doi: 10.1093/nar/gkw902

SZGR 2.0: a one-stop shop of schizophrenia candidate genes

Peilin Jia[1], Guangchun Han[1], Junfei Zhao[1], Pinyi Lu[1] and Zhongming Zhao[1,2,*]

[1]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA and [2]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

Received August 1, 2016; Revised September 17, 2016; Accepted October 06, 2016

ABSTRACT

SZGR 2.0 is a comprehensive resource of candidate variants and genes for schizophrenia, covering genetic, epigenetic, transcriptomic, translational and many other types of evidence. By systematic review and curation of multiple lines of evidence, we included almost all variants and genes that have ever been reported to be associated with schizophrenia. In particular, we collected ∼4200 common variants reported in genome-wide association studies, ∼1000 *de novo* mutations discovered by large-scale sequencing of family samples, 215 genes spanning rare and replication copy number variations, 99 genes overlapping with linkage regions, 240 differentially expressed genes, 4651 differentially methylated genes and 49 genes as antipsychotic drug targets. To facilitate interpretation, we included various functional annotation data, especially brain eQTL, methylation QTL, brain expression featured in deep categorization of brain areas and developmental stages and brain-specific promoter and enhancer annotations. Furthermore, we conducted cross-study, cross-data type and integrative analyses of the multidimensional data deposited in SZGR 2.0, and made the data and results available through a user-friendly interface. In summary, SZGR 2.0 provides a one-stop shop of schizophrenia variants and genes and their function and regulation, providing an important resource in the schizophrenia and other mental disease community. SZGR 2.0 is available at https://bioinfo.uth.edu/SZGR/.

INTRODUCTION

Schizophrenia is a chronic and socially disabling disorder characterized with cognitive impairment, increased mortality and low fecundity (1). It is one of the major mental disorders, impacting ∼1% of the general population worldwide. Previous studies have indicated a variety of heterogeneous risk factors underlying schizophrenia, including a large genetic component as well as environmental and other factors (2). However, the etiology of schizophrenia has not been fully understood. In 2009, we developed SchiZophrenia Gene Resource database (SZGR), which collected association studies, linkage data, expression and other annotation data that were available for schizophrenia (3). Since its release, SZGR has served as a main platform and repository for the research community of schizophrenia and other neuropsychiatric diseases.

During the past several years, we have witnessed an unprecedented growth in the generation of genetic, genomic and functional data in order to better understand the etiology of schizophrenia, thanks to the rapid evolution of high-throughput technologies and increasing funding to many large-scale studies. Data volumes have accumulated more than exponentially. While genome-wide association studies (GWAS) continue to uncover common genetic variants that are statistically associated with the disease, the applications of next-generation sequencing (NGS) technologies have enabled the discovery of rare variants and *de novo* mutations (DNMs). In addition to single nucleotide polymorphisms (SNPs), many complex structural variants, including copy number variants (CNVs), were also reported as the risk factors of schizophrenia. Similar to other mental disorders, schizophrenia is now widely recognized to involve a combination of common variants, rare variants and DNMs underlying its genetic architecture. Moreover, epigenetic studies and transcriptome-wide NGS analyses using disease relevant tissues (e.g. various brain regions) have revealed a large number of genomic regions and genes that show abnormal changes in schizophrenia patients compared to the unaffected subjects.

Importantly, functional annotation data, especially for brain and the nervous system, have grown dramatically. As schizophrenia has been commonly believed as a neurodevelopmental disorder, the data from brain and nervous system is critical for deeper understanding and accurate interpre-

[*]To whom correspondence should be addressed. Tel: +1 713 500 3631; Fax: +1 713 500 3929; Email: zhongming.zhao@uth.tmc.edu

Disclaimer: The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

tation of risk variants in schizophrenia, as well as in other psychiatric diseases. These resources included deeply categorized temporal and spatial brain tissue expression profiles (4), expression quantitative trait loci (eQTL) (5) of brain tissues, methylation QTLs (meQTLs) of brain tissues (6) and functional annotations specially for brain regions, all of which have undergone dramatic growth during the past several years. In particular, the Genotype-Tissue Expression (GTEx) project conducted eQTL analysis for many tissues, including 10 brain regions from hundreds of donors (7). The Roadmap Epigenomics project also included brain samples (8) although its main aim was to provide functional annotations such as enhancers.

With the massive amounts of schizophrenia research data being generated in the past years and even more being expected in the near future, it is important to systematically collect, curate and analyze the data so that disease causal markers can be identified for better understanding its genetic architecture and potential treatment. In this work, we present SZGR 2.0, a one-stop shop of schizophrenia genes. Our new database has a substantial expansion from the previous SZGR (3), including novel data types, substantial addition of data content, useful online tools and a more user-friendly interface. By systematically searching and manually reviewing the literature in PubMed, we included almost all the variants and genes that have ever been linked to schizophrenia. To facilitate better interpretation, we included various functional annotation data from disease relevant tissues. Finally, we performed cross-study, cross-type and integrative analyses with supportive evidence.

FROM SZGR TO SZGR 2.0

The updates in SZGR 2.0 compared to SZGR were at three levels and substantial. First, data type and data volume were significantly expanded. SZGR was initially released in 2009 when the majority of schizophrenia GWA studies had not been released. In SZGR 2.0, the new data types included GWAS results, CNV, DNMs, brain specific expression, epigenetics data and antipsychotic drugs and drug–gene interaction data. For many data types, the content has been substantially updated, including the genes that were co-mentioned with schizophrenia in literature, linkage data and functional annotations. Second, the functional annotations and supportive lines of evidence have been largely expanded. We included brain specific eQTL, meQTL and gene expression data, which were generated after our first release of SZGR, for the annotations and interpretation of genes and variants in the current version of the database. We expanded the canonical pathway annotations and manually collected a total of 39 supportive gene sets that had been implicated in neurodevelopment, cognition and brain functions. Third, new online tools and new web interfaces were developed and implemented. For example, the new gene page provides a comprehensive annotation for genes. The SNP page, which was not supported in SZGR, enables the exploration of SNPs reported in GWAS. We also updated the web interfaces to make it more user-friendly. The content was organized in six groups: Genetics, Transcriptome, Epigenetics, Translational, Statistics and Document.

