De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability

SE McCarthy[1], J Gillis[1], M Kramer[1], J Lihm[1], S Yoon[1], Y Berstein[1], M Mistry[2], P Pavlidis[2], R Solomon[1], E Ghiban[1], E Antoniou[1], E Kelleher[3], C O'Brien[3], G Donohoe[3], M Gill[3], DW Morris[3], WR McCombie[1,4] and A Corvin[3]

Schizophrenia is a serious psychiatric disorder with a broadly undiscovered genetic etiology. Recent studies of *de novo* mutations (DNMs) in schizophrenia and autism have reinforced the hypothesis that rare genetic variation contributes to risk. We carried out exome sequencing on 57 trios with sporadic or familial schizophrenia. In sporadic trios, we observed a ~ 3.5-fold increase in the proportion of nonsense DNMs (0.101 vs 0.031, empirical $P = 0.01$, Benjamini–Hochberg-corrected $P = 0.044$). These mutations were significantly more likely to occur in genes with highly ranked probabilities of haploinsufficiency ($P = 0.0029$, corrected $P = 0.006$). DNMs of potential functional consequence were also found to occur in genes predicted to be less tolerant to rare variation ($P = 2.01 \times 10^{-5}$, corrected $P = 2.1 \times 10^{-3}$). Genes with DNMs overlapped with genes implicated in autism (for example, *AUTS2*, *CHD8* and *MECP2*) and intellectual disability (for example, *HUWE1* and *TRAPPC9*), supporting a shared genetic etiology between these disorders. Functionally *CHD8*, *MECP2* and *HUWE1* converge on epigenetic regulation of transcription suggesting that this may be an important risk mechanism. Our results were consistent in an analysis of additional exome-based sequencing studies of other neurodevelopmental disorders. These findings suggest that perturbations in genes, which function in the epigenetic regulation of brain development and cognition, could have a central role in the susceptibility to, pathogenesis and treatment of mental disorders.

INTRODUCTION

Schizophrenia is a complex brain disorder affecting perception, thinking, behavior, cognition and social functioning. The disorder affects about 1% of the adult population and is a huge burden for those diagnosed, their families and society. As is the case with most psychiatric disorders, schizophrenia is a syndromal diagnosis based on observed behavior, duration of symptoms and impaired function rather than on a biological understanding of disease etiology. This has significantly hindered progress in developing more precise diagnosis and better therapeutics to improve patient outcomes.

Schizophrenia is substantially heritable making it a target for human genetics research. As genomic technologies have improved, a wide spectrum of genetic risk factors has emerged, encompassing common and rare risk variants, but also suggesting significant genetic heterogeneity within the patient population. Published genome-wide association studies (GWASs) have confirmed at least 20 common loci of small effect,[1,2] with many more likely to be detected as sample sizes increase.[3] From genome-wide association study data, it has also been possible to estimate that common risk variants account for at least a quarter of the genetic contribution to schizophrenia risk[4] and that genetic risk

overlaps with other psychiatric disorders, in particular bipolar disorder.[5,6] Studies of rare variation identified recurrent copy-number variants (CNVs) that have a moderate or large effect on schizophrenia risk[7–9] but also implicate *de novo* mutation (DNM) mechanisms as a critical source of private, large effect risk variants in schizophrenia.[10,11] Significantly, almost all of the confirmed CNVs are also risk factors for other neurodevelopmental disorders including autism, intellectual disability (ID) and seizure disorder. In many instances, for example, the 1q21.1 deletion originally identified as a risk factor for schizophrenia, the CNV actually has a substantially greater effect on risk for developmental delay, ID and autistic spectrum disorder.[12–14]

Exome sequencing studies of parent offspring trios[15,16] and the accrued risk associated with greater paternal age[17] suggest that an increased rate of DNMs disrupting gene function (for example, missense and nonsense mutations), could have a significant role in schizophrenia susceptibility. Similar findings have been reported for other severe neurodevelopmental disorders, including autism[18–21] and ID.[22–24] Although the rate of functional DNMs may be increased in neurodevelopmental disorders, the genetic heterogeneity, the abundance of loss of function mutations in the genome of healthy individuals and the abridgement in our

understanding of immediate mutational effects on gene function and downstream biological processes make pinpointing or prioritizing specific mutations difficult. In addition to more exome sequencing studies of trios with neurodevelopmental disorders, analytical approaches that overcome inherent analytical biases (for example, the limited curation of biological resources) are necessary to elucidate disease pathogenesis.

In this study, we have sequenced the exome of 171 individuals representing 42 sporadic and 15 familial trios with schizophrenia or a related psychotic condition to identify additional risk mutations. In our primary analysis, we test the hypothesis that the rate of functional DNM is increased in the sporadic and familial trios group compared with the expected rate in unaffected individuals. We perform a hypothesis-free over-representation analysis using the ontology and annotations from Neurocarta to determine whether the genes with DNM were enriched in other neurodevelopmental disorders.[25] We assess whether genes with DNMs in our data set are over-represented among highly specific chromatin remodeling genes based on protein domain data[26,27] and in chromatin remodeling genes previously implicated in mental disorders.[28] Finally, we evaluate the robustness of our findings in recently published exome data sets of schizophrenia, autism and ID trios.

