A scan-statistic based analysis of exome sequencing data identifies *FAN1* at 15q13.3 as a susceptibility gene for schizophrenia and autism

Iuliana Ionita-Laza[*], Bin Xu[†], Vlad Makarov[*] , Joseph D. Buxbaum[‡], J. Louw Roos[§], Joseph A. Gogos[¶], and Maria Karayiorgou[†]

[*]Department of Biostatistics, Columbia University, New York, NY USA,[†]Department of Psychiatry, Columbia University, New York, New York, USA,[‡]Department of Psychiatry, Mount Sinai School of Medicine, New York, NY USA,[§]Weskoppies Hospital, Pretoria RSA, and [¶]Departments of Neuroscience, Physiology & Cellular Biophysics, Columbia University, New York, New York, USA

We used a family-based cluster-detection approach designed to localize significant rare disease-risk variants clusters within a region of interest to systematically search for schizophrenia (SCZ) susceptibility genes within 49 genomic loci previously implicated by de novo copy number variants (CNVs). Using two independent whole exome sequencing (WES) family datasets and a follow-up autism spectrum disorder (ASD) case/control WES dataset, we identified variants in one gene, *FAN1*, as being associated with both SCZ and ASD. *FAN1* is located in a region on chromosome 15q13.3 implicated by a recurrent CNV predisposing to an array of psychiatric and neurodevelopmental phenotypes. In both SCZ and ASD datasets, rare nonsynonymous risk variants cluster significantly in affected individuals within a 20kb window that spans several key functional domains of the gene. Our finding suggests that *FAN1* is a key driver in the 15q13.3 locus for the associated psychiatric and neurodevelopmental phenotypes. *FAN1* encodes a DNA repair enzyme, thus implicating for the first time abnormalities in DNA repair in the susceptibility to SCZ or ASD.

Whole exome sequencing | Copy number variation | Schizophrenia | Autism

Significance Schizophrenia and autism are severe, lifelong brain disorders with complex etiology and high prevalence. A strong link has been established between both disorders and de novo copy number variants but the culprit genes remain unknown. This study uses whole exome sequencing data and a new statistical method based on detecting clusters of rare disease associated variants to identify the responsible gene(s) within genomic regions affected by de novo copy number variants. We discovered a new gene on chromosome 15q13.3, *FAN1*, which contains rare risk variants for both schizophrenia and autism. *FAN1* encodes a DNA repair enzyme, thus implicating for the first time abnormalities in DNA repair in the genetic component and as potential drug targets in psychiatric and neurodevelopmental disorders.

Section Introduction

SCZ is a severe psychiatric disorder characterized by positive, negative and cognitive symptoms and is associated with increased mortality and severely reduced fecundity. It is associated with high heritability and it is now widely accepted that multiple rare de novo and inherited genetic variants contribute to the genetic component of the disease, which is characterized by high locus and allelic heterogeneity[1, 2]. Genetic studies designed to elucidate the forces that shape the genetic risk for SCZ and facilitate identification of variants in specific genes as risk factors for the disorder may help to elucidate the underlying pathophysiology and the identification of novel treatment targets[3]. Genomes are under mutational pressure, thus constantly giving rise to many disease-predisposing variants that are under strong negative selection when leading to diseases with increased mortality and low fecundity and therefore remain rare[4].

Advances in next-generation sequencing technologies make possible to comprehensively explore the contribution of rare variants, both point and structural, on the risk of developing complex psychiatric and neurodevelopmental disorders. A comprehensive investigation of the role that rare variants play both in patients as well as in animal and cellular models can advance our understanding of psychiatric and neurodevelopmental disease[3]. Focusing on de novo variation in appropriate subsets of well-characterized cases can ameliorate the confounding influences of the large number of neutral transmitted variation[1, 2]. However, taking advantage of transmitted variation is also critical in comprehensive gene discovery efforts and a number of analytical methods and strategies have been proposed toward this end[5, 6]. We have previously described a method based on scan statistics for case-control designs that is specifically designed to identify small regions within larger genomic loci enriched with rare disease-risk variants[7]. Unlike conventional association tests that test for association with variants within a specific region, such as a gene, scan statistics approaches test for both association and clustering of variants in a small window of a larger region; in other words, the scan statistic approaches are only powerful if the disease risk variants cluster significantly in a small window, and tend to lose power as the clustering becomes weaker (see results in SI Appendix). Large CNVs, that can extend several megabases in length and which have been shown to influence disease risk, represent a natural application setting for the

scan statistic approaches, since disease risk variants, by definition, reside within the underlying disease gene(s), and therefore cluster in a small region of the large CNV. In order to identify risk genes in genomic loci previously implicated by de novo CNVs we adapted this cluster-detection method to family-based designs (Materials and Methods). We applied such methods to two independent WES SCZ family datasets and one WES ASD case/control dataset. The SCZ family samples analyzed here comprise of trios collected from two European descent populations, the Afrikaner population from South Africa (mostly Dutch descent) and the U.S. population (Northern European descent) and have been sequenced and processed under the same conditions, in the same lab. We analyzed these datasets individually but we also combined them, to allow for an increase in power because a priori, there is no reason to believe that variants in a given CNV cannot play a role in both study populations. The ASD case/control dataset consists of 860 unrelated individuals of European descent (488 ASD subjects and 372 controls), sequenced using an Illumina platform at the Broad Institute.