SCHIZOPHRENIA DATA COLLECTION, INTEGRATION AND ANALYSIS

By both text-mining and manual curation, we collected a variety of different types of data and, for each data type, we systematically searched the literature in PubMed for schizophrenia studies. This often resulted in multiple datasets for each data type. To better organize the data, we developed the format *type*:*study* to name a dataset, where *type* indicates a general data type and *study* refers to a particular publication or study, which is often labeled by the first author and publication year. Throughout this work, a dataset refers to the collection of genes and variants from a single study. For example, DEG:Zhao_2015 refers to a dataset including the genes reported in a whole transcriptomic study of schizophrenia brain tissue in 2015 (9), belonging to the data type DEG (differentially expressed genes). In total, we organized all datasets as eight types of evidence supporting genetic variants or genes in association with schizophrenia (Table 1). A detailed description of each dataset was presented in Supplementary Table S1. For genetic variants, we mapped them to the corresponding genes using wANNOVAR (10). A gene would have the same categories as the genetic variants which are located in the gene or nearby the gene (50 kilobase pairs upstream or downstream). A gene would have multiple categories if it contained variants from different *type:study* categories.

Common genetic variants (data type: CV)

We curated a total of 4261 schizophrenia-associated SNPs from five independent sources and organized them as five datasets: GWAS Catalog (11,12) (labeled as CV:GWAScat in SZGR 2.0, where CV indicated common variants and GWAScat denoted GWAS Catalog), GWASdb (13,14) (CV:GWASdb), PheWAS (15) (CV:PheWAS), Psychiatric Genomics Consortium (PGC) (16) (CV:PGC128) and a multi-stage GWA study (17) (CV:Ripke_2013). In all cases, we included only variants with genome-wide significance, e.g. $P < 1 \times 10^{-8}$ or meeting the specific threshold cutoff in the original studies. A comparison of the five sets of variants showed that each dataset had many unique variants and genes, though many were overlapped as well (Supplementary Figure S1). This feature reflects the complex nature of the mental disease and the importance to integrate these datasets. We categorized the SNPs from the five datasets into the tier 1 SNP set, reflecting their strong evidence in association with schizophrenia.

In addition, considering the polygenic nature of schizophrenia and numerous lines of evidence for regulatory variants enriched in GWAS results surpassing liberal significance thresholds of association (e.g. $P < 0.05$ or $P < 1 \times 10^{-3}$) (18–20), we included SNPs with nominal significance from PGC GWAS summary data (nominal P-value < 0.05, CV:PGCnp) and referred them as tier 2 SNP set. Although these SNPs did not reach the stringent genome-wide significance, some of them may still be involved in schizophrenia as causal or regulatory SNPs (21).

Table 1. Description of major data types

Data type	# publications reviewed	# datasets	# variants	# genes	# samples
ADT	1	1		49	
Common variant (CV)	Manual curation	6	>900 000 SNPs (after imputation)	2367 from tier 1 SNPs; 10856 from tier 2 SNPs	
***De novo* mutation (DNM)**	∼80	11	∼1000 DNMs	806	>1680 trios, >800 cases/controls
Copy number variation (CNV)	500	4	15 rare and replicated CNVs	215	
Differentially expressed gene (DEG)	>140	3		240	70 brain and ∼800 blood samples
Differentially methylated gene (DMG)	>60	5	4651 DMGs	4651	
Linkage	1	1	22 regions	99	Swedish national sample (5001 cases and 6243 controls)
Co-occurrence	Systematic search	1		3027	

De novo mutation (data type: DNM)

DNMs recently attracted much attention for their potential roles in schizophrenia. With the rapid development of NGS, more and more schizophrenia patients and their parents have been sequenced at the whole exome or genome level, with the aim to discover sporadic mutations that are only present in the schizophrenia probands but not in the parents. We searched in PubMed for large-scale sequencing projects that could detect DNMs in schizophrenia patients, using the keywords 'schizophrenia and DNM'. We required the publication date to be between 1 January 2008 and 30 June 2016. A total of 105 publications were obtained. Most of these studies were conducted using trio samples while a few employed the case-control design. We particularly focused on *de novo* coding mutations including single nucleotide variants and small insertions and deletions (indels). We reviewed these publications manually and collected 10 whole exome sequencing projects and one targeted sequencing based project (Supplementary Table S1). All of these projects utilized the design of trios or complex families (with additional family members or including unaffected trios as controls), with the sample sizes ranging between 14 and 623 families. For those studies that sequenced schizophrenia as well as other psychiatric disease samples, we extracted DNMs that only occurred in schizophrenia patients. Although schizophrenia is mainly an adult mental disorder, there is a rare form of the disease called childhood-onset schizophrenia (COS). Because we used the keyword 'schizophrenia' for searching, our list of publications included studies of all forms of schizophrenia, including one study that sequenced COS. That is, our data included ∼20 DNMs in COS patients. The targeted sequencing study (22) employed Sanger sequencing to examine the coding regions and splice site junctions of 401 synapse-expressed genes in 142 autism and 143 schizophrenia patients. Notably, these 401 genes were also recruited into our new SZGR 2.0 database as a supportive functional annotation set. We remapped all DNMs to the human reference genome (hg19) whenever applicable and annotated them using wANNOVAR (10) for their functional roles. In total, we collected 1057 records with DNMs in schizophrenia. Among them, 845 were amino-acid changing DNMs involving 806 genes.

Rare and replicated copy number variations in schizophrenia patients (data type: CNV)

To collect schizophrenia candidate CNVs, we searched PubMed using the terms [schizophrenia and 'CNVs'] or [schizophrenia and 'CNV'] for publications before 1 June 2016. We manually read the 341 resultant publications and summarized 15 replicated CNV regions associated with schizophrenia. The number of CNVs might change slightly depending on how the CNVs were counted. These regions were intensively studied (23,24), replicated and reviewed (1,25) through large-scale genomic analyses. Recently, an additional study performed systematic evaluation of these regions in independent samples of schizophrenia patients ($n = 6882$) and controls ($n = 6316$), and successfully validated 13 of them (26), indicating that these CNVs were most likely associated with schizophrenia. We labeled the genes overlapping with CNVs as CNV:YES.

Linkage data (data type: LK)

We had previously collected meta-analysis of linkage studies in our initial release of SZGR (27). There have been only few new linkage studies after 2010. However, the 22 regions reported in the Ripke *et al*. study had been conventionally referred to as linkage loci in many publications, although it was a multi-stage GWA study. In SZGR 2.0 database, we followed the convention (17) and included these 22 regions as linkage regions. In total, we found 99 genes spanned by these loci and labeled them with LK:YES. Notably, we had included the 24 leading SNPs in the data type CV (labeled as CV:Ripke_2013). The genes in CV:Ripke_2013 and the genes in LK:YES were considered as different datasets in our SZGR 2.0 database.

