Transcriptome outlier analysis implicates schizophrenia susceptibility genes and enriches putatively functional rare genetic variants

Jubao Duan1,3,*, Alan R. Sanders1,3, Winton Moy1,3, Eugene I. Drigalenko4, Eric C. Brown2, Jessica Freda1, Catherine Leites1, Harald H. H. Göring4, MGS5,† and Pablo V. Gejman1,3

1Center for Psychiatric Genetics and 2Center for Biomedical Research Informatics, NorthShore University HealthSystem, Evanston, IL, USA, 3Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, IL, USA, 4Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, USA and 5Molecular Genetics of Schizophrenia (MGS) Collaboration

*To whom correspondence should be addressed at: 1001 University Place, NorthShore University HealthSystem, Evanston, IL 60076, USA. Email: jduan@uchicago.edu or jduan@northshore.org

Abstract

We searched a gene expression dataset comprised of 634 schizophrenia (SZ) cases and 713 controls for expression outliers (i.e., extreme tails of the distribution of transcript expression values) with SZ cases overrepresented compared with controls. These outlier genes were enriched for brain expression and for genes known to be associated with neurodevelopmental disorders. SZ cases showed higher outlier burden (i.e., total outlier events per subject) than controls for genes within copy number variants (CNVs) associated with SZ or neurodevelopmental disorders. Outlier genes were enriched for CNVs and for rare putative regulatory variants, but this only explained a small proportion of the outlier subjects, highlighting the underlying presence of additional genetic and potentially, epigenetic mechanisms.

Introduction

Schizophrenia (SZ; OMIM #181500) is a severe brain disorder with ~0.5–1% prevalence and ~80% heritability estimated from twin studies (1). Recent SZ genome-wide association studies (GWAS) have implicated >100 risk loci with common associated variants (2–8) each contributing very small effects. Several rare (frequency <1%) and large (>100 kb) genomic deletions or duplications (copy number variants, CNVs), such as 1q21.1del, 2p16.3del (NRXN1), 15q13.2del, 16p11.2dup and 22q11.21del have also been found associated, each with larger effect sizes (9–11). A polygenic model, i.e., many common variants each contributing a very small effect on risk (6,12), has been proposed as an explanation of a large portion of the remaining genetic liability (6). Yet, the biological mechanisms underlying most of these genetic findings remain at best only vaguely delineated. Coding variants seem to explain little of the disease associations (7,13,14). On the other hand, GWAS-implicated disease risk variants tend to be in non-coding sequences (7). We hypothesize that the study of gene expression profiles may contribute to the identification of functional regulatory variants involved in the etiology and pathophysiology of SZ.
Transcriptional studies generally examine gene expression averages, which can miss infrequent but large differences in gene expression that may be present only in a small proportion of the sample. On the other hand, expression quantitative trait locus (eQTL) analyses have low power to identify rare regulatory variants (15). We report here the utilization of an approach to identify expression ‘outliers’ enriched for SZ cases, i.e., gene-specific expression outliers located in the expression distribution tails beyond two standard deviations (SD) from the mean (16) with a bootstrapping (17) P > 0.95. These analyses detect a subset of genes that are more likely to harbor regulatory variants associated with SZ (Fig. 1). We hypothesize that DNA variants with low minor allele frequency (MAF) that makes them undetectable in GWAS in the coding or regulatory sequences are responsible for pathogenic transcriptional dysregulation, and therefore these variants should be found enriched in SZ cases. Specifically, we hypothesize that the study of the set of genes presenting case-enriched aberrant abundances associated with genetic variants may provide biological insight into disease mechanisms.

We have analyzed RNAseq data from lymphoblastoid cell lines (LCLs) of 634 SZ cases and 713 controls (none of these subjects carry an aforementioned CNV known to be associated with SZ) (9–11). We report here that the outlier genes with expression distribution tails enriched for SZ cases are also enriched with brain expression and for being located within CNVs associated with neurodevelopmental disorders, including SZ. SZ cases overall had a higher outlier burden than controls for genes within these same CNVs. In a limited follow up, we have analyzed by DNA sequencing a subset of case-enriched outlier genes with plausible involvement in SZ, examining their coding and also putative non-coding regulatory regions, and found enrichment of rare regulatory variants in outlier subjects.

Results

Expression outlier analysis reliably detected transcriptional effects of known CNVs

As shown in Figure 1, we have carried out expression outlier analysis in an RNAseq data set of 634 SZ cases and 713 controls. After a series of quality control procedures (see method), we identified 8355 autosomal genes expressed at a level of RPKM (reads per kilobase of transcript per million reads mapped) ≥1 in 100% of LCL samples. Subsequently, separately for each gene, we compared a given gene’s observed expression level for every sample to the gene’s mean expression level across all samples, identifying as ‘outliers’ those samples where the gene’s expression level was at least 2 (or 3) SD higher than average (which we refer to as an upper tail expression outlier) or lower than average (a lower-tail expression outlier). All the expressed genes had ≥1 expression outlier/s at the 2SD cut-off. The average number of outlier subjects per gene was 60.8 (1–105; Fig. 2A), or ~4.5% of the samples, which matches the expectation under the assumption that each gene’s expression level is normally distributed among individuals. At the 3SD cut-off, ~99.8% of expressed genes had ≥1 outlier subject/s, with an average of 7.5 per gene (1–35; Fig. 2B), or ~0.56% of the samples, which is elevated compared with the expectation of 0.27% under a normal distribution. We found no correlation between the number of outlier subjects per gene and its log2RPKM (Fig. 2C and D), indicating that our outlier detection had no bias towards lowly expressed genes.