MATERIALS AND METHODS

Sample collection
All participants gave written informed consent according with local research ethics committee approval. Participants were screened for psychiatric disorder by a trained clinician and cases were interviewed using a structured clinical interview (Structured Clinical Interview for DSM-IV (SCID-P) (ISBN:0880489324).[29] Diagnosis of a major psychotic disorder was made by the consensus lifetime best estimate method using DSM-IV criteria with all available information including interview, family history (or staff) report and chart review. All cases were over 18 years of age, of Irish origin (having all four grandparents born in Ireland) and had been screened to exclude substance-induced psychotic disorder or psychosis owing to a general medical condition. Family history was defined by the absence or presence of psychosis in 1° or 2° relatives. Further details on ascertainment methodology are provided in Strange et al.[30] and information on the family history, age at onset, illness course and other clinical indices of included trios are provided in Supplementary Information.

Exome capture and sequencing
Exome capture DNA library was performed using the Solution Phase Exome Capture method,[31] which is a compilation and optimization of the Bioo Scientific NEXTflex (Illumina Compatible, San Diego, CA, USA) Sequencing Kit and NimbleGen (Madison, WI, USA), SeqCap EZ Library and Technology Note: Targeted Sequencing with NimbleGen SeqCap EZ Libraries and Illumina TruSeq DNA Sample Preparation kit. Briefly, sonicated genomic DNA ranging from 1–5 μg was used to create TruSeq Barcoded libraries. Approximately 1 μg of the precapture library was hybridized with the NimbleGen's SeqCap EZ Human Exome Library v2.0 probes (NimbleGen SeqCap EZ Exome User's Guide and TechNote for paired-end libraries.) for 72 h at 47 °C. Following dual PCR enrichment and quality control evaluation, samples with Bioanalyzer traces resulting in broad peaks ranging from 250–850 bp and producing the highest peak around 400 bp (DNA insert plus adaptors) were pooled. Libraries were sequenced on three or four lanes on a HiSeq 2000, with a Paired-End 101 run including a seven-reads indexing run for the barcode detection. Additional methodological details are provided in the Supplementary Information.

Data processing and variant calling
Sequence reads from the Illumina HiSeq 2000 runs were demultiplexed using the Illumina Casava v1.8 pipeline, aligned to hg19 using the BWA aligner,[32] allowing two mismatches in the 30-base seed. Alignments were then paired, imported to binary (bam) format, sorted and indexed using SAMtools.[32] Picard was then used to fix any mate-pair information altered by the sorting. Bamtools[33] was used to filter alignments to retain only properly paired reads (reads aligned with appropriate insert size and orientation). PCR duplicates were removed using Picard. Bamtools[33] was then used to select alignments with a minimum mapping quality score of 20. Target coverage for each NimbleGen exome capture was assessed using Picard's HSmetrics utility, and both depth and breadth of coverage were reviewed for each sample. The Genome Analysis Toolkit,[34] GATK, was used for local read realignment around indels, and for base quality score recalibration using corrections for base position within the Illumina read, for sequence context and for platform-reported quality. Filtering criteria are provided in the Supplementary Materials. Finally, the variant calls were processed with snpEff v 2_0_5b[35] to provide annotation.

De novo single nucleotide variant (SNV) and INDEL discovery
Any proband SNV that was not present in either parent was considered for further analysis. All proband variants required >10× in all members of the trio. Variants were also filtered for segmental duplications and presence in the Exome Variant Server (EVS6500). Indels were filtered similar to Iossifov et al.,[18] any indel called in a proband was removed from further analysis if at least one read from either parent backed the indel call. In addition, 15 reads supporting the reference allele were required in all three members of the trio. Finally, at least 5% of the reads in the proband were required to support the indel. Further details are provided in the Supplementary Materials.

Variant annotation
Putative de novo variants affecting protein function were selected based on the SNPEFF annotations of these data. In line with the annotation of existing data sets ANNOVAR was used to annotate the selected SNVs and Indels.[36] Scores from SIFT,[37] PolyPhen2 (Adzhubei et al.[38]), LRT and[39] MutationTaster[40] were used to predict the amino-acid effects on protein function. PhyloP,[41] GERP ++ (Davydov et al.[42]) and phastCons46wayscores were determined to measure conservation of variant sites. The variants were then tested for allele frequency in the 1000 Genomes and again by the Exome Variant Server (6500).

Confirming relatedness
Trio relatedness was confirmed using a combination of two measures, the Glaubtiz score[43] and the Square Sum of the difference in the number of alternative alleles between two individuals. We used a set of 17 855 common single-nucleotide polymorphisms with alternative allele frequency between 0.45 and 0.55 with thresholds for relatedness scores at 0.79 and 9000 for the Glaubitz Score and Square Sum, respectively.

SNV mutation rate and differences in functional annotations
Mutation rates were determined within targets with a mean coverage >10× in each proband and respective parents jointly. Confidence Intervals for the mutation rates were determined using a two-sided binomial exact test. These rates were compared with prior estimates described and tabulated in the Supplementary Note using binomial exact tests. We compared the distribution of missense, nonsense and silent mutations with data ascertained in recent studies further described in the Supplementary Materials (Supplementary Table S8). Binomial exact tests were used to determine the significance of the missense to silent and nonsense to missense ratios.