Differentially expressed genes in schizophrenia patients (data type: DEG)

Many studies have attempted to decode the transcriptome of schizophrenia patients. However, most of them used blood samples, mainly because brain tissues were hard to obtain or store (28). To ensure the quality of DEGs being deposited into our database, we required that studies would only be included if they used brain tissues with sample size good for statistical test (e.g. >20 cases and controls, respectively) or if they used non-brain tissues, the number of samples was more than 100. After reviewing the related literature, we recruited three studies: DEG:Maycox_2009 (brain tissue) (29), DEG:Zhao_2015 (brain tissue) (9) and DEG:Sanders_2013 (hundreds of blood samples) (30,31). Specifically for DEG:Sanders_2013, we selected DEGs from this study because it had large sample size, even though lymphoblastoid cell lines were used. The measurements for the DEGs in SZGR 2.0 included *P*-values, *q*-values, β and fold change (FC) n both the discovery dataset and the replication dataset, whichever was available.

Differentially methylated sites and genes in schizophrenia patients (data type: DMG)

We searched PubMed using the term ‘methylation and schizophrenia and genome-wide’. After careful review, a total of ten genome-wide methylation studies were retained, although only five of them had differentially methylated genes (DMGs) or probes available to us. We remapped the probes from these studies using the Illumina Infinium HumanMethylation450K v1.2 for their locations in hg19. We also filtered for genes that were in the current RefSeq annotation. In total, we obtained 8045 records, 7897 CpG probes and 4651 nearest genes. These genes were named as DMGs.

Large-scale text-mining of genes reported for schizophrenia (data type: PMID)

The number of publications per gene in which the gene’s symbol co-occurs with schizophrenia keywords reflects how frequently the gene has been studied in schizophrenia. We started with all RefSeq genes (accessed on 23 May 2016) and searched PubMed using four keywords related to schizophrenia: *schizophrenia*, *schizophrenic*, *schizotypy* and *schizotypal*. We obtained 3073 genes that had been co-mentioned with these keywords in the titles or abstracts. The ten most frequently studied genes were *BDNF* ($n = 579$ publications), *COMT* ($n = 555$), *DISC1* ($n = 516$), *NRG1* ($n = 388$), *DRD2* ($n = 284$), *DTNBP1* ($n = 221$), *TNF* ($n = 209$), *FOS* ($n = 206$), *CYP2D6* ($n = 174$) and *ERBB4* ($n = 166$). Of note, we did manual check of the 3073 genes and removed those whose related abstracts were false associations. For example, some genes were collected due to unrelated abbreviations, such as *CAMP* ($n = 271$, frequently used in cAMP pathway in literature) and *NNT* ($n = 231$, matched to ‘number needed to trait’). Some genes were collected due to apparent false meaning, such as the genes *TRD* (abbreviation for Treatment Resistant Depression), *BP1* and *BP2* (abbreviation for break point 1 and 2), *CA1*, *CA2* and *CA3* (brain regions in hippocampus), among others. After manual check, a total of 3027 genes remained and collectively named as PMID:cooccur.

Antipsychotic drugs and targets (data type: ADT)

Antipsychotic drugs are medications primarily used in the treatment of schizophrenia and other related mental disorders. In our previous work, we had collected a total of 43 antipsychotic drugs and their target genes based on DrugBank annotations (32). In this study, we utilized the same strategy and queried the most recent release of DrugBank data (version 5.0). In total, we obtained 49 target genes of these antipsychotic drugs, collectively referred as ADT:Sun_2012, where ADT denotes antipsychotic drug target. Both drugs and their targets were included in SZGR 2.0.

SUPPORTIVE DATA

While the schizophrenia data collected above provided direct evidence for candidate variants and genes to be associated with schizophrenia, supportive data can be used to explore how they function or interact in the cellular systems. This provides indirect evidence. Such data was often generated using disease-relevant tissue. To this end, we collected functional annotation data especially in normal brain tissues, including eQTL, gene expression, meQTL and promoters and enhancers, among others.

Brain eQTL data is relatively rare compared to the eQTL data for other tissues due to the difficulty in acquiring the related tissue. We managed to collect 12 eQTL datasets from three sources specifically for brain tissues, including the Myers’ dataset (5) downloaded from SeeQTL (33), which included both *cis*- and *trans*-eQTL (5), 10 eQTL datasets for different brain regions from GTEx and the results from a meta-analysis of five eQTL datasets (34). These datasets were used for annotations of both SNPs and genes. In this study, we defined eSNPs as those that were significantly associated with gene expression changes of at least one gene, according to the criteria listed in the original studies. Similarly, we defined eGenes as those that were significantly associated with at least one SNP. SNPs collected in SZGR 2.0 were annotated as the eSNPs if they were observed in any of the 12 datasets. The same strategy was applied to the genes for eGenes.

We collected four sets of expression data to provide comprehensive and complementary information for our schizophrenia candidate genes. Our rationale to collect these datasets were based on the facts that schizophrenia has been widely considered as a brain disorder with abnormal activities in certain brain regions as well as abnormalities during the critical stages of brain development (35). Three of our annotation datasets were for brain tissues while the fourth one was for comparison across multiple tissues including brain. The three brain datasets (namely DS1, DS2 and DS3, where DS indicates dataset) provided transcriptomes of deep classification of brain areas (DS1, downloaded from the BrainSpan database) (36), developmental stages (DS2, downloaded from the BrainCloud) and a spatiotemporal classification of brain (DS3, downloaded from the BrainSpan database) (35). The multiple-tissue gene expression data was retrieved from GTEx v6 (referred to as DS4). We included 27 tissues that had more than 30 samples. For each gene, a boxplot was presented for the GTEx tissue-specific gene expression so that it could provide a di-