We first examined the ability of our outlier approach to detect large transcriptional differences in LCLs due to DNA dosage differences by examining whether known CNV-carriers (11) fell within the expression outlier tail of the deleted or duplicated genes (Supplementary Material, Text). Consistent with their gene dosage, 96% of genes entirely in a CNV deletion (i.e., hemizygous) were detected as 2SD lower-tail expression outliers (Supplementary Material, Table S1). For genes in a duplication, 81% were detected as upper expression tails (Supplementary Material, Table S2). For known CNVs only partially overlapping a particular gene involving 1 or 2 deleted exons, which often leads to decay of the truncated transcripts (18–20), we detected 64% as outliers in the 2SD lower expression tails (Table 1), showing a significant enrichment of expression outliers in such small deletion carriers (empirical P = 2.0 × 10−5 from 105 permutations), which suggests that small exon deletions have detectable transcriptional effects (i.e., decreased expression) in our analyses. However, for CNVs involving 1 or 2 duplicated exons of a gene, we only identified ~10% (2 out of 21 subjects, each with a different CNV) as outliers in 2SD upper tails [perhaps because duplications may generate different lengths of transcript or alter the transcript splicing resulting in transcripts with different RNA stabilities, some with decay of the truncated transcripts (18–20)].

Thus, our analyses detected transcriptional effects for a very high proportion of genes known to be disrupted by rare CNVs. Furthermore, a higher proportion of known CNV genes were detected as expression outliers in 2SD tails than in 3SD tails (Fig. 3A). We also observed stronger transcriptional effects (i.e., more extreme Z-scores) from whole-gene deletions than from 1 or 2 exons deletions (P = 0.009; 2-tailed unpaired Student’s t-Test) and than from whole gene duplications (P = 1.03 × 10−17; 2-tailed unpaired Student’s t-Test) (Fig. 3B). The high proportion of outlier subjects from known CNV-carriers (versus non-CNV carriers) demonstrates non-random detection of expression outliers [see also (16)].

Genes with case-preponderant expression tails are enriched for brain expression and for CNV genes associated with neurodevelopmental disorders

We first examined genes with 2SD expression tails enriched for SZ cases at false discovery rate, FDR, <0.05 (Fisher’s Exact Test). The analysis was performed for 2SD ‘lower’ or ‘upper’ expression tail separately. One gene (LIM domains containing 1; LIMD1) with
lower 2SD expression tails and two genes (Semaphorin 6A; SEMA6A and WD repeat and FYVE domain containing 1; WDFY1) with upper 2SD expression tails showed enrichment of SZ cases in their respective tails (Supplementary Material, Table S3). SEMA6A is a transmembrane protein that plays a role in dendrite growth of spinal motor neurons (21) and development of visual pathways (22).

Because of the small number of outlier subjects in expression tail, we further relaxed the statistical threshold of defining SZ-case enrichment for 2SD expression tails by using nominal significance (Fisher's Exact Test \( P < 0.05 \)) rather than FDR < 0.05 (Table 2). We identified an over-representation of brain-expressed genes (23) (1.1-fold enrichment, \( P = 0.002 \); Fisher's Exact Test, corrected for 16 tests) in case-enriched upper tails (Table 2). There was also a 1.6-fold enrichment of genes within CNVs associated with neurodevelopmental disorders (24) (\( P = 0.011 \); Fisher's Exact Test, corrected for 16 tests) in case-enriched lower expression tail, the latter remaining nominally significant (\( P = 0.013 \); Fisher's Exact Test) even after excluding the SZ-risk CNV genes (i.e., 1q21.1, NRXN1, 15q13.2, 16p11.2 and 22q11.21) from the list of known neurodevelopmental-associated CNVs (Table 2). To further determine whether the observed over-representation of brain-expressed genes and CNVs associated with neurodevelopmental disorders in case-enriched 2SD expression tails was by chance, we also carried out the same analysis for control-enriched gene expression tails. Although there were more genes with control-enriched lower or upper expression tails than with case-enriched tails, we did not observe significant over-representation of brain-expressed genes or CNVs associated with neurodevelopmental disorders in control-enriched tails.

---

**Table 1. Outlier subjects in genes with one or two deleted exons**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Deletion exons</th>
<th>Transcripts All</th>
<th>w/del</th>
<th>Expression Z-score</th>
<th>Outlier</th>
</tr>
</thead>
<tbody>
<tr>
<td>OAS2</td>
<td>1</td>
<td>6</td>
<td>4</td>
<td>−0.89</td>
<td>No</td>
</tr>
<tr>
<td>CHST11</td>
<td>1</td>
<td>6</td>
<td>5</td>
<td>1.39</td>
<td>No</td>
</tr>
<tr>
<td>FARS2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>−3.75</td>
<td>Yes</td>
</tr>
<tr>
<td>SCFD2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>−3.51</td>
<td>Yes</td>
</tr>
<tr>
<td>TIFA</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>−5.66</td>
<td>Yes</td>
</tr>
<tr>
<td>ANAPC10</td>
<td>2</td>
<td>16</td>
<td>16</td>
<td>1.71</td>
<td>No</td>
</tr>
<tr>
<td>UQCR10</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>−0.04</td>
<td>No</td>
</tr>
<tr>
<td>KIAA1586</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>−5.82</td>
<td>Yes</td>
</tr>
<tr>
<td>TRIB3</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>−2.82</td>
<td>Yes</td>
</tr>
<tr>
<td>ARID1B</td>
<td>2</td>
<td>11</td>
<td>8</td>
<td>−2.57</td>
<td>Yes</td>
</tr>
<tr>
<td>NCOA7</td>
<td>2</td>
<td>13</td>
<td>8</td>
<td>−2.51</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Genes listed (\( n = 11 \)) are those non-pseudogenes with a predicted CNV (Affymetrix 6.0) >5 kb deleting 1-2 exons in >50% of ENSEMBL transcripts. Z-scored expression outlier subjects at 2SD lower-tails were counted. The significance of enrichment of these genes with 1 or 2 deleted exons was estimated through 100,000 permutations (empirical \( P = 1.5 \times 10^{-4} \) and \( P = 2.0 \times 10^{-5} \), respectively). Only 2/21 subjects with duplications disrupting 1 or 2 exons were outliers in upper tails and thus duplications were not listed here.
Duplications did not show significant unpaired Student outlier events for genes known to have 1 duplicated (Affymetrix 6.0 SNP array intensity data from the MGS sample (11)), (Table 2). We do not have a definitive explanation for the observation of an increased number of genes enriched for control outliers mainly in the upper 2SD expression tails (1200 versus 765). Although our study has more controls than cases (713 versus 634), the tests of enrichment in gene expression tails are corrected for the case/control imbalance. The increased number of genes enriched for controls in upper 2SD expression tails does not appear to be caused by a systematic expression distribution bias to lower expression in cases because, if this were the case, we would have expected to also observe a corresponding increase in the number of case-enriched lower 2SD expression tails, which was not found. However, we cannot rule out technical limitations such as our inability to detect very low transcript abundances (i.e., missed due to not reaching the threshold of detection). Finally, we did not find any differences of the average gene expression level or cell culture characteristics, such as viral load, energy status and growth rate (data not shown), between SZ cases and controls.