A higher incidence of de novo CNVs in individuals with SCZ compared with controls[8, 9, 10] highlights a significant contribution of rare de novo copy number mutations to risk of SCZ. Most of the identified de novo CNVs span multiple genes and highlight regions in the human genome likely containing SCZ susceptibility genes. Therefore, focusing on identifying the underlying SCZ genes in these CNV regions is a powerful strategy because of the high prior probability that such genes exist in the regions. Under the assumption that risk genes may be disrupted in variable ways (either by de novo or inherited variation), previous efforts to identify SCZ-susceptibility genes within pathogenic CNV loci focused primarily on analysis of common genetic variants[11, 12]. By taking advantage of rare inherited variants from newly generated whole exome sequencing data and a new cluster detection methodology, we have attempted to identify the responsible gene(s) within the 49 genomic regions that have been previously implicated in SCZ by systematic de novo CNV studies.

Section Results

We included in our analysis sequence data consisting of all single nucleotide variants (SNVs) identified within 49 genomic regions previously implicated in SCZ by systematic de novo CNV studies (SI Appendix, Table S1). The sequencing data have been generated by a recent whole exome sequencing scan of two SCZ datasets, one from the European origin homogeneous Afrikaner population in South Africa (SA, $n = 146$) and one from the more genetically heterogeneous US population (US, $n = 85$)[2]. In our cluster detection approach, we used a 20 kb sliding window, as the default window size, and performed analyses using nonsynonymous SNVs with frequency less than 0.01 (i.e. variants that are likely to have functional impact). Among 49 CNV regions screened, one CNV, on chromosome 15q13.3, carried a strong association signal with a p value in the combined SA and US dataset that remains significant after adjusting for multiple testing (CNV level p value, adjusted for multiple scanning windows considered in the approach, $p = 0.0012$, Bonferroni adjusted threshold is $0.05/49 = 0.001$). Several other CNVs had a nominally significant p value (SI Appendix, Table S2). Although it is likely that, at least, for several of the remaining, non-significant CNVs disease risk variants also cluster in small regions within the CNVs, lack of power precludes their detection, and future, larger studies may elucidate the underlying genetic causes in these CNVs.

For the rare nonsynonymous variants within the 15q13.3 CNV region, we determined that a 20 kb window contained within *FAN1* (Fanconi-associated nuclease 1, also named *MTMR15*), had the highest significant score in the US and SA datasets combined (CNV level $p = 0.0012$; unadjusted or window level p value is 0.00014) (Table 1 and Figure 1). The signal comes from both the SA and US datasets, with the highest peak being consistent across the two datasets. As a comparison, we tested for association using conventional gene-based tests, such as Burden and SKAT tests for family designs[6], for both *FAN1* and the highest scoring 20 kb window (Table 2). These analyses showed that the 20 kb window does indeed contain variants that are significantly associated with SCZ (unadjusted window p for Burden= 0.001 when only nonsynonymous variants are considered). Concordant with our simulation results (see SI Appendix) the Burden and SKAT tests resulted in less significant p values than the cluster test (Table 2, and SI Appendix, Table S3), that are no longer significant after multiple testing adjustment (multiple windows/genes within a CNV). Furthermore, results for the other known genes within 15q13.3 were not significant (not even nominally; SI Appendix, Table S4). In summary, there is evidence of significant clustering of rare associated variants in *FAN1* for the combined SA and US dataset, with the evidence coming from both the SA and US datasets.

The 15q13.3 region directly abuts the 15q11.2-q13 Prader-Willi syndrome locus which is known to contain a cluster of imprinted genes[13]. Although the evidence for parent-of-origin effects in the 15q13.3 locus remains scarce and equivocal[14, 15] it is well known that imprinted genes often cluster in megabase-sized chromosomal domains[16]. Therefore, we have also performed parent-of-origin effect analyses for *FAN1*. When restricting to transmissions of nonsynonymous rare variants from mothers in the combined SA and US datasets, the p value for the Burden test becomes 0.0005 (vs. 0.025 in the original analysis), with the corresponding p value for the transmission from the fathers being 0.46 (Table 2). The transmitted to untransmitted ratio for nonsynonymous rare variants in *FAN1* is 8:0 for mothers, and 3:4 for fathers (SI Appendix, Table S5). These results suggest that the signal is primarily driven by transmissions from mothers. Although *FAN1* is not included among the known imprinted human genes (http://igc.otago.ac.nz/home.html) the possibility of tissue- and developmental stage-specific imprinting at this genetic locus remains to be determined[17].

FAN1 encodes a DNA repair nuclease involved in the repair of highly cytotoxic DNA interstrand cross-links, which prevent strand separation and block replication during mitosis. *FAN1* has at least four isoforms (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/av.cgi?db=human&l=FAN1) (SI Appendix, Figure S1). The longest isoform spans about 40 kb and encodes for a protein with four functional domains: a UBZ (Ubiquitin-binding zinc finger)-type ubiquitin-binding domain, a SAP (Scaffold attachment factor-A/B, Acinus and Protein Inhibitor Of Activated STAT motifs)-type DNA binding domain, a protein-protein interaction (TPR) motif, and a putative nuclease domain termed the "VRR_nuc" domain (uniprot database, SI Appendix, Figure S1)[18]. A list of rare variants (MAF less than 0.01) identified by our WES in *FAN1* is shown in SI Appendix, Table S6 and Figure S1. For each variant, the number of transmissions/untransmissions (T/U) of the minor allele from heterozygous parents is also shown (SI Appendix, Table S6), along with the SIFT and GERP scores for each variant. Among these variants, there are 10 predicted nonsynonymous variants (SI Appendix, Tables S6 and S7). All 10 variants occur in different families (SI Appendix,