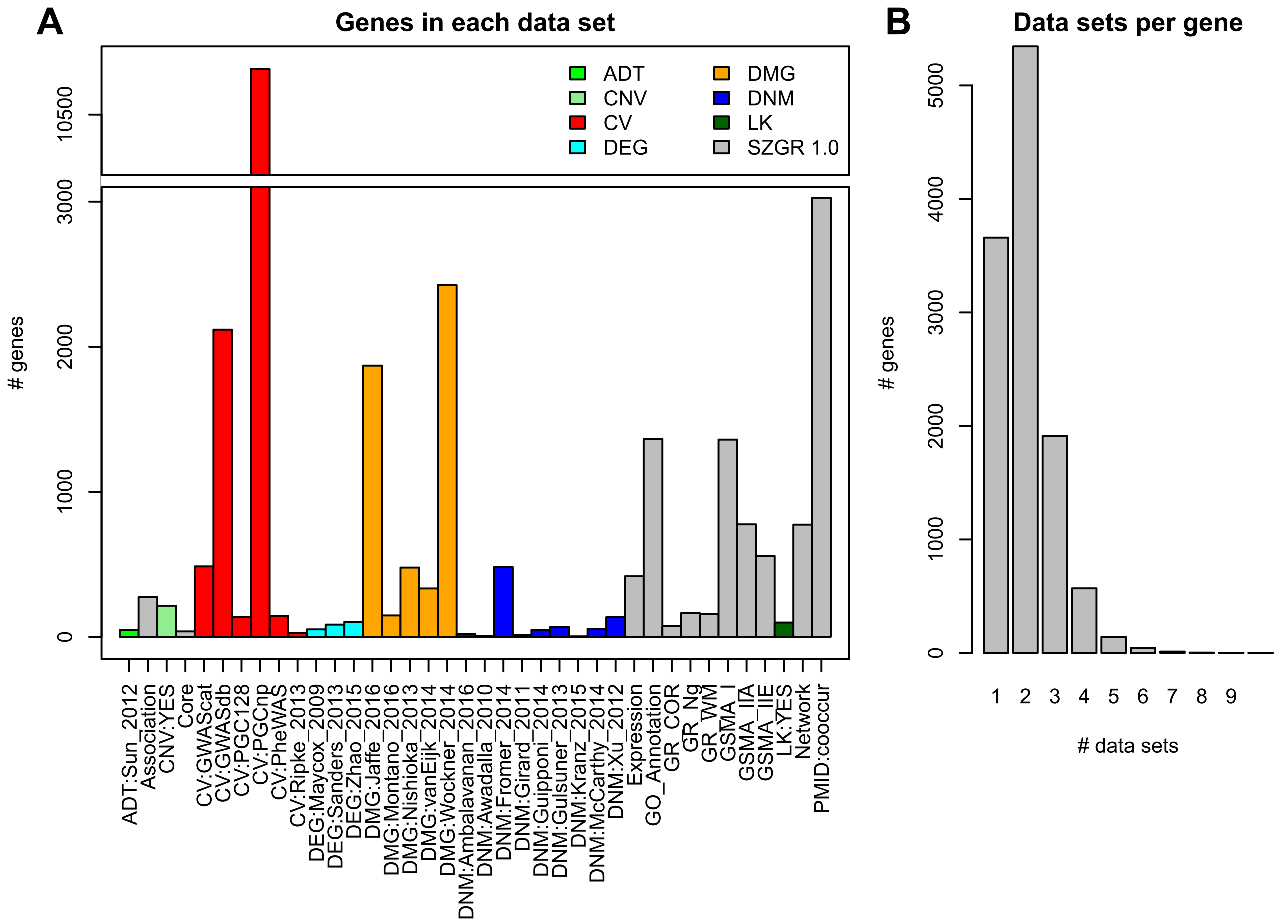

Figure 1. Overview of the datasets currently deposited in the SZGR 2.0 database. Datasets are organized by the format '*type:study*' (see details in the main text). (**A**) Distribution of genes in each dataset. The datasets with the same data type are labeled in the same color. (**B**) Histogram plot of the number of datasets per gene.

rect comparison of each gene's expression in brain with that in other tissues.

We collected three sets of meQTL data, two of which were performed using brain tissues (6,37), while the other one used the blood from schizophrenia patients (38). These datasets were used for annotations of both SNPs and genes, following the same strategy as we used for eQTL annotations.

We collected annotations of brain enhancers and promoters from the Roadmap Epigenomics project. BED files with coordinates for each region type were downloaded from http://egg2.wustl.edu/roadmap/data/byDataType/dnase/. Ten brain samples were obtained in this dataset, including both fetal and adult samples (39). These annotations were used for genetic variants and were displayed dynamically in the SNP page.

DATA INTEGRATION

Combining the genes from all sources with direct evidence resulted in a total of 11 736 genes (Figure 1A), each with at least one line of evidence supporting their association with schizophrenia. We collectively referred these genes as SZGR genes. There were 2367 genes harboring tier 1 SNPs in CV, 806 genes with non-silent DNMs, 215 genes overlapping with CNVs, 99 genes overlapping with linkage regions, 240 DEGs, 4651 DMGs, 3027 genes that had been studied in schizophrenia and 49 genes as targets of antipsychiatry drugs. The majority of SZGR genes (8025, 68.7%) were present in two or more datasets (Figure 1B), indicating that most genes included in our database had multiple lines of evidence supporting their association with schizophrenia.

Development related expression patterns of schizophrenia genes

There have been repeated reports that susceptibility genes with schizophrenia and other psychiatric diseases were preferentially differentially expressed during fetal life (40,41). Among the four gene expression datasets, DS2 and DS3 were brain expression data. We used these two datasets to explore gene expression patterns during developmental stages. With DS2, we calculated the log2 FC for each gene by comparing its fetal and postnatal expression (40). We then examined the log2 (FC) distribution of SZGR genes to determine if they were over-represented with genes upregulated or downregulated before and after birth. As