We also evaluated the gene set enrichment by using more stringent statistical cut-off (P < 0.01 and 0.001; Fisher’s Exact Test) for selecting case- or control-enriched expression tails. Consistent with results as described above, we found nominally significant enrichment of brain-expressed genes (23) and neurodevelopmental disorders-associated CNV genes (24), which did not survive multiple testing correction largely due to the smaller number of outlier genes in each gene set than using a less stringent P-value cut-off (Supplementary Material, Table S8). The enrichment of genes within CNVs involved in neurodevelopmental disorders in SZ case-enriched gene expression tails suggests that genes identified by outlier expression analyses may be relevant to SZ pathogenesis. 

**SZ cases showed higher outlier burden for genes spanned by CNVs associated with SZ or other neurodevelopmental disorders** 

Geneset enrichment analysis as described above relies on an arbitrary P-value cut-off for selecting case- or control-enriched expression tails. Furthermore, it examines only the impacted genes rather than the number of events, i.e., multiple outlier genes might be a result of only a single event (e.g., a large CNV). To complement the above-described gene set enrichment analysis, we therefore carried out a burden test (16,25) by examining whether cases appeared to be in expression tails more often than controls. We calculated the number of times (events) of being an outlier expression per case versus per control for all the expressed genes or for different subsets of genes (Table 2). We

<table>
<thead>
<tr>
<th>Gene group</th>
<th># Genes</th>
<th>Brain-expressed (#, %)</th>
<th>GWS CNVs (#, %)</th>
<th>SZ CNVs (#, %)</th>
<th>P value &lt; 0.05</th>
<th>Develop. CNVs (#, %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expressed genes</td>
<td>8355</td>
<td>6744</td>
<td>80.7</td>
<td>154</td>
<td>1.8</td>
<td>43</td>
</tr>
<tr>
<td>2SD Lower (case enriched)</td>
<td>728</td>
<td>578</td>
<td>79.4</td>
<td>20</td>
<td>2.7</td>
<td>7</td>
</tr>
<tr>
<td>2SD Upper (case enriched)</td>
<td>765</td>
<td>651</td>
<td>85.1</td>
<td>12</td>
<td>1.6</td>
<td>2</td>
</tr>
<tr>
<td>2SD Lower (control enriched)</td>
<td>880</td>
<td>735</td>
<td>83.5</td>
<td>25</td>
<td>2.8</td>
<td>1</td>
</tr>
<tr>
<td>2SD Upper (control enriched)</td>
<td>1200</td>
<td>979</td>
<td>81.6</td>
<td>30</td>
<td>2.5</td>
<td>7</td>
</tr>
</tbody>
</table>

8355 autosomal genes are expressed (RPKM ≥ 1) in 100% of the studied sample of 634 SZ cases and 713 controls (known SZ-risk CNV-carriers removed). Genes are classified as brain-expressed (n = 14295, ~81% of well-annotated genes) in adult brain per (23). Here, genes (n = 107) in SZ-risk CNVs are 1q21.1, NRXN1, 15q13.2, 16p11.2 and 22q11.21. Developmental disorder CNV genes (n = 940; 615 deletions and 325 duplications) per (24). The enriched subgroup of genes with a Fisher’s Exact Test P < 0.05 (after Bonferroni correction for 16 tests; one-sided) are highlighted in gray. GWS: genome-wide significant.
examined the significance of the case-control burden differences empirically in 10,000 permutations (16), controlling for gene number and size. For all 8355 expressed autosomal genes, we found very similar rates (events/subject) in cases and controls (202.8 versus 208.4, respectively; Table 3). For genes within SZ-associated CNVs, we found an increased outlier burden in SZ cases versus controls at the lower 2SD tails (1.20 versus 0.90, respectively), representing a 1.33-fold enrichment of outlier events (empirical P = 0.0002; Table 3). For genes within neurodevelopmental disorder-associated CNVs (after excluding SZ-CNVs), we also observed a higher case-burden of outlier events (1.16-fold enrichment; empirical P = 0.024).