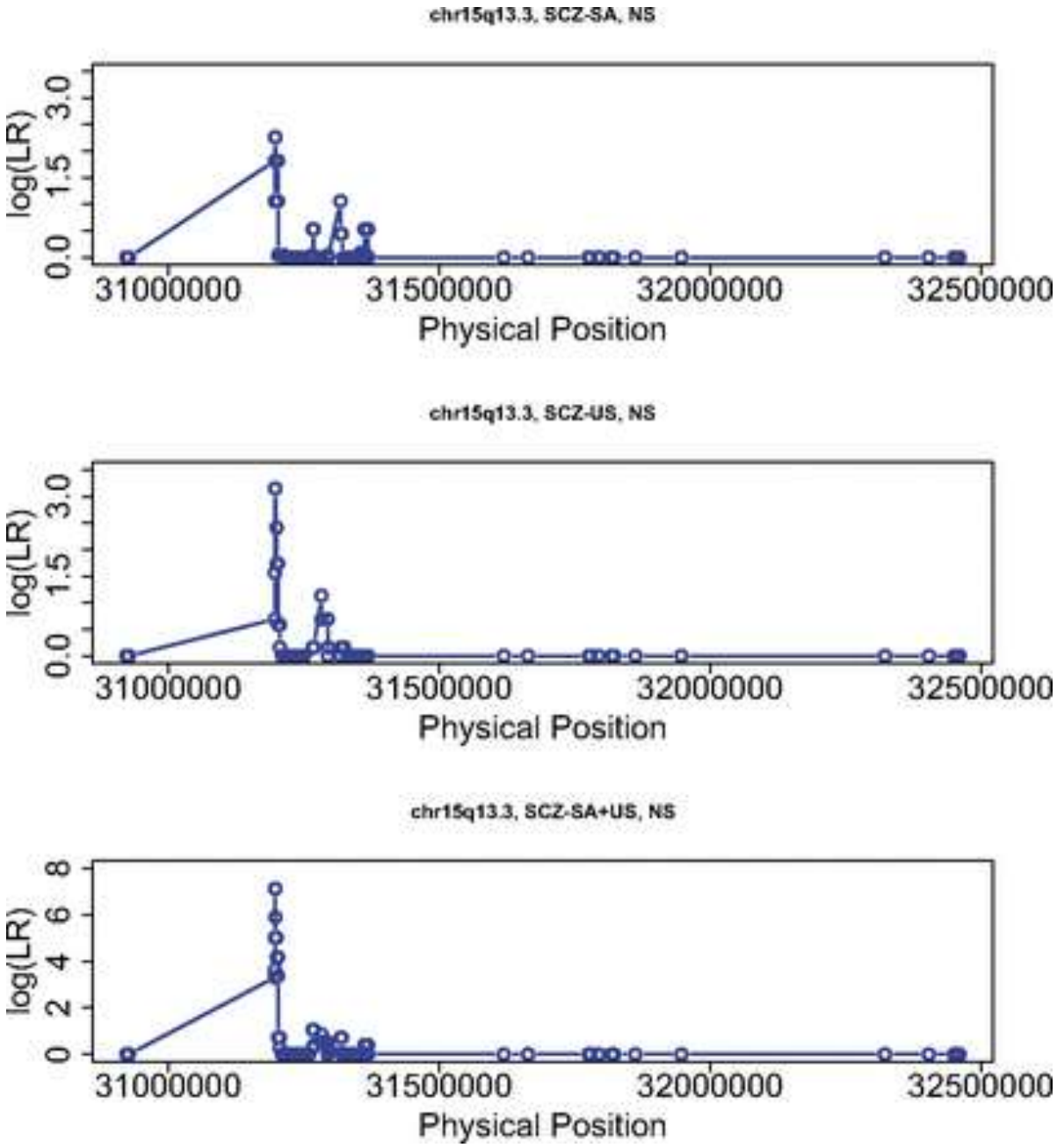

Fig. 1. Application of the scan statistic approach to identify clustering of rare, nonsynonymous (NS) associated variants in the 15q13.3 CNV region in SCZ study. A sliding window of size 20 kb has been used, and for each such window a LR score has been calculated. LR scores for all windows with at least one variant are shown, with window position on the x-axis being the mid-position of the window. Results are shown for the two datasets (SA and US) separately, and combined (SA+US).

Table 1. Chromosome 15q13.3 scan statistic results for the SA, US and SA+US datasets in SCZ study. All rare variants (All) or only the rare nonsynonymous ones (NS) were analyzed. 'Window' corresponds to the 20 kb window with the highest score in the scan statistic procedure, and the gene that contains the window is also reported. The p values in the table are CNV level p values, and are adjusted for multiple scanning windows considered in the approach

Dataset	n	Variants	P	Window (hg19)	Gene
SA	146	All	1.2E-02	31.197.564-31.217.564	$FAN1$
		NS	2.3E-01	31.197.976-31.217.976	$FAN1$
US	85	All	9.4E-02	31.198.043-31.218.043	$FAN1$
		NS	4.8E-02	31.198.043-31.218.043	$FAN1$
SA+US	231	All	7.3E-04	31.197.564-31.217.564	$FAN1$
		NS	1.2E-03	31.197.976-31.217.976	$FAN1$

Table 2. Conventional gene based tests (Burden and SKAT) results for $FAN1$ in SCZ and ASD studies. Parent-of-origin effect (POE) analysis results are also shown, with transmissions from mothers (Mo) and fathers (Fa) being considered separately. Only rare nonsynonymous (NS) variants were included in the analyses. 'Window' corresponds to the 20 kb window with the highest score in the scan statistic procedure. The p values in the table are gene or window level p values, hence they are unadjusted for multiple genes or windows within a CNV region

					POE-Mo		POE-Fa	
Dataset	n	Variants	Burden	SKAT	Burden	SKAT	Burden	SKAT
SCZ	231	NS	0.025	0.14	0.00049	0.02	0.46	0.69
(SA+US)		NS-window	0.001	0.022	0.003	0.023	0.13	0.90
ASD	860	NS	0.022	0.014	-	-	-	-
(Broad - C/C)		NS-window	0.055	0.010	-	-	-	-

Table S7). Six rare variants were predicted to be damaging by the SIFT software (SIFT score < 0.05). Notably one of them (p.R507H), that falls within the 20 kb window, shows a 4/0 T/U ratio. This variant is located within the SAP domain, as is another predicted damaging variant (p.K505I) with a 1/0 T/U ratio. The SAP domain is required for potential DNA-binding.