Analysis of Haploinsufficiency and Residual Variation Intolerance Score (RVIS)
Functional analyses were limited to DNMs defined by the following categories (1) broadly damaging: de novo missense variants considered by one or more prediction algorithms to possibly alter or damage gene function; (2) nonsense mutations; (3) likely gene disruptive (LGD) mutations such as nonsense, frameshifts and canonical splice sites; (4) LGD+broadly damaging. For comparison, we also assessed silent variants. Gene-based probability of exhibiting haploinsufficiency and RVIS scores based on ESP6500 (All_0.1%) were obtained from Huang et al.[44] from Petrovski et al.,[45] respectively. The distribution of haploinsufficiency probabilities and ranked RVIS scores for genes in the defined mutational categories were compared with the remainder of the genome using a two-sided Wilcoxon rank sum test. To rule out a potential bias in our observation owing to gene size, 10 000 permutations were performed selecting genes randomly controlling for gene length and GC content.

Disease gene ontology

We used 26 762 human-phenotype associations across 6788 human genes in 2178 phenotype categories from Neurocarta[46] to access the enrichment of disease ontologies in our list of genes. We treated the functional annotation of genes as being drawn from a pool of 19 099 genes in UCSC GoldenPath 'known genes' table. We used the hypergeometric distribution to calculate the significance of overlaps between Neurocarta phenotypes and DNM gene sets. The ontology data were filtered to disorders with 10 or more reliable gene–phenotype relationships (>0.4) in order to flatten representation biases, leaving 72 of the original 2178 phenotypes. Bootstrap permutations (10 000) were performed sampling genes with probabilities determined by coding length and GC content. Reported P-values were multiple-test corrected using the BH correction.

Enrichment of DNMs in chromatin modifier genes

Pfam domain information for all genes in the genome was obtained from the UCSC genome browser and matched to logs odds ratio (LOR) for 3469 Pfam protein domains obtained from Pu et al.[26] Pfam domains with no LOR were assigned the value min(LOR)-1. A list of chromatin modifiers implicated in mental disorders was obtained from Ronan et al.[28] We used the hypergeometric test to calculate the significance of overlaps between chromatin modifers (LOR>5 and the disease list) and DNM sets. Bootstrap permutations (10 000) were done using the same parameters as in the Gene Ontology analysis.

Assessing additional neurodevelopmental and healthy trio exome data sets

Our cross-disorder analysis included data from published trio-based exome studies representing two schizophrenia cohorts (Xu et al.[47] and Gulsuner et al.[48]); four autism data sets (O'Roak et al.,[20] Neale et al.,[19] Iossifov et al.,[18] Sanders et al.[21]) and two ID cohorts (Rauch et al.[23] and de Ligt et al.[22]). For the purpose of this analysis, Afrikaner and US cohorts reported by Xu et al.[47] in their schizophrenia study were analyzed separately. De novo SNV calls in these data were annotated and filtered using the same pipeline applied to our data. Additional analytical details are provided in the Supplementary Materials.

URLS

RESULTS

De novo variant discovery in schizophrenia trios

We performed whole-exome sequencing on 57 complete parent–parent–offspring trios with schizophrenia or a related psychiatric condition, composing 42 'sporadic' trios and 15 'familial trios' defined by the absence or presence of psychosis in 1° or 2° relatives, respectively (Supplementary Information; Supplementary Table S1). On average, 94.2M properly paired reads mapped to the human exome reference (target size ~36 MB) for sporadic trios providing a mean coverage of 67× with over 90% of the exome covered at 10× or greater (Supplementary Information; Supplementary Table S2). The number of mapped reads and mean coverage was higher for familial trios, however with little gain in the breadth of coverage at 10× (Supplementary Information; Supplementary Table S2).

Proband calls were filtered for coverage (>10×), for parental variants and for presence in the Exome Variant Server 6500 and 1000 Genomes (URL). Fifty-nine exonic de novo variants validated by Sanger Sequencing including 58 de novo SNVs, one de novo dinucleotide variant (dnDNV) (Supplementary Table S3). The combined effect of both adjacent dnDNV nucleotide substitutions

introduced a stop codon in SEC31A and therefore the dnDNV was considered a nonsense variant in down stream analysis.

Of the 59 exonic de novo SNVs, 47 and 12 were present in sporadic and familial trios, respectively. In sporadic trios, 28/47, 5/47 and 14/47 DNMs were classified as missense, nonsense and silent mutations, respectively. In familial trios the, 10/12 and 2/10 DNMs were classified as missense and silent, respectively. The number of de novo SNVs per sporadic trio was higher (1.12) than for familial trios (0.8), however, the difference was not consistently significant (Supplementary Information). In both cohorts the distribution of DNMs was consistent with an expected Poisson distribution (Supplementary Information; Supplementary Table S4). The overall mutation rates observed in sporadic (1.62×10^{-8}) and familial trios (1.16×10^{-8}) were within the range of rates observed in previous studies (Supplementary Information; Supplementary Table S6). We did not see a correlation between paternal age and the number of DNMs per trio in this relatively small data set. Likely de novo INDELs were filtered similarly to Iossifov et al. (Supplementary Information). Six de novo INDELs were detected and validated by Sanger sequencing five of which were predicted to generate amino-acid frameshifts and present in the sporadic trios