**Sequencing expression outliers revealed enrichment of putative regulatory variants**

To explore whether the expression outliers could be explained by DNA variants, we sequenced 17 genes with 2SD tails enriched for SZ cases, selected by a priori evidence of their associations in SZ GWAS, SZ or autism spectrum disorder studies (ASD; OMIM #209850) CNV studies, or sequencing studies of rare mutations (Supplementary Material, Table S9). In 275 subjects (157 cases and 118 controls), we sequenced all coding exons and putative regulatory sequences (DNasel hypersensitive [HS] sites) within 50 kb upstream of these 17 genes. We identified a total of 728 DNA variants. The common variants (MAF ≥ 5%; n = 154) correlated well with those from public whole-genome sequencing (WGS) data sets (UK10K-Twins; 1000 Genomes CEU; Fig. 4A) for the targeted sequencing regions. To evaluate whether there was an enrichment of rare to low frequency [MAF < 5%; n = 574] variants in expression outlier subjects, we compared the number of DNasel HS variants, frameshift or nonsense variants, missense and synonymous variants in sequenced outliers versus non-outliers for each gene and in aggregate (Supplementary Material, Table S10). DNasel HS variants showed the strongest enrichment in outliers (P = 7.6 × 10−4, one-side Fisher exact test), with 7/17 genes showing nominally significant enrichments. In aggregate, missense, synonymous and frameshift/nonsense variants also showed enrichment in sequenced outliers (Supplementary Material, Table S10). Consistently, we also found enrichment of rare to low frequency putative regulatory variants and frameshift mutations in our outlier subjects versus the UK10K-Twins sample (P = 1.6 × 10−5 for variants in DNasel HS; P = 0.013 for nonsense or frameshift mutations; one-sided Fisher exact test; Fig. 4B; 1000 Genomes CEU were not used for analyzing rare variants due to small sample size). Putative regulatory variants were not enriched in the common variants (MAF ≥ 5%; Fig. 4C). The proportion of missense and synonymous coding variants (whether MAF ≥ 5% or MAF < 5%) in both sequencing datasets was similar (our dataset and UK10K-Twins; Fig. 4B and C). This suggests that rare to low frequency regulatory variants in aggregate likely explain some of the expression outliers.

We next estimated the proportion of the expression outliers in case-enriched tails that could be explained by known CNV and rare to low frequency sequence variants located in regulatory and coding sequences (Supplementary Material, Table S11) at an individual gene level, assuming these variants are all functional. Two (tubulin, gamma complex associated protein 5 [TUBGCP5] and adenosine deaminase [ADA]) and out of the 17 genes showed nominally significant enrichment of putative regulatory variants (within H3K4MeK3, H3K4MeK3, DNasel HS or CCCTC-binding factor [CTCF]) and coding variants of rare to low frequency. TUBGCP5 had the highest proportion (85.7%) of outliers that could be tentatively explained by genomic and bioinformatic analyses of the genetic variants (P = 0.035; by 100 000 permutations): out of seven cases (no controls) that were upper 2SD expression outliers, one with a known duplication, three with putative regulatory variants, two with missense and one with synonymous variants (Supplementary Material, Table S11). For ADA, out of the seven cases and one control that were in lower 2SD expression tails, two cases had frameshift variants and a third case had a putative regulatory variant in the H3K4meK1 peak (P = 0.046; by 100 000 permutations). Across the 17 sequenced genes, ~17.4% (0.85.7%) of expression outliers could be explained by rare to low frequency variants (Supplementary Material, Table S11).

Such a proportion (~17.4%) of expression outliers explained by genetic variants seemed to be in accordance with the known average heritability of gene expression in LCLs (h² = 0.21) from studying 856 monozygotic and dizygotic female twins (MuTHER report) (26). The low proportion of DNA variants potentially explaining expression outliers suggests that stochastic variations in cell culture may affect the transcriptional signatures. It is noteworthy that the quality of our expression data is high as indicated by mean correlation of gene expression between biological replicates (R = 0.992) and between technical replicates (R = 0.995). To assure the data quality in our expression outlier analysis, we have also carefully regressed out possible confounding effects of 12 covariates on expression (see methods). Although some other potential covariates like medication and smoking for each study subject were not included, these factors are not expected to pose large effects on expression variation in LCLs, because of the LCL culturing (cell divisions) time, making any epigenetic

**Table 3. Burden of genes with outlier subjects, in SZ cases and controls from analyzing 8355 autosomal genes expressed in 100% of the samples**

<table>
<thead>
<tr>
<th>Gene sets (expressed genes)</th>
<th>2SD lower-tail Events per SZ case</th>
<th>Events per control</th>
<th>2SD lower-tail Events per SZ case</th>
<th>Events per control</th>
<th>P</th>
<th>2SD upper tail Events per SZ case</th>
<th>Events per control</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>All genes (8355)</td>
<td>200.73</td>
<td>201.47</td>
<td>0.996</td>
<td>0.53</td>
<td>171.00</td>
<td>180.23</td>
<td>0.95</td>
<td>0.81</td>
</tr>
<tr>
<td>Brain-expressed (6700)</td>
<td>158.82</td>
<td>161.01</td>
<td>0.986</td>
<td>0.60</td>
<td>139.30</td>
<td>145.40</td>
<td>0.96</td>
<td>0.77</td>
</tr>
<tr>
<td>PGC2-SZ-GWS (154)</td>
<td>3.78</td>
<td>3.80</td>
<td>0.994</td>
<td>0.53</td>
<td>2.88</td>
<td>3.31</td>
<td>0.87</td>
<td>0.98</td>
</tr>
<tr>
<td>SZ-CNV (43)</td>
<td>1.20</td>
<td>0.90</td>
<td>1.33</td>
<td>0.0002</td>
<td>0.81</td>
<td>0.89</td>
<td>0.91</td>
<td>0.81</td>
</tr>
<tr>
<td>Developmental disorder dup&amp;del (326)</td>
<td>8.22</td>
<td>7.09</td>
<td>1.16</td>
<td>0.024</td>
<td>6.45</td>
<td>7.15</td>
<td>0.90</td>
<td>0.86</td>
</tr>
</tbody>
</table>