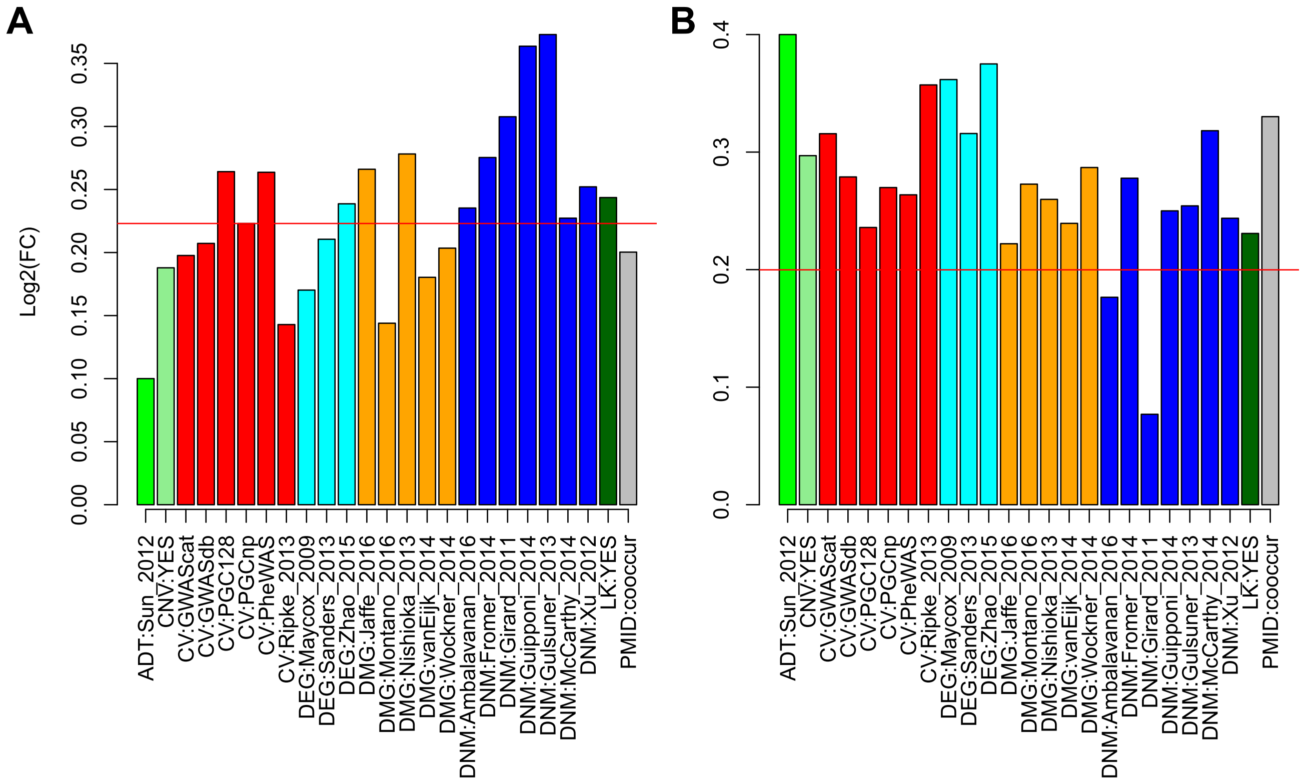

Figure 2. Overlap between schizophrenia genes and developmentally abnormal genes as determined by the log2 fold change (FC, see main text). (**A**) Distribution of overlapping genes between each dataset and genes that were downregulated during fetal life (log2 (FC) < -0.5). (**B**) Distribution of overlapping genes between each dataset and genes that were upregulated during fetal life (log2 (FC) > 0.5). The red line indicates the average proportion of genes downregulated in fetal life (A) or upregulated in fetal life (B).

shown in Figure 2, DEGs from three datasets tended to be over-represented with both upregulated genes and downregulated genes in fetal life compared to postnatal stage. Genes targeted by antipsychotic drugs were particularly over-represented in those downregulated genes in fetal life ($P = 1.86 \times 10^{-4}$). This is interesting because these genes, as the drug targets, are high confidence genes in schizophrenia. The observed over-representation supported the hypothesis that these drug target genes were not preferentially expressed during fetal life, consistent with previous reports that schizophrenia associated genes tended to be dysregulated in fetal life (40).

We used two genes as examples here: *RGS4* and *PLXNA2*. *RGS4* encodes a regulator of G-protein signaling 4. As shown by the GTEx data, it was mainly expressed in brain (Figure 3A). *RGS4* was down regulated during fetal life compared to postnatal life (DS2, Figure 3B). This down regulation pattern of *RGS4* in early developmental stages was also observed in the DS3 data in three brain regions (Figure 3C). Similarly, for the gene *PLXNA2*, although it was upregulated during fetal life (Figure 3D), its expression in adult brain (GTEx) was quite moderate (Figure 3E). Its upregulation was replicated in DS3 in four brain regions (Figure 3F).

Schizophrenia genes as the eQTL target genes

We next interrogated the proportion of SZGR genes that were also brain eGenes using brain eQTL data. As aforementioned, we collected a total of 12 brain eQTL datasets, including one with both *cis*- and *trans*-eQTLs and 11 with only *cis*-eQTLs. As shown in Figure 4, with a few exceptions, almost all datasets contained a proportion of genes that were also brain eGenes. As expected, the proportion of eGenes from the Myers' annotation in SZGR genes was the maximum, likely because both *cis*- and *trans*- eGenes were included in this annotation dataset. Among the eQTL annotations from GTEx for 10 brain regions, cerebellum had the highest proportion of eGenes included in our SZGR genes, followed by cerebellar hemisphere. This is interesting because cerebellum plays critical roles in cognitive function. For example, stimulation of cerebellar influences several regions in frontal cortex and impacts cognition in schizophrenia. The high proportion of our schizophrenia genes in cerebellum eQTL further confirmed the important role of cerebellum in schizophrenia.

DESCRIPTION OF THE WEBSITE AND THE TOOLS

In SZGR 2.0 database, genes and variants were firstly organized by the data type (genetics studies, epigenetics studies, transcriptomic studies and translational studies) and then by each specific dataset. We deposited all variants and genes for which there was good evidence to support their associa-