Distribution of exonic DNMs in schizophrenia trios

In contrast to the distribution of exonic DNMs in healthy trios previously used in exome sequencing studies of autism (Supplementary Information; Supplementary Table S7), there was no significant difference in the proportion of missense DNM in the sporadic trios; however, the proportion of nonsense DNMs was increased by ~3.5-fold (0.101 vs 0.031, empirical $P = 0.01$, Benjamini–Hochberg (BH)-corrected $P = 0.044$, Supplementary Information; Supplementary Table S8). Furthermore, the ratio of nonsense to missense DNMs in sporadic trios was also significantly greater than expected ($P = 0.01$) (Supplementary Table S9). In familial trios, although there was proportionally more missense than silent variants (ratio 5:1), the difference was not significant. Nonsense mutations were not identified in the familial schizophrenia trios.

Haploinsufficiency and intolerance analysis of genes with exonic DNMs

To prioritize likely candidate mutations based on functional impact, the distributions of haploinsufficiency and RVIS[45] were analyzed in five mutation groups defined by (1) broadly damaging missense: DNMs that are potentially damaging by one or more prediction algorithms; (2) nonsense; (3) LGD: nonsense, frameshifts and splice sites; (4) LGD and broadly damaging missense and (5) silent mutations.

Relative to genome-wide predications, genes with nonsense DNM in sporadic trios had significantly higher probabilities of haploinsufficiency (Supplementary Figure 2, $P = 0.0029$, BH-corrected $P = 0.015$). This remained significant after simulations controlling for gene size and GC content ($P = 0.0012$, BH-corrected $P = 0.006$). All but one gene ranked in the top 15% of probable haploinsufficient genes (Table 1). The genes in other mutational groups did not show significantly higher probabilities of haploinsufficiency.

Similarly, RVIS scores of genes with nonsense DNMs ranked significantly higher relative to genome-wide predictions ($P = 0.0013$, BH-corrected $P = 0.0022$). This effect was even more evident for genes with broadly damaging missense DNMs ($P = 2 \times 10^{-4}$, BH-corrected $P = 5 \times 10^{-4}$). These results remained significant in simulations controlling for gene size and GC content (nonsense: $P = 0.0027$, BH-corrected $P = 0.0067$; broadly damaging: $P = 0.0081$, BH-corrected $P = 0.014$). Collectively, the RVIS scores for genes with LGD and broadly damaging mutations ranked significantly higher relative to the remainder of the genome

Table 1. *De novo* mutations in the top 15% ranked RVIS genes

Chr[a]	Pos	Function[b]	Gene	AA change	Base change	Trio	FH[c]	Diagnosis[d]	AAO[e]	Sex[f]	Disease[g]	HI[h]	LOR>5[i]	RVIS rank[j]
12	49420670	nonSNV	MLL2	p.R5027X	C>T	51	F	SCZ	21	M	KS	0.647|9.9%	+	0.06
1	12316498	misSNV	VPS13D	p.R260W	C>T	19	F	SCZ	18	F		0.231|38.1%	−	0.14
11	68204419	misSNV	LRP5	p.C1355R	T>C	53	F	SCZ	17	F	EV	0.066|88.0%	−	0.24
X	53654776	misSNV	HUWE1	p.V433I	G>A	4	F	SCZ	31	F	MR	0.816|5.4%	−	0.4
14	21860919	nonSNV	CHD8	p.S2173X	C>A	37	F	SCZ	18	M	ASD	0.637|10.3%	+	1.18
7	69364416	nonSNV	AUTS2	p.R152X	C>T	58	F	SCZ	15	F	ASD, EP	0.741|7.3%	−	1.82
8	140922446	misSNV	TRAPPC9	p.R970Q	G>A	37	F	SCZ	18	M	ID,MR	0.143|56.4%	−	2.77
16	14340403	misSNV	MKL2	p.R429H	G>A	32	F	SCZ	19	M	ASD	0.316|27.9%	−	3.27
11	3697753	misSNV	NUP98	p.R1650P	G>C	16	F	SCZ	14	M	AML	0.932|2.6%	−	5.01
11	67267698	misSNV	PITPNM1	p.R279W	C>T	53	F	SCZ	17	F	SCZ	0.106|70.4%	−	8.2
4	83770051,2	nonSNV	SEC31A	p.P764X	CC>AT	15	F	SCZ	24	M		0.328|26.7%	−	8.63
11	62414716	misSNV	INTS5	p.V946F	G>T	24	F	SCZ	15	M		0.274|32.4%	−	8.95
10	48390369	misSNV	RBP3	p.R170P	G>C	12	F	SA	19	M		0.340|25.8%	−	9.02
X	153296711	misSNV	MECP2	p.R202C	C>T	17	F	SCZ	24	F	ASD,RS	0.359|24.3%	+	10.46
2	43958706	misSNV	PLEKHH2	p.E970K	G>A	39	F	SCZ	18	F		0.104|71.0%	−	11.68
10	111642192	misSNV	XPNPEP1	p.K347Q	A>C	49	F	SCZ	19	M		0.155|52.8%	−	14.02