All are 8355 autosomal genes expressed (RPKM ≥ 1) in 100% of the studied sample of 634 SZ cases and 713 controls [known SZ risk CNV-carriers removed]. Genes are classified as brain-expressed (n = 14295; ~81% of well-annotated genes) in adult brain per (23). Here, genes (n = 107) in SZ-risk CNVs are: 1q21.1, NXN1, 15q13.3, 16p11.2 and 22q11.21. Developmental disorder CNV genes (n = 940; 615 deletions and 325 duplications) per (24). GWS: genome-wide significant. Outlier burden was calculated as the number of genes (with outlier subjects) per individual in non-SZ-associated-CNV carriers, and the statistical significance of higher burden in SZ cases was empirically derived from 10,000 permutations (by shuffling the case-control labels).
changes associated with medications, smoking or other clinical measurements reduced to minimal influences. Nonetheless, we expect our outlier detection for variants of large regulatory effect to be stable as shown above for known CNVs (Fig. 3). We have, however, found significant fluctuation of expression distribution of biological replicates (i.e., from different cell culture batches) or

Figure 4. Rare to low-frequency regulatory variants are enriched in sequenced outliers of 17 selected genes, and two frameshift variants identified in two ADA expression outliers cause NMD of the transcript. (A) Venn diagram shows the overlap of common variants (MAF ≥ 5%) identified in outlier DNA sequencing and in public whole-genome sequencing data sets (UK10K TwinsUK and 1000 Genomes). (B) Variants in DNaseI HS and nonsense/frameshift indels are enriched in our outlier-seq data set versus UK10K TwinsUK data set for rare to low frequency (MAF < 5%) variants, but not for those common variants (MAF ≥ 5%) (C). Fisher’s exact test was used and two-tailed P values are presented. 1000 Genomes was not used for comparisons in (B) and (C) because it is a small data set (only 85 CEU subjects). (D) Gene structure of ADA and its exons where the two frameshift mutations (chr20_43249673_CTCTT_del and chr20_43254291_T_ins) reside. (E) The two frameshift mutations were detected by multiplex outlier DNA sequencing (middle panel) and confirmed by Sanger sequencing (bottom panel), but the frameshift allele was absent in RNAseq (top panel). T allele of rs61737144 adjacent to the insertion chr20_43254291_T_ins and in the same haplotype was also not detected at the transcript level. Both findings support that the frameshift variants create premature stop codons that caused NMD of the transcript. (F) Distribution of Z-scored expression of ADA in 312 SZ cases and 322 controls. Two outlier subjects with the frameshift variants are indicated by arrows at the at lower 2SD tail.
Two frameshift variants in ADA caused expression outliers through nonsense-mediated decay (NMD)

ADA is a gene with a plausible role in SZ and ASD since it modulates the release of multiple neurotransmitters (27–29), and it is noteworthy that two ADA expression outliers each had a frameshift variant (chr20_43254291_T_ins in exon 5 and chr20_43249673_CTCTT_del in exon 10; Fig. 4D–F), representing an enrichment of frameshift variants (P = 0.003; by 10^3 permutations). These two frameshift variants are predicted to cause NMD of the transcript by creating a premature stop codon, which was confirmed by the absence of the frameshift allele in our RNAseq data (Fig. 4E). RNAseq data at exon 5 also showed the absence of allele T of rs61737144, the allele on the same haplotype as the frameshift allele of chr20_43254291_T_ins, further supporting NMD of the transcript as a result of chr20_43254291_T_ins (Fig. 4E and F). For chr20_43249673_CTCTT_del, for which a Taq-NMD of the transcript as a result of chr20_43254291_T_ins frameshift allele of chr20_43254291_T_ins, further supporting allele T of rs61737144, the allele on the same haplotype as the ancestry (EA) and tested its association with SZ. However, we identified one more SZ case and three controls carrying the frameshift variant, which did not support the association of chr20_43249673_CTCTT_del with SZ in the MGS EA sample.

Neurodevelopmental disorder-associated CNV genes often had case-enriched expression tails

We have found that the genes with case-enriched lower expression tails are enriched for neurodevelopmental disorder-associated CNV genes. This provided supportive transcriptomic evidence for likely neurodevelopmental abnormalities in SZ (33). Genetic overlap between SZ and various neurodevelopmental disorders has been reported (34,35). Some genetic variants associated with SZ, ASD, X-linked intellectual disability (XLID) and attention deficit hyperactivity disorder (ADHD) occur in the same molecular pathways and functional domains (35). For instance, genes associated with ASD, XLID, ADHD and SZ are mainly distributed in gene networks involved in the regulation of transcription, synaptic transmission, cell-cell communication, intracellular signaling pathways, cell cycle, metabolic processes and nervous system development (35). Furthermore, loci with balanced chromosomal abnormalities (BCAs, e.g., balanced translocation) in patients with neurodevelopmental disorders often harbor genes that confer risk across diagnostic boundaries (36), i.e., exhibit pleiotropy. Our findings are consistent with the growing evidence that SZ is a highly heterogeneous disorder.

Complexities and limitations of interpreting expression outliers

Our transcriptome outlier analysis did enrich for subjects carrying rare or low frequency putatively regulatory variants of large effect (DNaseI HS and frameshift). This is consistent with a recently reported excess of rare regulatory variants in expression outliers identified by using a similar Z-scoring approach (37) or by a multivariate metric Mahalanobis distance (MD) approach (38). Combined with the selection of expression tails enriched for SZ cases, outlier analysis can help identify gene sets relevant to SZ and neurodevelopmental disorders. However, the proportion of outliers that can be explained by a rare putative regulatory variant in most sequenced outlier genes was low (an average of 17%) even under the assumption that all the rare to low frequency variants for functional, implying the complexities of the genetic and epigenetic underpinnings of being an expression outlier. Outliers may result from recurrent mutations of large effect, de novo mutations (39), or simply the secondary physiologic-al consequence of the primary dysregulation of another gene. An evolving approach to detect and functionally interpret novel types of genetic variants (e.g., large multiallelic CNVs, mCNVs) resulted in mCNVs accounting for seven times the combined contribution of deletions and biallelic duplications to gene expression variation (40), and such mCNVs may also help explain expression outliers. Furthermore, since the functional annotation of non-coding sequences remains a major limitation (41,42), the genomic coverage of the chosen regulatory sequences in our target resequencing of the selected outlier genes may be incomplete. Although we have used a bootstrapping procedure to evaluate genes with outlier subjects enriched for SZ cases may help to prioritize genes in these SZ-associated CNV regions. For instance, one such gene, DCCRS at the 22q11.21 deletion region, is involved in plasticity of the prefrontal cortex (30–32), a brain region hypothesized to be fundamental to SZ pathophysiology. We did not find enrichment of Psychiatric Genomics Consortium, part 2 (PGC2) for SZ (7) genome-wide significant (GWS) genes among genes with outlier subjects over-represented by SZ cases in their expression tails, which is consistent with small effect sizes of common SNPs associated with SZ at PGC2-SZ loci.