Overall, there are four patients with the one recurrent damaging variant that falls within the 20 kb window. Notably, three of these patients report co-morbid depression (meeting formal diagnostic criteria for Dysthymia, Major Depressive Disorder (MDD), or Severe depressive episode), while the fourth is diagnosed with schizoaffective disorder of the bipolar type with prominent features of depression accompanying the psychotic episodes ($p = 0.02$ based on a prevalence of co-morbid depression in our combined samples of 17.6%). In fact, of nine carriers of missense variants in the 20 kb window, four of the five patients that are diagnosed with DSM-IV schizophrenia have some form of co-morbid depression ($p = 0.055$), while the remaining four are diagnosed with schizoaffective disorder where a strong component of depression is present (SI Appendix, Table S8) suggesting that mutations in the *FAN1* gene could be associated with a SCZ phenotype characterized by prominent depressive symptoms. Along these lines, while analysis of psychiatric GWAS did not provide any evidence for a significant association with common *FAN1* variants and SCZ, multiple marginally significant associations exist between *FAN1* SNPs, MDD and response to antidepressants (SI Appendix, Figure S2).

Recently, Vacic et al.[19] performed a CNV association study based on three different datasets (8,394 SCZ cases and 7,431 controls). Based on their analyses, a peak of association (i.e. a segment with minimal p value in the CNV region) in the 15q13.3 region is achieved in the interval 31.094.316-31.203.815 that encompasses only *FAN1*. Across the three datasets, there were 18 deletions in 8,394 SCZ cases and one deletion in 7,431 controls, with the p value for this peak region being 0.0001.

A number of CNVs, including 15q13.3, show diagnostic pleiotropy, increasing risk across a number of neurodevelopmental disorders such as ASD, intellectual disability and epilepsy[20, 21]. It is highly probable that the same gene in 15q13.3 confers increased risk to both SCZ and ASD. We therefore tested for association between variants in *FAN1* and ASD using a publicly available WES case/control dataset ($n = 860$; more details on this dataset are given in the Materials and Methods). Specifically, using the scan statistic approach, we find that the highest scoring 20 kb window within the entire 15q13.3 locus is similar to the one in SCZ (31.202.961-31.222.961), and overlaps *FAN1* (Figure 2). The p value for the 15q13.3 CNV in the ASD data is 0.065 (adjusted for scanning the region using 20 kb windows; unadjusted or window level p value is 0.004). Applications of conventional Burden and SKAT tests showed that rare nonsynonymous *FAN1* variants are also associated with ASD (Table 2). Since the ASD dataset comprises unrelated cases and controls, a parent-of-origin effect analysis is not possible in this dataset. A list of variants and their frequencies in cases vs. controls is shown in SI Appendix, Table S9. Interestingly, the same damaging variant that is transmitted four times and untransmitted 0 times to SCZ patients also has increased frequency in ASD cases compared with controls (0.018 vs. 0.008; Barnard's test one-sided $p = 0.039$; SI Appendix, Tables S6 and S9). In fact, the frequency of this variant based on data from 4,600 European Americans in the NHLBI Exome Variant Server (http://evs.gs.washington.edu/EVS/) is 0.0079, very similar to the frequency estimated in the controls from the ASD study (Barnard's test one-sided $p = 0.004$ when comparing frequency in ASD cases vs. ESP controls). Therefore, *FAN1* is likely to be a previously unappreciated risk gene predisposing to both neurodevelopmental and psychiatric disorders.

Because risk genes often act together in pathways or networks, we have looked at additional genes that interact directly or indirectly with *FAN1* and asked whether genes connected with *FAN1* carry rare loss-of-function (LOF) variants more often than would be expected by chance. Using the STRING database (http://www.string-db.org), we have constructed a network of genes that interact with *FAN1*. We started with *FAN1* and expanded the network to include all genes that have been directly or indirectly connected to *FAN1* with a high confidence score greater than 0.9, resulting in a network of 27 genes (Figure 3). Among these genes, we have identified 10 rare LOF variants in SCZ cases from the SA+US dataset (SI Appendix, Table S10), suggesting a significant enrichment of LOF variants in this *FAN1*-based network (permutation-based $p = 0.037$). Identification of recurrent mutations that presumably exert the same deleterious effect at the scale of a pathway provides additional support for a functional contribution of *FAN1* deficiency to the risk of the 15q13.3-associated psychiatric and neurodevelopmental phenotypes. It is interesting to note here that one of the genes in this network, *FANCL*, resides in the same linkage disequilibrium block as a genome-wide significant SNP in a recent GWAS study on schizophrenia[22].