[a]Chromosome. [b]Exonic functions. misSNV, missense SNV; nonSNV, nonsense SNV. The *de novo* dinucleotide in *SEC31A* creates one stopcodon; however, ANNOVAR annotations for both individual positions are provided in Supplementary Table S3. [c]FH: Family history. F: Sporadic; T: Familial. [d]Diagnosis. SA, Schizoaffective; SCZ, Schizophrenia. [e]AAO, age at onset. [f]Sex: F, female; M, male. [g]Associated diseases. AML, acute myeloid leukemia; ASD, austism spectrum disorder; EP, epilepsy; EV, exudative vitreoretinopathy; ID, intellectual disability; KS, Kabuki syndrome; MR, mental retardation; RS, Rett syndrome; SCZ, schizophrenia; SNV, single nucleotide variant. [h]HI: probability of haploinsufficiency and percentile rank from Huang *et al*.[44] [i]Genes with Pfam domains with LOR>5 that they are present or absent in chromatin modifiers.[26] [j]Residual variation intolerance ranks based on ESP6500 (ALL_0.1%).[45]

before and after controlling for gene size and GC content ($P = 2.01 \times 10^{-5}$, GC-size-BH-corrected $= 2.1 \times 10^{-3}$). Genes with silent DNMs did not show any difference relative to the genome. Genes ranked in the top 15% RVIS intolerant scores are shown in Table 1.

Notably, among genes ranked in the top 15% of RVIS scores, nonsense DNMs in sporadic trios were identified in chromodomain helicase DNA-binding protein 8 (*CHD8*), autism susceptibility locus 2 (*AUTS2*), histone lysine methyltransferase 2 gene (*MLL2*). Prior genetic evidence suggest that *CHD8*, *AUTS2* and *MLL2* may have an important role in the risk and pathogenesis of neurodevelopmental disorders. Broadly damaging DNMs discovered in other genes implicated in neurodevelopmental disorders such methyl DNA-binding protein, *MECP2*, E3 ubiquitin-protein ligase *HUWE1* and trafficking protein particle complex 9 (*TRAPPC9*) may also be of etiological relevance in this cohort. Recurrent broadly damaging missense DNMs in *PITPNM1* observed in this study and another schizophrenia cohort[47] suggest this gene may also be of importance.

Enrichment analysis of DNMs in austism spectrum disorder/ID-implicated genes

In a hypothesis-free over-representation analysis using high-quality disease ontology annotations (≥10 genes, >0.4 quality score and yielding 72 phenotypes) from Neurocarta,[25] the ontologies, 'autism spectrum disorders', 'autistic disorders' and 'ID' were the most significantly over-represented disorders in all mutational categories assessed except for silent mutations. After correcting for assessment across this broad set of phenotypes, these disorders remained the highest ranked disorders in all mutation categories, with significant evidence for enrichment of LGD+broadly damaging *de novo*'s in 'autistic disorders' (BH-corrected $P = 0.02$). The enrichment of autism and ID among genes with DNMs was significant in repeated simulations controlling for gene size and GC content. This provides support for a specific overlap between schizophrenia and autism at the gene level. Genes factoring in this enrichment were, *CHD8*, *MECP2*, *AUTS2*, *HUWE1* and *TRAPPC9*.

DNMs in chromatin modifiers

The convergent molecular functions of *CHD8, MECP2 and HUWE1* support growing hypotheses that epigenetic regulation of transcription could represent a shared molecular 'risk' mechanism in neurodevelopmental disorders, including autism and ID.[28,49] Indeed, across all mutational categories, except silent mutations, there was a significant over-representation of genes associated with mental disorders involved in chromatin organization, the most significant of which was observed for genes with LGD +broadly damaging *de novos* (BH-corrected $P = 7 \times 10^{-6}$). This was significant in repeated simulations that also controlled for gene size and GC content (BH-corrected >0.0001). Overall, nonsense DNMs were significantly enriched among a set of 419 genes characterized by domains highly specific (LOR>5) to chromatin modification ($P = 0.0046$, BH-corrected $P = 0.023$) (Supplementary Information) This association was largely contributed by *CHD8* and *MLL2*.

Consistency across exome sequencing studies of other neurodevelopmental disorders

To validate this finding, we assessed the robustness of our results using data from nine larger exome sequencing studies of trios with neurodevelopmental disorders including schizophrenia ($n = 3$), autism ($n = 4$), ID ($n = 2$) as well as six 'healthy' siblings/controls (Supplementary Information).

Following consistent annotation and filtering, a significant increase in the proportion of *de novo* nonsense mutations was observed in three of nine neurodevelopmental data sets but only one of six unaffected data sets. (Supplementary Information; Supplementary Table S10). Including the current study, this increase in the proportion of nonsense variation was observed in 40% (4/10) of the neurodevelopmental cohorts. Six of nine additional disease data sets (66%) had significantly more haploinsufficient genes in one of the functional classes with potentially damaging or disrupting mutations, compared with only one control data set (Supplementary Table S11). Similarly, 7/9 additional disease data sets had significantly more RVIS-based intolerant genes with broadly damaging or disrupting mutations

compared with 2/6 control data sets. Combined with the current study this increase was observed in 80% of exome-based neurodevelopment cohorts (Supplementary Table S12).