Expression outlier analysis implicates SZ-associated CNV region genes

Both our gene set enrichment analysis and the case-control outlier burden test suggested that SZ-associated CNV genes were enriched for outlier genes with more SZ cases in the expression tails. Because we have excluded subjects carrying known SZ-associated CNVs, these outlier events cannot be explained by known SZ-associated CNVs previously detected from SNP array intensity data (11), suggesting that SZ-associated CNV genes are likely to have additional small rare CNVs beyond the detection resolution of SNP intensity arrays and/or other types of genetic variants of large functional effect on gene expression. As SZ-associated CNVs mostly span multiple genes, there is a need to determine which genes are ‘causal’. Identification of technical replicates (i.e., from the same RNA sample) (Supplementary Material, Figs S6 and S7). For instance, when one of the 22 biological replicates was found in a 2SD expression tail, we observed only ~10% chance that there were >3 replicates in the same expression tail (Supplementary Material, Fig. S6).

Although the outlier detection for technical replicates appeared to be more stable with ~3/4 of the 2SD expression tails having >3 replicates in the same expression tail, only ~36% of outlier genes have all the 11 replicates in the same expression tail (Supplementary Material, Fig. S7).

Discussion

We present here an approach to identify disease-related functional consequences of rare to low frequency genetic variants (e.g., CNVs, small indels or SNPs) on gene expression. Similar analyses from patients with ASD had previously provided insight into the functional consequences of de novo CNVs (16). We found that expression tails with more SZ cases than controls are enriched for genes spanned by CNVs associated with neurodevelopmental disorders (with or without inclusion of those CNVs associated with SZ) as well as CNVs associated with SZ (even though the studied individuals were not themselves carriers of SZ-associated CNVs). Furthermore, we have found that 17 selected outlier genes were enriched for rare regulatory variants.
the probability of a detected outlier subject in 2SD expression tails to enhance the confidence of outlier detection, we acknowledge that an unknown proportion of the detected expression outliers may still be false positives.

The use of LCLs as a cellular model versus brain tissues for transcriptome outlier analysis presents some clear limitations. Some of the gene expression changes in LCLs differ from brain, the latter presumably being the most relevant tissue for SZ pathogenesis. However, LCLs are the most accessible tissue with a sizable sample compared with other alternatives such as postmortem brain, or neuronal cell lines. The use of LCLs as a cellular model (versus brain tissue) for transcriptome outlier analysis presents some limitations. Some of the gene expression changes in LCLs differ from brain, the latter presumably being the most relevant tissue for SZ pathogenesis. However, LCLs are the most accessible tissue with a sizable sample compared with other alternatives such as postmortem brain or neuronal cell lines. Furthermore, although the correlation of mean expression levels between brain and blood is moderate (r = 0.24–0.64), about 35–80% of known transcripts are expressed in both tissues (43,44). A large proportion of gene expression signatures are shared between different tissues (45–52), and we thus expect that many of the LCL-detected outlier genes also show relevant transcription in the brain. Indeed, we have found an enrichment of brain-expressed genes among those outlier genes with more SZ cases in their expression tails (Table 2). Moreover, maternal immune activation (e.g., via an infectious exposure such as influenza) has been historically associated with risk of SZ and autism, and blood levels of cytokines (e.g., IL1β, IL2RA, IL6 and TNF) are elevated in SZ patients [see reviews (53–55)]. Large population-based cohort research suggests a shared etiology between SZ and other autoimmune diseases, with increased risks of 1.1–1.6 for SZ (56). Besides the epidemiological supporting evidence, GWAS results strongly suggest the involvement of immune mechanisms in SZ pathogenesis. The strongest and most replicable SZ GWAS finding is at the extended MHC (xMHC) region (2–5,7,8). Even without considering the xMHC region, key immune processes such as TGF-beta signaling, B-cell activation and T-cell activation have been implicated (57). Furthermore, immune-related genes were enriched among the transcripts differentially expressed by SZ affection status in our transcriptome profiling study (58).

Another potential difficulty resulting from using LCLs as a model is that some outliers may be due to lab-induced de novo mutations during the Epstein–Barr virus (EBV) transformation and cell passages (i.e., not present in the native blood cells of an individual). However, we expect any such effect to be minor, as it has been shown that whole-exome sequencing (WES) data of DNAs from peripheral blood mononuclear cells (PBMC) and from EBV-transformed lymphocytes (early passages, such as with MGS LCLs) from the same donor are >99% concordant (59–61). Furthermore, we did not find a correlation between the number of outliers per individual and the EBV copy number of a LCL, and we have included EBV copy number as a covariate in the analyses (and also excluded monochonal or pauciconal LCLs) to further minimize any confounds from that source. Indeed, the gene set enrichment analysis gave very similar results after excluding the genes with EBV-associated expression (62) (data not shown).