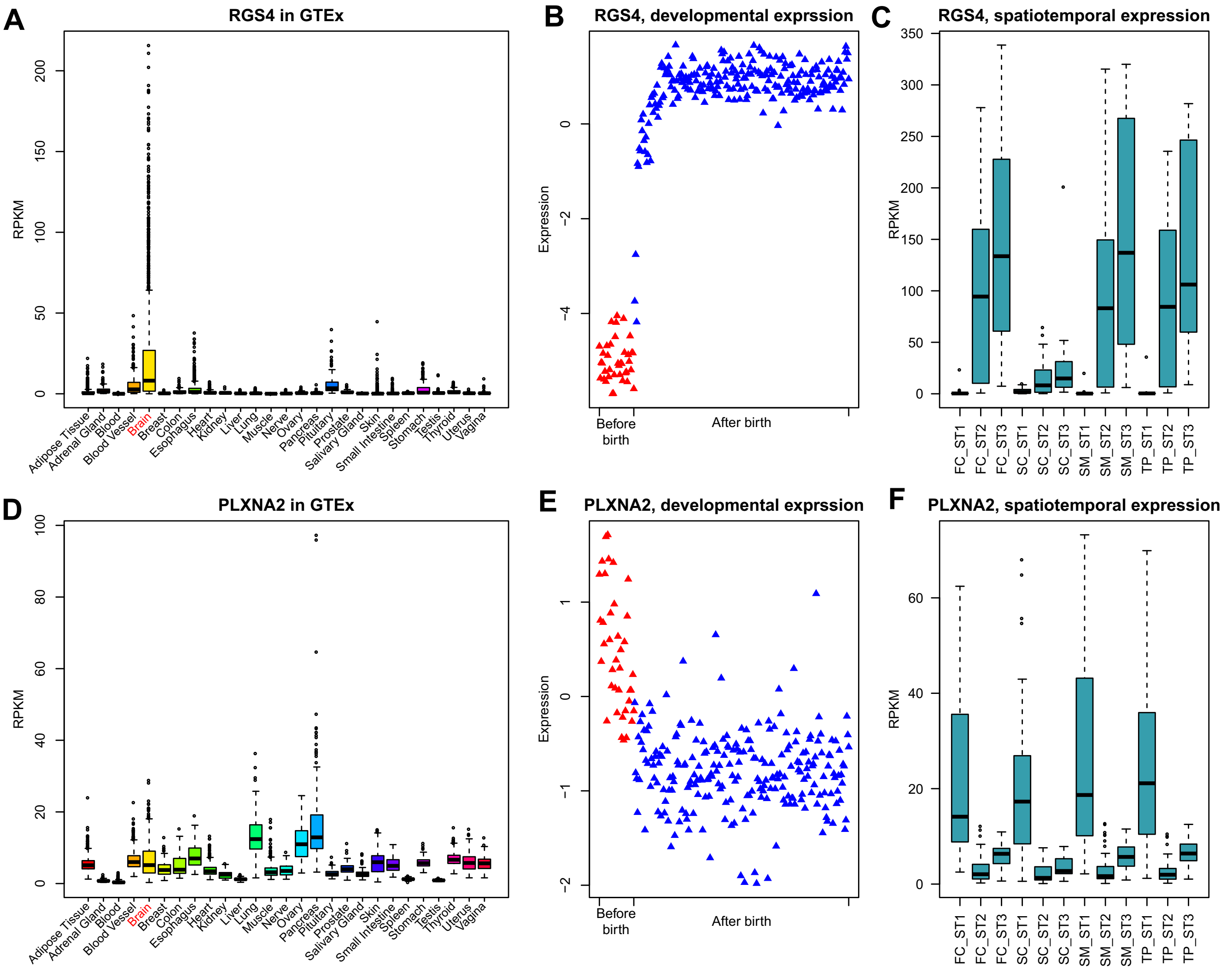

Figure 3. Demonstration of two example genes for their gene expression patterns. (**A**) Tissue-specific gene expression from GTEx data for *RGS4*. (**B**) Developmental expression pattern for *RGS4*. Red dots indicate the samples before birth and blue dots for the samples after birth. (**C**) Spatiotemporal expression pattern for *RGS4* in four brain regions (SC: sub-cortical regions; SM: sensory-motor regions; FC: frontal cortex; and TP: temporal-parietal cortex) and three developmental stages (ST1: 13–26 post-conception weeks, ST2: 4 months to 11 years and ST3: 13–23 years). (**D**) Tissue-specific gene expression from GTEx data for *PLXNA2*. (**E**) Developmental expression pattern for *PLXNA2*. Red dots indicate the samples before birth and blue dots for the samples after birth. (**F**) Spatiotemporal expression pattern for *PLXNA2* in four brain regions and three developmental stages.

tion with schizophrenia, and we built high schematic structures to facilitate easier access to every dataset. In general, we provided an overall page for each data type, including web pages for CV, CNV, DNM, DEG, DMG and ADT. In each of these pages, we provided summarized annotations and analysis results. For example, in each CV page, we mapped the variants to brain enhancers, promoters and genes, and presented figures and tables for the overall features of these variants. For user's convenience, the organization of datasets could be accessed from any page in the SZGR 2.0 web interface. Furthermore, we provided both raw records and our second-analysis results wherever available. Briefly, two major pages were developed to display information, one for genes and the other for SNPs (Supplementary Figures S2 and 3). The gene page holds general information, schizophrenia association evidence, raw and processed records for each line of evidence and comprehensive annotation that could be helpful for users to interpret the function of the gene. The SNP page is similarly designed, including both general information and schizophrenia association evidence. One advantage of the information in SNP page includes the brain-specific enhancer and promoter annotations. A detailed description of the website was provided in the Supplementary Data and was also available through the document page of SZGR.

We provided different browsing options (by data type, study and datasets) and search options (by SNPs, genes or regions) for all the data. Users can browse by data type and then by studies, using the aforementioned labels. For each dataset, we also provided both summary and initial analysis of the data, whenever applicable. For example, for the SNPs collected from GWAS Catalog (CV:GWAScat), we pro-