The only disease ontologies that remained significant in these additional data sets after correction for multiple tests and gene size were 'ID', 'autistic disorder', 'autism spectrum disorder' and 'infantile epilepsy'. An over-representation of functional mutations was specifically observed in 3/9 neurodevelopmental data sets (Supplementary Table S15). Including our data, these ontologies were enriched in 4/10 (40%) disease data sets analyzed, but not in any control data set.

Finally, three out of nine additional neurodevelopmental data sets had an over-representation of LGD mutations in chromatin remodeling genes implicated in mental disorders (Supplementary Table S14). Including our study, this enrichment was observed in 40% (4/10) of the neurodevelopmental data sets analyzed but not in any control data set. Although, the broader but highly specific chromatin modifier gene set was not over-represented among nonsense mutations in any of the additional neurodevelopmental data sets analyzed, they were enriched in other functional mutation classes in 5/9 neurodevelopmental data sets (6/10 including our data) compared with one control data set (Supplementary Table S12).

DISCUSSION

Motivated by the growing interest in identifying ultra rare, potentially highly penetrant, genetic variants underlying the pathogenesis of psychiatric and neurodevelopmental disorders, we describe the exome sequencing of 57 parent–offspring trios with schizophrenia or a related psychotic disorder.

In our analysis of sporadic trios, we observed a higher than expected proportion of nonsense DNMs. We also found that genes with potentially functional mutations ranked significantly less intolerant to rare variation, complementing recently proposed hypotheses that DNMs may be significant risk factors for sporadic schizophrenia.[16,47] We also provide supporting evidence that schizophrenia shares a genetic etiology with autism and ID, and highlight specific genes with roles in chromatin modification proposing a potential molecular disease mechanism shared by these diseases. The analysis of additional exome studies of neurodevelopmental disorders supports these findings.

Although categorical family history information represents a relatively crude measure of genetic liability, in our data set, this distinction was sufficient to identify significant increases in the rate of nonsense mutations in sporadic trios relative to familial and healthy controls reinforcing recently emerging hypotheses that DNMs with a greater likelihood of disrupting gene function could have a significant etiological role in neurodevelopmental disorders.[16,18–24] Although DNM in sporadic cases is insufficient to confirm causality, four of five nonsense DNMs occurred in genes with high probabilities of haploinsufficiency, and three of five genes have been previously implicated in other neurodevelopmental disorders (CHD8, MLL2 and AUTS2) increasing the possibility that they are highly sensitive to inactivating mutations and significant risk factors for schizophrenia. Larger studies will be important in guiding future gene discovery and widening our perspective on the genetic architecture and allelic diversity of the disorder.[50]

Consistent with the growing epidemiological[51,52] and genetic evidence for a shared etiology between neurodevelopmental disorders, we identified an overlap between genes in several trio-based exome studies sets and autism as well as ID. Interestingly, over-representation of genes with potentially functional mutations was restricted to affected trios and not unaffected trios. Although the enrichment of these diseases was not evident across all of the neurodevelopmental disorder data sets analyzed, autism, ID and schizophrenia are unlikely to represent single-disease entities, and

a substantial genetic and etiological heterogeneity is captured by current neurodevelopmental disorder classification. The analysis of much larger cohorts will be required to identify key points of similarity and difference within these patient groups and as well as support for continuing revisions of gene–phenotype ontologies and quality assignments in Neurocarta.

In contrast to genome-wide association study or CNV studies, the granularity of exome data allows the ability to pinpoint, at a higher resolution, potential molecular risk genes and mechanisms tractable to further investigation.

We found a significant over-representation of potentially functional DMNs in genes containing domains necessary for editing, reading, writing of histone post-translational marks and DNA methylation complementing neurobiological findings that epigenetic and retrotransposition regulation have an important role in neurodevelopment.[53–55] Mutations were especially enriched in chromatin-modifying genes already implicated in mental disorders such as CHD8, MECP2 and HUWE1, suggesting the importance of genes that have evolutionarily impacted the epigenetic regulation of brain development and cognitive function in humans[26,56] as having a shared central role in the susceptibility to, pathogenesis and treatment of neurodevelopmental diseases.

Few of the de novo variants discovered in sporadic trios are expected to be highly penetrant mutations[21] (Supplementary Material). Consequently, gene prioritization based on disease and function ontologies (for example, Neurocarta and Gene Ontology), and network inference (for example, protein–protein interaction), is important. However, these approaches are not without limitations. Despite systematic ascertainment, the population of gene–phenotype databases with insufficient literature evidence or weak experimental validity can mislead gene–disease ontologies. Furthermore the ambiguity between mutational effects on gene function and biological processes can bias variant prioritization based on protein–protein interactions and molecular pathways toward genes that are well studied, are of high node degree or are multifunctional. Probabilistic approaches based on recurrence may circumvent biological information but if the likelihood of recurrence is low, this approach could have limited application in diseases with broad genetic heterogeneity.[57,58] Alternatively, developing a network-level understanding of genetic mechanisms by combining complementary genomic data sets that are incognizant of the underlying biology of disease may be critical to prioritize genes and provide novel insights into disease pathogenesis (Supplementary Information).

