1* and *ACACB*, were observed in ASD cases (Iossifov et al., 2012; Sanders et al., 2012), while *SYNCRIP* was hit by a *de novo* LOF variant in ID (Rauch et al., 2012).

From the list of curated candidate genes for neuropsychiatric and neurodevelopmental disorders, a total of 27 genes were identified as harboring private LOFs in our sample (Table S5). 15 of these genes were found in a database of SCZ candidate genes. Eight of these genes were found in curated candidate gene lists for ASD and seven of these genes were among a list of genes known to be involved in ID. Interestingly, patients carrying private LOF variants in genes solely implicated in ID (*SYNCRIP*, *CDH15*, *CDK5RAP2*, *ERCC2*, *ERCC3*, *RDH12* and *SLC2A1*) appear more likely to demonstrate childhood learning difficulties (42.9% [3/7] in patients with private LOFs in ID genes and 15.2% [24/158] in the rest of the patients whose history for childhood learning difficulties was available, $p = 0.044$, one-tailed Fisher's exact test).

SUPPLEMENTAL REFERENCES

Abrahams, B.S., Arking, D.E., Campbell, D.B., Mefford, H.C., Morrow, E.M., Weiss, L.A., Menashe, I., Wadkins, T., Banerjee-Basu, S., and Packer, A. (2013). SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Molecular autism *4*, 36.

Allen, N.C., Bagade, S., McQueen, M.B., Ioannidis, J.P., Kavvoura, F.K., Khoury, M.J., Tanzi, R.E., and Bertram, L. (2008). Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. Nature genetics *40*, 827-834.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T*., et al.* (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156-2158.

Gudjonsson, T., Altmeyer, M., Savic, V., Toledo, L., Dinant, C., Grofte, M., Bartkova, J., Poulsen, M., Oka, Y., Bekker-Jensen, S*., et al.* (2012). TRIP12 and UBR5 suppress spreading of chromatin ubiquitylation at damaged chromosomes. Cell *150*, 697-709.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Makarov, V., O'Grady, T., Cai, G., Lihm, J., Buxbaum, J.D., and Yoon, S. (2012). AnnTools: a comprehensive and versatile annotation toolkit for genomic variants. Bioinformatics *28*, 724-725.

Simons, Y.B., Turchin, M.C., Pritchard, J.K., and Sella, G. (2013). The deleterious mutation load is insensitive to recent population history. arXiv:13052061.

Vissers, L.E., de Ligt, J., Gilissen, C., Janssen, I., Steehouwer, M., de Vries, P., van Lier, B., Arts, P., Wieskamp, N., del Rosario, M., et al. (2010). A de novo paradigm for mental retardation. Nature genetics 42, 1109-1112.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research *38*,

e164.

Xu, B., Woodroffe, A., Rodriguez-Murillo, L., Roos, J.L., van Rensburg, E.J., Abecasis, G.R., Gogos, J.A., and Karayiorgou, M. (2009). Elucidating the genetic architecture of familial schizophrenia using rare copy number variant and linkage scans. Proceedings of the National Academy of Sciences of the United States of America *106*, 16746-16751.

Zuberi, K., Franz, M., Rodriguez, H., Montojo, J., Lopes, C.T., Bader, G.D., and Morris, Q. (2013). GeneMANIA prediction server 2013 update. Nucleic acids research *41*, W115-122.