Section Discussion

Copy number variation at the 15q13.3 locus results in disturbed brain development contributing to an increased risk for different neuropsychiatric disorders with variable expressivity, dependent on additional genetic and environmental factors. In particular, heterozygous and homozygous 15q13.3 microdeletions exist in both inherited and de novo forms and predispose to a spectrum of clinical phenotypes, including SCZ[23, 24], ASD, ADHD, intellectual disability and epilepsy[21]. The underlying mechanism leading to these microdeletions is a high density of low copy repeats (LCR), LCR-mediated nonallelic homologous misalignment and unequal recombination resulting in a common 2.0 Mb deletion, which includes deletion of a 1.5 Mb of unique sequence as well as an additional $\sim$ 500 Kb of segmental duplications. 15q13.3 microdeletions disrupt at least seven genes: *ARHGAP11B*, *FAN1/MTMR15*, *MTMR10*, *TRPM1*, *KLF13*, *OTUD7A*, and *CHRNA7*[20]. Among them, the alpha7-nicotinic receptor subunit (*CHRNA7*) has been discussed as a prime candidate gene for at least some of the disturbances such as seizures, which have been observed in individuals with smaller microdeletions comprising only *CHRNA7* and *OTUD7A*[25]. However, no patients with pathogenic point mutations in *CHRNA7* have been reported[20]. The genetic contribution of the other genes within this deleted region remains largely unexplored. Here we report the identification of a cluster of rare nonsynonymous variants located within a 20kb window that spans several key functional domains of *FAN1*, which are associated with SCZ and ASD in two independent datasets. Our findings suggest that *FAN1* is a key susceptibility gene in this region for the psychiatric and neurodevelopmental phenotypes associated with the structural mutations at 15q13.3.

Based on the multitude of diagnoses associated with 15q13.3 microdeletions we envision two possible interpretations of our results. One possibility is that deleterious rare variants in the *FAN1* gene represent primary genetic deficits shared by a number of disorders such as schizophrenia and ASD. Accordingly, given the variable phenotypic spectrum associated with these rare

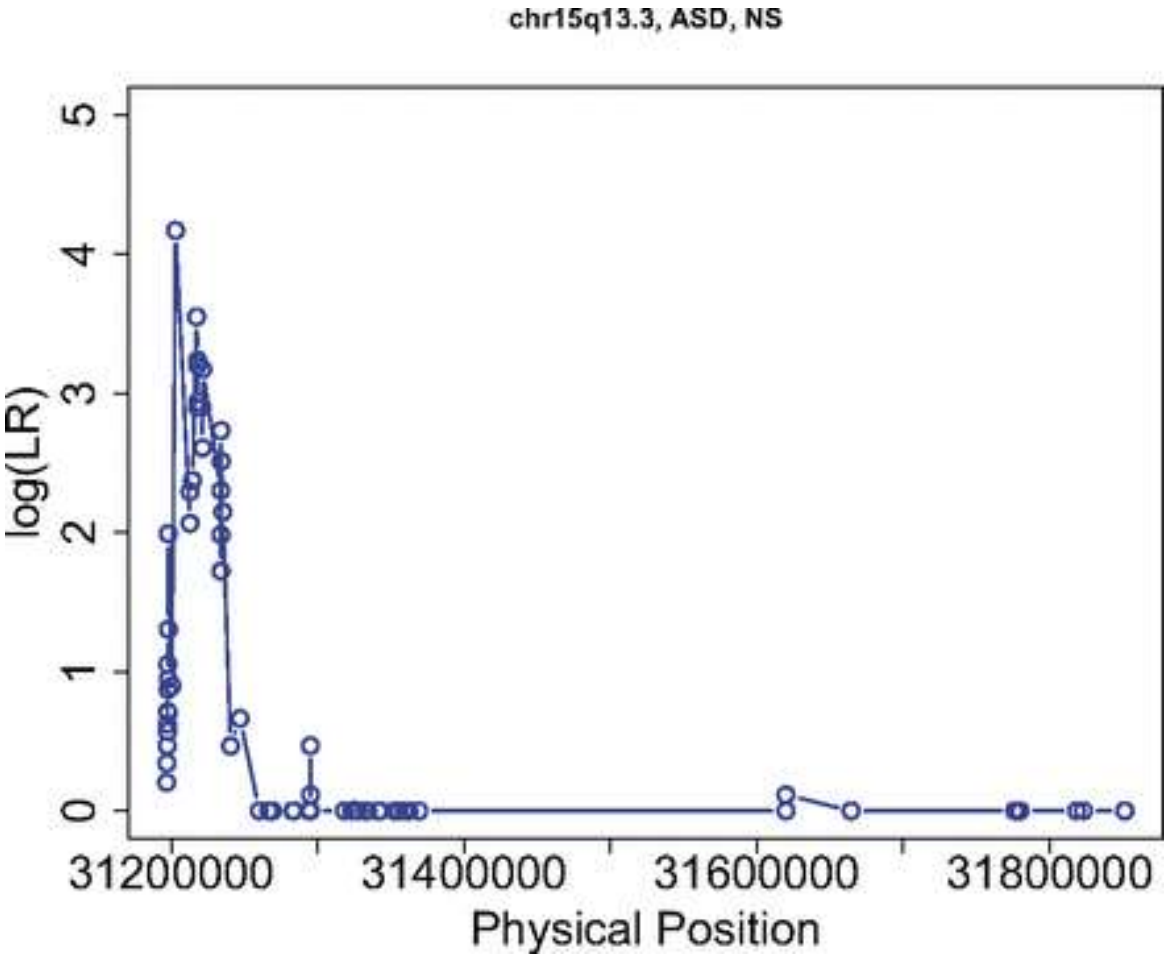

Fig. 2. Application of scan statistic approach to identify clustering of rare, nonsynonymous (NS) associated variants in the 15q13.3 CNV region in the ASD study. A sliding window of size 20 kb has been used, and for each such window a LR score has been calculated. LR scores for all windows with at least one variant are shown, with window position on the x-axis being the mid-position of the window.