We have preliminarily used gene expression data from the prefrontal cortex, brain and non-brain to build novel co-expression networks and identified a brain region-specific role for other DNMs, suggesting a new approach to prioritizing future candidate disease-causing variants (Supplementary Information). We found that genes with putatively highly damaging variants were preferentially not co-expressed (low connectivity) with one another in the prefrontal cortex in both the control and schizophrenia conditions. This low connectivity did not extrapolate to the full brain or non-brain network, suggesting that genes such as AUTS2 and NIP7 may have particular, independent functions in networks specifically in the prefrontal cortex. The low connectivity of likely disrupting variants that we observed is consistent with the low node degree of 'nonessential' disease genes,[59] which are likely under less but still purifying selection than essential higher node degree hub genes.[60,61] Furthermore the specificity of this low connectivity in the prefrontal cortex, a region of the brain highly relevant to schizophrenia, is concordant with the confined expression of nonessential disease genes to specialized tissues. This is in contrast to hub genes of high node degree, which show widespread expression in multiple tissues. We suggest tailoring of co-expression networks to control for both brain region and disease state represents a potentially interesting approach to determine molecular mechanisms for further research and

supports the utility of complementing standard methods with agnostic approaches to prioritize novel disease candidates on genome-wide scales.

In conclusion, our results indicate that potentially functional and deleterious DNMs may contribute to the risk of schizophrenia and are consistent with prior exome studies. Genetic and phenotypic diversity represent a challenge for population-based association approaches and may require a broader inclusion of neurodevelopmental phenotypes in assessment of identified risk genes. Our results provide a defined set of genes that support the genetic overlap between schizophrenia and autism, some of which may have a role in chromatin modeling and epigenetic regulation. The identification of these biological functions as potential contributors to the etiology of schizophrenia is, until very recently, very unexpected but is consistent with recent observations in autism genetics (ref. CHD8).[58] A caveat of our study is sample size and it will be necessary to assess these findings in larger cohorts. However, given the genetic and biological heterogeneity of neurodevelopmental disorders, novel findings require increased granularity, which may be shadowed by the reliability of larger cohort studies on formal network analysis that innately rely on limited and biased annotations. As the number of exome studies increase, a more refined set of genes spanning the broad heterogeneity of autism, schizophrenia and other psychiatric disorders will emerge and the incorporation of complementary genomic data may elucidate pathways and mechanisms such as epigenetic regulation that are critical to the development, and ultimately the treatment, of neuropsychiatric conditions.

AUTHOR CONTRIBUTIONS

SEM, WRM and AC designed the study. DWM, EK, CO'B, GD, MG and AC collected all genomic DNA samples for the study. AC made the final diagnoses. MK, RS, EG and EA organized and performed exome capture, sequencing, data handling and validation genotyping. MM and PP provided data from Neurocarta and co-expression meta-analysis data. SEM, JG, JL and SY performed the analysis. SEM, JG, SY, RS, WRM and AC wrote the paper with comments and critique from all co-authors.

REFERENCES

1 Ripke S, Sanders AR, Kendler KS, Levinson DF, Sklar P, Holmans PA et al. Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* 2011; **43**: 969–976.

2 Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kahler AK, Akterin S et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* 2013; **45**: 1150–1159.

3 Sullivan PF. The psychiatric GWAS consortium: big science comes to psychiatry. *Neuron* 2010; **68**: 182–186.

4 Lee SH, DeCandia TR, Ripke S, Yang J, Sullivan PF, Goddard ME et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet* 2012; **44**: 247–250.

5 Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009; **460**: 748–752.

6 Smoller JW, Craddock N, Kendler K, Lee PH, Neale BM, Nurnberger JI et al. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* 2013; **381**: 1371–1379.

7 Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, Steinberg S et al. Large recurrent microdeletions associated with schizophrenia. *Nature* 2008; **455**: 232–236.

8 Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R et al. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* 2008; **358**: 667–675.

9 McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S et al. Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet* 2009; **41**: 1223–1227.

10 Millar JK, Wilson-Annan JC, Anderson S, Christie S, Taylor MS, Semple CA et al. Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Hum Mol Genet* 2000; **9**: 1415–1423.

11 Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos JA, Karayiorgou M. Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet* 2008; **40**: 880–885.

12 Malhotra D, McCarthy S, Michaelson JJ, Vacic V, Burdick KE, Yoon S et al. High frequencies of de novo CNVs in bipolar disorder and schizophrenia. *Neuron* 2011; **72**: 951–963.

13 Grayton HM, Fernandes C, Rujescu D, Collier DA.. Copy number variations in neurodevelopmental disorders. *Prog Neurobiol* 2012; **99**: 81–91.

14 Haldeman-Englert C, Jewett T. 1q21.1 Microdeletion. In: Pagon RA, Adam MP, Bird TD, Dolan CR, Fong CT, Stephens K (eds). *GeneReviews*. University of Washington: Seattle, WA, USA: 1993.

15 Xu B, Roos JL, Dexheimer P, Boone B, Plummer B, Levy S et al. Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat Genet* 2011; **43**: 864–868.

16 Girard SL, Gauthier J, Noreau A, Xiong L, Zhou S, Jouan L et al. Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat Genet* 2011; **43**: 860–863.

17 Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 2012; **488**: 471–475.

18 Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J et al. De novo gene disruptions in children on the autistic spectrum. *Neuron* 2012; **74**: 285–299.

19 Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 2012; **485**: 242–245.

20 O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 2012; **485**: 246–250.

21 Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 2012; **485**: 237–241.