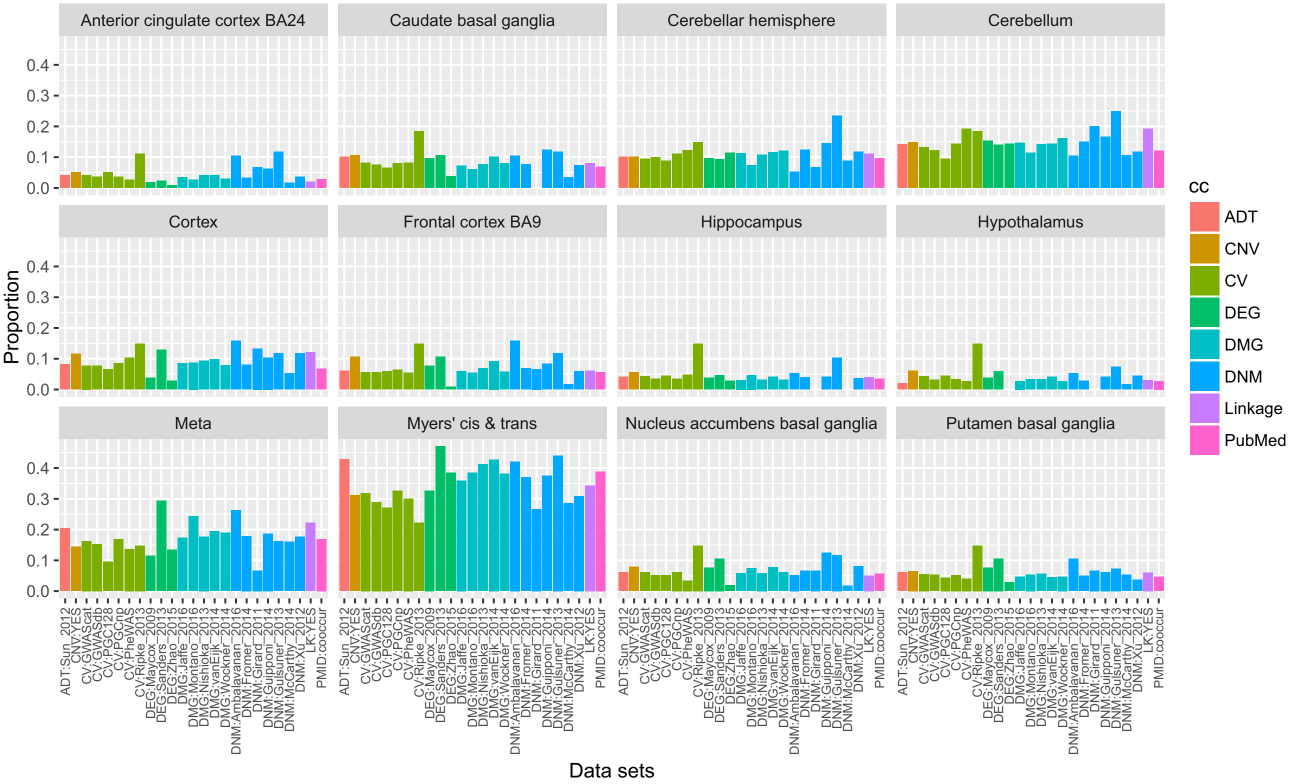

Figure 4. Overlap between schizophrenia genes and eGenes from 12 brain eQTL datasets. Here, eGenes are defined as genes with at least one significantly associated SNPs in the corresponding eQTL datasets.

vided detailed annotations (mapped genes, eSNP/meSNP relationship, functional impact, etc.), the distribution of their distances to the nearest promoters or enhancers in brain and a Circos plot showing their distribution in the human genome. Users can browse all datasets in the statistics page and find easy access to any dataset. In addition, we provided comprehensive searching functions, enabling users to search by SNP ID, gene name or genomic region.

DATABASE DESIGN AND UPDATES

All data in the new database was included in a MySQL database and the web interface was developed using JavaServer Pages (JSP). The web site was built on an Apache Tomcat server. This enables us to easily expand the database when new data types become available. As shown in the document page, new data can be easily added without disturbing existing datasets.

DISCUSSION

SZGR 2.0, to our knowledge, is the most comprehensive database dedicated to schizophrenia. After our substantial expansion from our SZGR 1.0 (https://bioinfo.uth.edu/SZGR1/), it was designed as a one-stop shop of schizophrenia variants and genes. By the time of this writing, the risk factors that have been reported for schizophrenia have all been screened and deposited in this new SZGR database, covering genetics studies, epigenetics studies, transcriptomic studies, translational studies and functional annotations. These curated datasets, as well as our numerous analysis results and plots, have all been deposited in SZGR 2.0 and made available to users in the research community.

Even with the massive amount of information, we expect a continuous rapid increase in data generation and improvement of data quality in schizophrenia research in the coming years. Our search for genomic studies is designed using automatic code for batch screening literatures in PubMed, followed by intensive manual check by experienced investigators. In future, we will update the database routinely (e.g. every quarter of a year). Our team has been active in schizophrenia research since 2003, enabling us to keep up with the rapidly growing schizophrenia research. With the new data and novel functions, SZGR is a dedicated repository, platform and communication warehouse for schizophrenia that aims for an enhanced understanding of the disease and an empowered facilitation of schizophrenia research.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Mr. Han Chen and Drs. Hua Sun, Yuan Cao and Pora Kim for valuable assistance in this project.

FUNDING

National Institutes of Health [R01LM011177, R21CA196508, in part]. Funding for open access charge: Dr. Doris L. Ross Professorship Funds to Dr. Zhao from the University of Texas Health Science Center at Houston.
Conflict of interest statement. None declared.