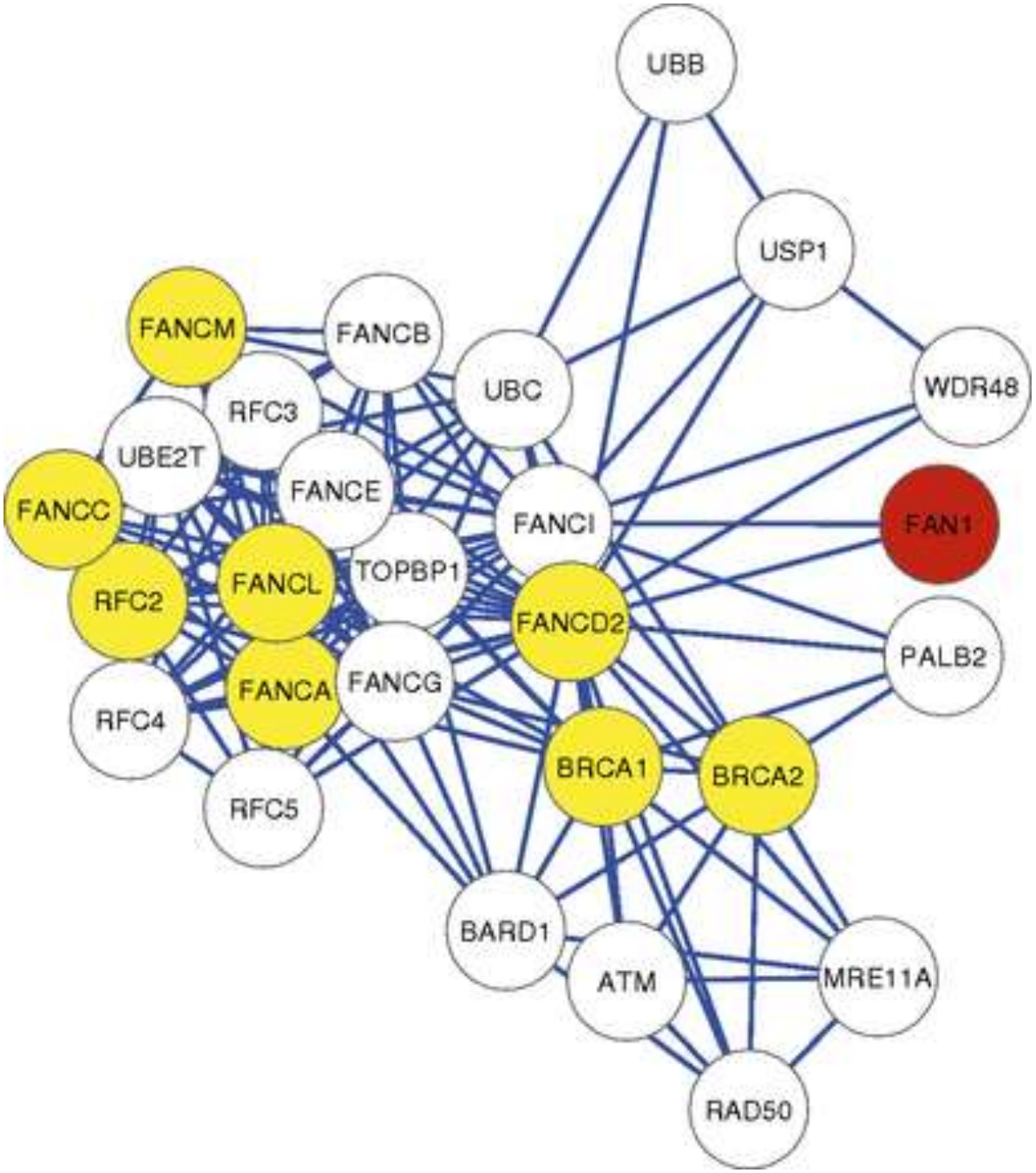

Fig. 3. Network with 27 genes interacting directly or indirectly with FAN1. Genes in yellow are the genes (excluding FAN1) containing at least one rare LOF variant transmitted to SCZ probands (see also SI Appendix, Table S10).

variants, it is likely that the expression of each disorder, and the form of neurodevelopmental phenotype taken, depends on an unknown combination of genetic (rare or common) and environmental factors, and possibly chance. Another possibility is that deleterious rare variants in the *FAN1* gene are promiscuous genetic lesions representing secondary modifier loci that modulate the penetrance and severity of the phenotype associated with other primary, disease-specific, genetic lesions. Additional genetic studies are needed to distinguish between these two possibilities.

As mentioned above, *FAN1* encodes for a recently discovered[18, 26, 27] DNA repair nuclease involved in the repair of highly cytotoxic DNA interstrand crosslinks (ICLs). The FAN1 protein is recruited to sites of ICL damage by interacting with a FANCI-FANCD2 complex through its UBZ domain. However, the exact role of FAN1 in this highly complex process is still poorly understood. A number of reports suggest that FAN1 functions in the DNA repair pathway affected in Fanconi anemia[18], a rare recessive disorder, which encompasses various consequences of DNA damage, including a range of developmental abnormalities. Indeed, *FAN1* mutations might generate broader developmental abnormalities related to DNA damage-induced loss of progenitor cells in various body tissues. Recently, mutations (6 heterozygous and 2 homozygous) in *FAN1* were reported as the cause of karyomegalic interstitial nephritis (KIN) in at least nine families, linking the accumulation of DNA damage with chronic kidney failure[28].