Finally, we restricted our outlier gene burden and enrichment analyses to a few gene sets of pathophysiologically relevance to SZ. However, testing even these gene sets is complex because of possible overlapping genes in different datasets (e.g., the neurodevelopmental-associated CNVs in Table 2) and consequently, appropriately correcting for multiple hypothesis testing is challenging (63). Genes with expression tails containing outlier subjects over-represented by SZ cases are more likely to harbor functional variants of large effect on gene expression relevant to SZ. The enrichment of such genes in gene sets spanned by neurodevelopmental disorder-associated CNVs offered further support for the neurodevelopmental hypothesis of SZ pathogenesis. The presented outlier gene analysis approach represents a novel alternative to conventional differential gene expression (i.e., comparing averages across the whole distribution of expression) for better understanding the regulatory effect of rare to low-frequency variants and resultant molecular mechanisms underlying the risk for complex disorders.

Materials and Methods

Samples

The analyzed sample consisted of 634 SZ cases and 713 controls. These EA subjects are from the GWAS- and CNV-studied portion of the MGS collection (2,5,11). We excluded the samples that carry a deletion or duplication at known SZ-associated CNV loci (i.e., 1q21.1, NRXN1, 15q13.2, 16p11.2 and 22q11.21) (9–11,64), which show expected large transcriptional changes associated with each CNV [e.g., (65)]. Detailed phenotypic data have been previously described (2,66). The NorthShore University HealthSystem Institutional Review Board approved the study.

Cell culture and RNA preparation

The source of LCLs was the Rutgers University Cell and DNA Repository (RUCDR) (58). For each LCL, we measured EBV (viral) load, viable cell count (to index growth rate) and ATP level (to index energy status) at cell harvest (for use as covariates in expression analyses). Among the analyzed 634 SZ cases and 713 controls, cell culture and RNA preparation of 312 cases and 322 controls were carried out as one large batch and the rest as a second batch. Detailed methodology was as previously described (58). We considered batch as a possible confounding variable in the analysis and also evaluated the bath effect by analyzing the two batches separately.

RNAseq and data processing

RNA sequencing was carried out at the University of Minnesota Genomics Center (UMGC) on an Illumina HiSeq2000 at a depth of ~10 M reads/sample. We aligned the 50-bp single reads to the human reference gene map (Gencode v20) using the mapping tool Tophat v2.0.5 (67), allowing for two mismatches. We counted the raw reads by using the HTseq-count script (www.huber.embl.de/users/anders/HTSeq/doc/overview.html) (68) and calculated gene level expression as RPKM (69) based on the exon model of the longest transcript of a gene (Gencode v20). We then quantile-normalized gene level RPKMs to account for batch/run bias (Supplementary Material, Fig S1). The mean Pearson correlations of gene level RPKM among 10 technical replicates (i.e., same RNA) and among 16 biological replicates (independent cell cultures of the same LCL) were 0.991 and 0.988, respectively, indicating good data quality. Based on the correlations between technical replicates (Supplementary Material, Fig S2), we excluded genes with extremely low expression and thus high fluctuation of expressions between replicates, we considered RPKM > 1 as a cut-off for an expressed gene. All samples included in the RNAseq had >6 M mappable reads. We performed sample QC.
by (1) checking for consistency between expression levels of chromosome X (XIST) and chromosome Y genes (RPS4Y1, ZFY, USP9Y, DDX3Y, UTY, KDM5D, EIF1AY) versus reported sex, and (2) by comparing RNAseq-called genotypes [using SAMTools mpileup function (70), requiring >8 reads at a called SNP site] with previous GWAS SNP genotypes (Affymetrix 6.0) (2,5) for a panel of 175 informative SNPs. After sample QC, 634 SZ cases and 713 controls remained for expression outlier analyses.

Expression outlier analysis

We include in the analyses only genes expressed in 100% of the samples passing QC to decrease bias introduced by genes with low expression levels. We ended with 8355 autosomal genes in all samples. We regressed out possible confounding effects of 12 covariates on expression, namely, age, sex, major ancestry principal components (EA PC 1–5) calculated from SNP genotypes, LCL transformation site (RUCDR and Australia), RNAseq batch, viral load, energy status and growth rate (2,58). We then calculated the Z statistic of the log-transformed gene expression value of each subject for each gene by using the ‘scale’ function in R (as described in (16)). Based on the Z-score distribution of a gene, we used a cutoff of 2SD or 3SD of the Z-score values of all samples to define genes with extreme expression Z-scores, i.e., ‘outliers’, where a negative Z-score denotes the ‘lower tail’ and a positive Z-score denotes the ‘upper tail’.

To calculate the probability of an individual being expression outlier at the 2SD cut-off, we applied a distribution-independent bootstrapping procedure (17). Bootstrapping can provide an indirect method to assess the properties of the distribution underlying the sample and the parameters of interest that are derived from this distribution (17). For each gene, we randomly removed 10% of the non-outlier (at 2SD cut-off) subjects and recalculate the Z-score based outlier analysis as described above 1000 times. We then calculated the 2SD outlier probability (for upper or lower expression tails) of an individual from the 1000 bootstraps, and defined outliers as those with >95% probability. For each gene with outliers, we then counted the number of SZ cases and controls in the lower and/or upper expression tails (at 2SD or 3SD cut-offs), and calculated enrichment of SZ cases or controls in a tail by applying the Fisher’s Exact Test (two-sided).

Gene set enrichment analyses for genes showing case or control-preponderant outliers

We assembled different gene sets (Supplementary Material, Table S9) which are deemed relevant to SZ pathogenesis based on public databases and literature searches: (1) Adult brain-expressed genes (n = 14 295) that account for ~81% of well-annotated protein-coding genes (23). (2) SZ-CNV genes (n = 107) that are within the following rare (<1%) and large (>100 kb) CNVs associated with SZ: 1q21.1, NRXN1, 15q13.2, 16p11.2 and 22q11.21 (9–11,64). (3) Genes (n = 435) with SNPs (within 500 kb) showing genome-wide significant (GWS) association to SZ in the Psychiatric Genomics Consortium (PGC2) sample (7). (4) Neurodevelopmental disorder-associated CNV genes (n = 940; 615 deleted genes and 325 duplicated genes, but only 837 unique genes) (24). For outlier genes with expression tails enriched by SZ cases or controls we counted the number of genes with such outliers (for both tails separately) in each gene set. We then estimated the enrichment of these genes with outlier subjects in a gene set using all the LCL-expressed genes in each gene set as the denominator. We evaluated the significance of the gene set enrichment using Fisher’s Exact Test (one-sided).