22 de Ligt J, Willemsen MH, van Bon BW, Kleefstra T, Yntema HG, Kroes T et al. Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* 2012; **367**: 1921–1929.

23 Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* 2012; **380**: 1674–1682.

24 Vissers LE, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P et al. A de novo paradigm for mental retardation. *Nat Genet* 2010; **42**: 1109–1112.

25 Portales-Casamar E, Ch'ng C, Lui F, St-Georges N, Zoubarev A, Lai AY et al. Neurocarta: aggregating and sharing disease-gene relations for the neurosciences. *BMC Genomics* 2013; **14**: 129.

26 Pu S, Turinsky AL, Vlasblom J, On T, Xiong X, Emili A et al. Expanding the landscape of chromatin modification (CM)-related functional domains and genes in human. *PLoS One* 2010; **5**: e14122.

27 Turinsky AL, Turner B, Borja RC, Gleeson JA, Heath M, Pu S et al. DAnCER: disease-annotated chromatin epigenetics resource. *Nucleic Acids Res* 2011; **39**, D889–D894.

28 Ronan JL, Wu W, Crabtree GR. From neural development to cognition: unexpected roles for chromatin. *Nat Rev Genet* 2013; **14**: 347–359.

29 American Psychiatric Association Task Force on DSM-IV*Diagnostic and Statistical Manual of Mental Disorders: DSM-IV*, vol. xxvii, 4th edn. American Psychiatric Association: Washington, DC, USA, 1994, 886pp.

30 Strange A, Riley BP, Spenser CA, Morris DW, Pirinen M, O'Dushlaine CT et al. Genome-wide association study implicates HLA-C*01:02 as a risk factor at the MHC locus in schizophrenia. *Biol Psychiatry* 2012; **72**: 620–628.

31 Green MR, Russell DW. *Molecular cloning: a laboratory manual* 4th edn. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, USA, 2012.

32 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**: 1754–1760.

33 Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* 2011; **27**: 1691–1692.

34 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011; **43**: 491–498.

35 Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 2012; **6**: 80–92.

36 Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010; **38**: e164.

37 Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009; **4**: 1073–1081.

38 Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P et al. A method and server for predicting damaging missense mutations. *Nat Methods* 2010; **7**: 248–249.

39 Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome Res* 2009; **19**: 1553–1561.

40 Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010; **7**: 575–576.

41 Siepel A, Pollard KS, Haussler D. New methods for detecting lineage-specific selection. In: *Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB 2006)*. Springer-Verlag: Berlin, Heidelberg, 2006, pp 190–205.

42 Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 2010; **6**: e1001025.

43 Glaubitz JC, Rhodes OE, Dewoody JA. Prospects for inferring pairwise relationships with single nucleotide polymorphisms. *Mol Ecol* 2003; **12**: 1039–1047.

44 Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet* 2010; **6**: e1001154.

45 Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* 2013; **9**: e1003709.

46 Zoubarev A, Hamer KM, Keshav KD, McCarthy EL, Santos JR, Van Rossum T et al. Gemma: A resource for the re-use, sharing and meta-analysis of expression profiling data. *Bioinformatics* 2012; **28**: 2272–2273.

47 Xu B, Ionita-Laza I, Roos JL, Boone B, Woodrick S, Sun Y et al. De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat Genet* 2012; **44**: 1365–1369.

48 Gulsuner S, Walsh T, Watts AC, Lee MK, Thornton AM, Casadei S et al. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* 2013; **154**: 518–529.

49 Houston I, Peter CJ, Mitchell A, Straubhaar J, Rogaev E, Akbarian S. Epigenetics in the human brain. *Neuropsychopharmacology* 2013; **38**: 183–197.

50 Walsh CA, Engle EC. Allelic diversity in human developmental neurogenetics: insights into biology and disease. *Neuron* 2010; **68**: 245–253.

51 Morgan VA, Croft ML, Valuri GM, Zubrick SR, Bower C, McNeil TF et al. Intellectual disability and other neuropsychiatric outcomes in high-risk children of mothers with schizophrenia, bipolar disorder and unipolar major depression. *Br J Psychiatry* 2012; **200**: 282–289.

52 Sullivan PF, Magnusson C, Reichenberg A, Boman M, Dalman C, Davidson M et al. Family history of schizophrenia and bipolar disorder as risk factors for autism. *Arch Gen Psychiatry* 2012; **69**: 1099–1103.

53 Muotri AR, Marchetto MC, Coufal NG, Oefner R, Yeo G, Nakashima K et al. L1 retrotransposition in neurons is modulated by MeCP2. *Nature* 2010; **468**: 443–446.

54 Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT et al. L1 retrotransposition in human neural progenitor cells. *Nature* 2009; **460**: 1127–1131.

55 Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F et al. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* 2011; **479**: 534–537.

56 Vogel C, Chothia C. Protein family expansions and biological complexity. *PLoS Comput Biol* 2006; **2**: e48.

57 Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* 2012; **151**: 1431–1442.

58 O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG et al. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* 2012; **338**: 1619–1622.

59 Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. *Proc Natl Acad Sci USA* 2007; **104**: 8685–8690.

60 Khurana E, Fu Y, Chen J, Gerstein M. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol* 2013; **9**: e1002886.

61 MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 2012; **335**: 823–828.