In terms of understanding the biological mechanisms underlying the genetic risk conferred by rare *FAN1* variants it is informative that *Fan1* knockdown in zebrafish induces developmental defects including microcephaly[28]. Interestingly, individuals carrying a homozygous microdeletion spanning *FAN1* show severe neurodevelopmental abnormalities, which also include microcephaly[29]. Micro- and macrocephaly phenotypes have also been described in conjunction with deletions and duplications of another locus (16p11.2) linked to high risk of psychiatric disorders and ASD, and were linked to a single gene at this locus[30]. Although the contribution of *FAN1* to the various neurodevelopmental phenotypes associated with 15q13.3 microdeletions remains to be determined, it is noteworthy that microcephaly is a feature common to a diverse range of DNA repair-defective disorders[31]. Microcephaly is most likely caused by increased cell death or failure of neuronal stem cells or their progenitors to divide, consistent with a fundamental role for the DNA damage response in maintaining proliferative potential and cellular survival in the developing nervous system in the face of exogenous and endogenous DNA damage. Interestingly, a recent gene expression profiling study of prefrontal cortex of postmortem brains from subjects with SCZ, bipolar disorder and depression showed that expression of 818 genes was significantly correlated with a decrease in the number of perineuronal oligodendrocytes and 600 genes were significantly correlated with a decrease in density of calbindin-positive interneurons across all patient samples[32]. *FAN1* was among these affected genes (the only gene from within the 15q13.3 CNV region). Overall, these observations are consistent with a microcephaly phenotype demonstrated in the *Fan1* knockout animal models. Therefore, it is likely that deleterious *FAN1* mutations increase risk for psychiatric and neurodevelopmental disorders by interfering with aspects of early neuronal development.

FAN1 is relatively widely expressed in various brain regions and its expression level appears to be constant along the human brain development as indicated by Human brain transcriptome (SI Appendix, Figure S3). Although the gene is widely expressed and at steady levels throughout development and adulthood, it is well known that genome maintenance deficiencies have a higher impact during neuronal development and on the embryonic brain rather than the adult brain[33]. Finally, cells unavoidably sustain DNA damage from their own metabolism[34, 35, 36, 37] but also from exogenous sources. It has been suggested that exposure to high levels of environmental toxins or chemotherapeutics that damage renal cells may underlie the association between *FAN1* mutations and KIN. The extent that (geno)toxic agents, which may induce DNA damage, influence neurodevelopmental phenotypes in FAN1-deficient individuals remains to be determined.

Materials and Methods

Sequencing Data for Schizophrenia Trios, and ASD Cases and Controls. The schizophrenia samples analyzed here comprise of families (trios) collected from two distinct populations, the Afrikaner population from South Africa (European, mostly Dutch origin) ($n = 146$ schizophrenia families) and the U.S. population (Caucasian, Northern European origin families) ($n = 85$ schizophrenia families). Of the 146 Afrikaner probands, 122 (83.6%) had a diagnosis of schizophrenia and 24 (16.4%) were diagnosed with schizoaffective disorder. Of the 85 U.S. probands, 46 (54.1%) had a diagnosis of schizophrenia, and 39 (45.9%) were diagnosed with schizoaffective disorder. Affected trios were recruited and characterized in the context of ongoing, large-scale genetic studies of schizophrenia and have been described previously[2, 8]. Informed consent was obtained from all participants and the Institutional Review Committees of Columbia University and University of Pretoria approved all procedures. Paternity and maternity were confirmed prior to sequencing via the Affymetrix Genome-Wide Human SNP Array 5.0 as well as via a panel of microsatellite markers. Carriers of large (≥ 30 kb) rare de novo CNVs were excluded based on prior CNV scans of the same datasets[8, 9]. DNA for all study subjects was extracted from whole blood and analysis was performed blind to affected status while maintaining knowledge of the parent-child relations. Exome capture and sequencing was performed as previously described[2]. The analytical pipeline and filters used in the exome sequence analysis have been previously described[2].

The ASD case-control dataset has been sequenced as part of the ARRA Autism Sequencing Collaboration (AASC), and is publicly available through dbGAP (http://www.ncbi.nlm.nih.gov/gap/?term=phs000298). The dataset consists of 488 ASD cases and 372 controls of European ancestry, and whole-exome sequencing was performed at the Broad Institute using an Illumina HiSeq2000 platform. Data was processed with Picard[38] and BWA[39] to map reads to hg19. Variants were called using the Genome Analysis Toolkit[40] and only those variants that passed standard quality control filters were analyzed.

Statistical Analyses. We use both conventional sequence-based association tests (such as Burden and variance component tests, e.g. SKAT), as well as clustering tests based on scan statistics to identify the gene(s) within CNVs that contain disease associated variants. Conventional gene-based tests, such as Burden and SKAT, have been proposed before and a detailed description can be found in [5] (case-control design) and [6] (family design). Briefly, they assess the evidence for association with variants in a particular region (such as a gene) by grouping variants in the region. The underlying test statistic is a weighted-sum of the individual variant score statistics, and statistical significance is assessed using either asymptotic approximations, or Monte Carlo simulations. We also perform parent-of-origin tests, by restricting analyses separately to transmissions from mothers, and fathers. The clustering tests we perform are designed to identify locations within a CNV where disease associated variants cluster. Unlike conventional gene-based tests, which are called *self-contained* tests, the cluster detection methods are *competitive* tests, that compare association signal in a region or gene, with that outside the region or gene. We have previously proposed a scan statistic based method for case-control designs[7]. We have extended it to trio designs, and details are described in the SI Appendix; conceptually it is very similar to the existing scan statistic method for the case-control design.

ACKNOWLEDGMENTS. We thank the families who contributed to the original studies that generated the data analyzed here. We also gratefully acknowledge support by National Science Foundation grant DMS-1100279 and National Institutes of Health grants R01MH095797 (II-L) and R01MH61399 (MK).