Case/control burden test for expression outliers

We calculated burden as the number of expression outlier events per individual (16,25). We computed the normalized outlier burden enrichment ratio as (total N of these genes in cases/total N of events in controls) × (control N/case N) to account for different case and control N’s. We assessed the statistical significance of this enrichment in cases by shuffling the case-control labels for 10 000 permutations (16).

Resequencing of the selected outlier genes

For part of the sample that was initially available and analyzed (312 cases and 322 controls), we sequenced 275 subjects (157 SZ cases and 118 controls) that fell within both the 2SD expression tails (upper and lower) of 17 selected genes with case-enriched 2SD expression tails at that time. The 17 genes were selected based on prior evidence of their associations in SZ GWAS, SZ or ASD CNV studies, or ASD-associated rare mutations. The sequencing targets were all the coding exons (not including 3’-UTR) and all the putative regulatory sequences 50 kb upstream of the transcript start as defined by DNase1 Hypersensitivity Uniform Peaks from ENCODE (ENCODE Jan 2011 Freeze; Sept 2012 Analysis Pubs). We used multiplex PCR to enrich target sequences by Fluidigm Access Array: a total of 800 amplicons in two multiplex pools (a multiplex of 480 amplicons for 48 samples per array) (Supplementary Material, Table S12). Average amplicon size ranged from 150 to 200 bp, and PCR primers were designed and synthesized by Fluidigm. Multiplex PCR, sample barcoding and sample pooling were carried out on Fluidigm’s BioMark system following the target resequencing protocol (Fluidigm) (Supplementary Material, Fig. S3). High throughput target resequencing was carried out on a HiSeq2000 (illumina) at UMGC. A total of 166 175 952 reads (paired-end reads; 2 × 100 bp) were produced from 1 HiSeq2000 lane. For mapping sequence reads and variant calling, we followed the GATK SNP/Indel calling pipeline similar to a previously described approach (71). Briefly, BWA was used for mapping reads to human reference genome (hg37: human_g1k_v37.fasta). Sequence variants were called by GATK Unified Genotyper (Filter: ≥25 >coverage; ≥Q30 quality; allele balance <0.75). No sample was excluded based on the average number of mapped reads/sample or the average base quality score (Supplementary Material, Fig. S4). A total of 32 amplicons (4% of the 800 targeted amplicons) with average read depth <20 × were excluded (Supplementary Material, Fig. S4). We had a 99.4% genotype concordance rate for eight available SNPs for comparison between the current resequencing dataset and the previous MGS GWAS genotype data (2).

Sanger sequencing confirmation and taqMan genotyping

The two frameshift variants (chr20:43249673_CTCTT_del and chr20:43254291_T_ins) in ADA identified by NGS were further confirmed by Sanger sequencing on ABI 3730DNA Analyzer (Life Technologies). TaqMan genotyping of chr20:43249673_CTCTT_del in the MGS EA sample (2663 cases and 2653 controls) was performed as described previously (72,73) on LightCycler 480 II (Roche). We did not genotype chr20:43254291_T_ins, because the TaqMan custom assay (Life Technologies) could only be designed for chr20:43249673_CTCTT_del. We called genotypes by each 384-well plate where we included at least two replicate DNAs of the known heterozygous rare allele carriers to help with genotype clustering. The genotyping completion rate was 98%.
**Supplementary Material**

Supplementary Material is available at HMG online.

**Acknowledgements**

MGS includes P. V. Gejman, A. R. Sanders, J. Duan (NorthShore University HealthSystem, and University of Chicago, IL, USA), D. F. Levinson (Stanford University, CA, USA), J. Shi (National Cancer Institute, MD, USA), N. G. Buccola (Louisiana State University Health Sciences Center, LA, USA), B. J. Mowry (Queensland Brain Institute, The University of Queensland, Australia), R. Freedman, A. Olintcy (University of Colorado Denver, CO, USA), F. Amin (Atlanta Veterans Affairs Medical Center and Emory University, GA, USA), D. W. Black (University of Iowa Carver College of Medicine, IA, USA), J. M. Silverman (Mount Sinai School of Medicine, NY, USA), W. F. Byerley (University of California at San Francisco, CA, USA), C. R. Cloninger, D. M. Svrakic (Washington University, MO, USA).

**Conflict of Interest statement:** The authors declare no conflicts of interest.

**Funding**

This study also makes use of whole-genome sequencing data ( TwinsUK) generated by the UK10K Consortium. A full list of the investigators who contributed to the generation of the data is available from UK10K Project homepage. Wellcome Trust award WT091310 provided funding for UK10K. We thank the study participants of MGS. The transcriptomic work was primarily supported by RC2MH090030 (to A.R.S.), R01MH094116 (to H.H. H.G.) and R01MH094091 (to P.V.G.). MGS was mainly supported by R01MH059571, R01MH081800 and U01MH079469 (to P.V.G.); and other NIH grants for other MGS sites (R01MH067257 to N.G.B., R01MH059588 to B.J.M., R01MH059565 to R.F., R01MH068757 to F.A., R01MH060870 to W.F.B., R01MH059566 to D.W.B., R01MH059586 to J.M.S., R01MH061675 to D.F.L., R01MH060879 to C.R.C., U01MH046276 to C.R.C. and U01MH079470 to D.F.L.). This work was also supported by NorthShore University HealthSystem Research Career Development Award (to J.D.).

**References**