REFERENCES

1. Sullivan,P.F., Daly,M.J. and O'Donovan,M. (2012) Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat. Rev. Genet.*, **13**, 537–551.
2. Riley,B. and Kendler,K.S. (2006) Molecular genetic studies of schizophrenia. *Eur. J. Hum. Genet.*, **14**, 669–680.
3. Jia,P., Sun,J., Guo,A.Y. and Zhao,Z. (2010) SZGR: a comprehensive schizophrenia gene resource. *Mol. Psychiatry*, **15**, 453–462.
4. Colantuoni,C., Lipska,B.K., Ye,T., Hyde,T.M., Tao,R., Leek,J.T., Colantuoni,E.A., Elkahloun,A.G., Herman,M.M., Weinberger,D.R. *et al.* (2011) Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature*, **478**, 519–523.
5. Myers,A.J., Gibbs,J.R., Webster,J.A., Rohrer,K., Zhao,A., Marlowe,L., Kaleem,M., Leung,D., Bryden,L., Nath,P. *et al.* (2007) A survey of genetic human cortical gene expression. *Nat. Genet.*, **39**, 1494–1499.
6. Hannon,E., Spiers,H., Viana,J., Pidsley,R., Burrage,J., Murphy,T.M., Troakes,C., Turecki,G., O'Donovan,M.C., Schalkwyk,L.C. *et al.* (2016) Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat. Neurosci.*, **19**, 48–54.
7. Consortium,G.T. (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.
8. Roadmap Epigenomics Consortium, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
9. Zhao,Z., Xu,J., Chen,J., Kim,S., Reimers,M., Bacanu,S.A., Yu,H., Liu,C., Sun,J., Wang,Q. *et al.* (2015) Transcriptome sequencing and genome-wide association analyses reveal lysosomal function and actin cytoskeleton remodeling in schizophrenia and bipolar disorder. *Mol. Psychiatry*, **20**, 563–572.
10. Chang,X. and Wang,K. (2012) wANNOVAR: annotating genetic variants for personal genomes via the web. *J. Med. Genet.*, **49**, 433–436.
11. Welter,D., MacArthur,J., Morales,J., Burdett,T., Hall,P., Junkins,H., Klemm,A., Flicek,P., Manolio,T., Hindorff,L. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.
12. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 9362–9367.
13. Li,M.J., Liu,Z., Wang,P., Wong,M.P., Nelson,M.R., Kocher,J.P., Yeager,M., Sham,P.C., Chanock,S.J., Xia,Z. *et al.* (2016) GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.*, **44**, D869–D876.
14. Li,M.J., Wang,P., Liu,X., Lim,E.L., Wang,Z., Yeager,M., Wong,M.P., Sham,P.C., Chanock,S.J. and Wang,J. (2012) GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.*, **40**, D1047–D1054.
15. Denny,J.C., Ritchie,M.D., Basford,M.A., Pulley,J.M., Bastarache,L., Brown-Gentry,K., Wang,D., Masys,D.R., Roden,D.M. and Crawford,D.C. (2010) PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*, **26**, 1205–1210.
16. Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421–427.
17. Ripke,S., O'Dushlaine,C., Chambert,K., Moran,J.L., Kahler,A.K., Akterin,S., Bergen,S.E., Collins,A.L., Crowley,J.J., Fromer,M. *et al.* (2013) Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.*, **45**, 1150–1159.
18. Nicolae,D.L., Gamazon,E., Zhang,W., Duan,S., Dolan,M.E. and Cox,N.J. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.*, **6**, e1000888.
19. Jiang,J., Jia,P., Shen,B. and Zhao,Z. (2014) Top associated SNPs in prostate cancer are significantly enriched in cis-expression quantitative trait loci and at transcription factor binding sites. *Oncotarget*, **5**, 6168–6177.
20. Gamazon,E.R., Badner,J.A., Cheng,L., Zhang,C., Zhang,D., Cox,N.J., Gershon,E.S., Kelsoe,J.R., Greenwood,T.A., Nievergelt,C.M. *et al.* (2013) Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Mol. Psychiatry*, **18**, 340–346.
21. International Schizophrenia Consortium, Purcell,S.M., Wray,N.R., Stone,J.L., Visscher,P.M., O'Donovan,M.C., Sullivan,P.F. and Sklar,P. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**, 748–752.
22. Awadalla,P., Gauthier,J., Myers,R.A., Casals,F., Hamdan,F.F., Griffing,A.R., Cote,M., Henrion,E., Spiegelman,D., Tarabeux,J. *et al.* (2010) Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. *Am. J. Hum. Genet.*, **87**, 316–324.
23. Vacic,V., McCarthy,S., Malhotra,D., Murray,F., Chou,H.H., Peoples,A., Makarov,V., Yoon,S., Bhandari,A., Corominas,R. *et al.* (2011) Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. *Nature*, **471**, 499–503.
24. Levinson,D.F., Duan,J., Oh,S., Wang,K., Sanders,A.R., Shi,J., Zhang,N., Mowry,B.J., Olincy,A., Amin,F. *et al.* (2011) Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. *Am. J. Psychiatry*, **168**, 302–316.
25. Malhotra,D. and Sebat,J. (2012) CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell*, **148**, 1223–1241.
26. Rees,E., Walters,J.T., Georgieva,L., Isles,A.R., Chambert,K.D., Richards,A.L., Mahoney-Davies,G., Legge,S.E., Moran,J.L., McCarroll,S.A. *et al.* (2014) Analysis of copy number variations at 15 schizophrenia-associated loci. *Br. J. Psychiatry*, **204**, 108–114.
27. Ng,M.Y., Levinson,D.F., Faraone,S.V., Suarez,B.K., DeLisi,L.E., Arinami,T., Riley,B., Paunio,T., Pulver,A.E., Irmansyah *et al.* (2009) Meta-analysis of 32 genome-wide linkage studies of schizophrenia. *Mol. Psychiatry*, **14**, 774–785.
28. Xu,J., Sun,J., Chen,J., Wang,L., Li,A., Helm,M., Dubovsky,S.L., Bacanu,S.A., Zhao,Z. and Chen,X. (2012) RNA-Seq analysis implicates dysregulation of the immune system in schizophrenia. *BMC Genomics*, **13**(Suppl. 8), S2.
29. Maycox,P.R., Kelly,F., Taylor,A., Bates,S., Reid,J., Logendra,R., Barnes,M.R., Larminie,C., Jones,N., Lennon,M. *et al.* (2009) Analysis of gene expression in two large schizophrenia cohorts identifies multiple changes associated with nerve terminal function. *Mol. Psychiatry*, **14**, 1083–1094.
30. Sanders,A.R., Goring,H.H., Duan,J., Drigalenko,E.I., Moy,W., Freda,J., He,D., Shi,J., MGS and Gejman,P.V. (2013) Transcriptome study of differential expression in schizophrenia. *Hum. Mol. Genet.*, **22**, 5001–5014.
31. Duan,J., Sanders,A.R., Moy,W., Drigalenko,E.I., Brown,E.C., Freda,J., Leites,C., Goring,H.H., Gejman,P.V. and MGS (2015) Transcriptome outlier analysis implicates schizophrenia susceptibility genes and enriches putatively functional rare genetic variants. *Hum. Mol. Genet.*, **24**, 4674–4685.
32. Sun,J., Xu,H. and Zhao,Z. (2012) Network-assisted investigation of antipsychotic drugs and their targets. *Chem. Biodivers*, **9**, 900–910.
33. Xia,K., Shabalin,A.A., Huang,S., Madar,V., Zhou,Y.H., Wang,W., Zou,F., Sun,W., Sullivan,P.F. and Wright,F.A. (2012) seeQTL: a searchable database for human eQTLs. *Bioinformatics*, **28**, 451–452.
34. Kim,Y., Xia,K., Tao,R., Giusti-Rodriguez,P., Vladimirov,V., van den Oord,E. and Sullivan,P.F. (2014) A meta-analysis of gene expression quantitative trait loci in brain. *Transl. Psychiatry*, **4**, e459.
35. Gulsuner,S., Walsh,T., Watts,A.C., Lee,M.K., Thornton,A.M., Casadei,S., Rippey,C., Shahin,H., Consortium on the Genetics of Schizophrenia *et al.* (2013) Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell*, **154**, 518–529.
36. Miller,J.A., Ding,S.L., Sunkin,S.M., Smith,K.A., Ng,L., Szafer,A., Ebbert,A., Riley,Z.L., Royall,J.J., Aiona,K. *et al.* (2014)

Transcriptional landscape of the prenatal human brain. *Nature*, **508**, 199–206.

37. Jaffe,A.E., Gao,Y., Deep-Soboslay,A., Tao,R., Hyde,T.M., Weinberger,D.R. and Kleinman,J.E. (2016) Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat. Neurosci.*, **19**, 40–47.
38. van Eijk,K.R., de Jong,S., Strengman,E., Buizer-Voskamp,J.E., Kahn,R.S., Boks,M.P., Horvath,S. and Ophoff,R.A. (2015) Identification of schizophrenia-associated loci by combining DNA methylation and gene expression data from whole blood. *Eur. J. Hum. Genet.*, **23**, 1106–1110.
39. Marbach,D., Lamparter,D., Quon,G., Kellis,M., Kutalik,Z. and Bergmann,S. (2016) Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. Methods*, **13**, 366–370.
40. Birnbaum,R., Jaffe,A.E., Hyde,T.M., Kleinman,J.E. and Weinberger,D.R. (2014) Prenatal expression patterns of genes associated with neuropsychiatric disorders. *Am. J. Psychiatry*, **171**, 758–767.
41. Parikshak,N.N., Luo,R., Zhang,A., Won,H., Lowe,J.K., Chandran,V., Horvath,S. and Geschwind,D.H. (2013) Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell*, **155**, 1008–1021.