1. Girard SL et al. (2011)— Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat Genet.* **43: 860-863.**
2. Xu B et al. (2012) De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat Genet* **44: 1365-1369.**
3. Arguello PA, Gogos JA (2012) Genetic and cognitive windows into circuit mechanisms of psychiatric disease. *Trends Neurosci* **35: 3–13.**
4. Keller MC, Miller G (2006) Resolving the paradox of common, harmful, heritable mental disorders: which evolutionary genetic models work best? *Behav Brain Sci* **29: 385–404.**
5. Lee S, Wu MC, Lin X (2012) Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13: 762-775.**
6. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X (2013a) Family-based association tests for sequence data, and comparisons with population-based association tests. *Eur J Hum Genet* **in press**
7. Ionita-Laza I, Makarov V; ARRA Autism Sequencing Consortium, Buxbaum JD (2012b) Scan-statistic approach identifies clusters of rare disease variants in LRP2, a gene linked and associated with autism spectrum disorders, in three datasets. *Am J Hum Genet* **90: 1002–1013.**
8. Xu B et al. (2008) Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet* **40: 880-885.**
9. Malhotra D et al. (2011) High frequencies of de novo CNVs in bipolar disorder and schizophrenia. *Neuron* **72: 951-963.**
10. Kirov G et al. (2012) De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol Psychiatry* **17: 142-153.**
11. Liu H et al. (2002) Genetic variation in the 22q11 locus and susceptibility to schizophrenia. *Proc Natl Acad Sci USA* **99: 16859-16864.**
12. Steinberg S et al. (2012) Common variant at 16p11.2 conferring risk of psychosis. *Mol Psychiatry*
13. Chamberlain SJ, Lalande M (2010) Neurodevelopmental disorders involving genomic imprinting at human chromosome 15q11-q13. *Neurobiol Dis* **39: 13–20.**
14. Stefan M et al. (2005) Genetic mapping of putative Chrna7 and Luzp2 neuronal transcriptional enhancers due to impact of a transgene-insertion and 6.8 Mb deletion in a mouse model of Prader-Willi and Angelman syndromes. *BMC Genomics* **6: 157.**
15. Ma J et al. (2008) Association study of a (TG)n dinucleotide repeat at chromosome 15q13.3 and schizophrenia in the Chinese population. *Psychiatry Res* **159: 245–249.**
16. Wan LB, Bartolomei MS (2008) Regulation of imprinting in clusters: noncoding RNAs versus insulators. *Adv Genet* **61: 207–223.**
17. Prickett AR, Oakey RJ (2012) A survey of tissue-specific genomic imprinting in mammals. *Mol Genet Genomics* **287: 621–630.**
18. Smogorzewska A et al. (2010) A genetic screen identifies FAN1, a Fanconi anemia-associated nuclease necessary for DNA interstrand crosslink repair. *Mol Cell* **39: 36-47.**
19. Vacic V et al. (2011) Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. *Nature* **471: 499-503.**
20. van Bon BWM, Mefford HC, de Vries BBA (2010) 15q13.3 Microdeletion.
21. Malhotra D, Sebat J (2012) CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* **148: 1223–1241.**
22. Steinberg S et al. (2011) Common variants at VRK2 and TCF4 conferring risk of schizophrenia. *Hum Mol Genet* **20: 4076–4081.**
23. Stefansson H et al. (2008) Large recurrent microdeletions associated with schizophrenia. *Nature* **455: 232-236.**
24. International Schizophrenia Consortium (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455: 237-241.**
25. Shinawi M et al. (2009) A small recurrent deletion within 15q13.3 is associated with a range of neurodevelopmental phenotypes. *Nat Genet* **41: 1269-1271.**
26. Kratz K et al. (2010) Deficiency of FANCD2-associated nuclease KIAA1018/FAN1 sensitizes cells to interstrand crosslinking agents. *Cell* **142: 77-88.**
27. MacKay C et al. (2010) Identification of KIAA1018/FAN1, a DNA repair nuclease recruited to DNA damage by monoubiquitinated FANCD2. *Cell* **142: 65-76.**
28. Zhou W et al. (2012) FAN1 mutations cause karyomegalic interstitial nephritis, linking chronic kidney failure to defective DNA damage repair. *Nat Genet* **44: 910–915.**
29. Trujillo JP et al. (2012) On the role of FAN1 in Fanconi anemia. *Blood* **120: 86–89.**
30. Golzio C et al. (2012) KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. *Nature* **485: 363–367.**
31. O'Driscoll M, Jeggo PA (2008) The role of the DNA damage response pathways in brain development and microcephaly: insight from human disorders. DNA Repair (Amst) **7: 1039-1050.**
32. Kim S, Webster MJ. Correlation analysis between genome-wide expression profiles and cytoarchitectural abnormalities in the prefrontal cortex of psychiatric disorders. *Mol Psychiatry* **15: 326–336.**
33. Barzilai A, Biton S, Shiloh Y (2008) The role of the DNA damage response in neuronal development, organization and maintenance. *DNA Repair* **7: 1010–1027.**
34. Lans H, Hoeijmakers JHJ (2012) Genome stability, progressive kidney failure and aging *Nature Genetics* **44: 836–838.**
35. Regnell CE et al. (2012) Hippocampal adult neurogenesis is maintained by Neil3-dependent repair of oxidative DNA lesions in neural progenitor cells. *Cell Rep* **2: 503–510.**
36. Suberbielle E et al. (2013) Physiologic brain activity causes DNA double-strand breaks in neurons, with exacerbation by amyloid-? *Nat Neurosci* **doi: 10.1038/nn.3356.**
37. Sykora P et al. (2013) Modulation of DNA base excision repair during neuronal differentiation. *Neurobiol Aging* **doi:pii: S0197-4580(12)00647-1.**
38. DePristo MA et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43: 491–498.**
39. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26: 589–595.**
40. McKenna A et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20: 1297–1303.**