Enrichment of rare genetic variants in astrocyte gene enriched co-expression modules altered in postmortem brain samples of schizophrenia

J. González-Peñas, J. Costas, M.J. Ginzo-Villamayor and B. Xu

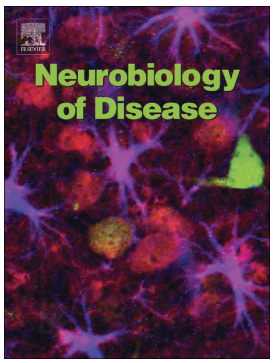
Version: Accepted manuscript

HOW TO CITE

González Peñas, J., Costas, J., Ginzo-Villamayor, M.J., Xu, B. (2019). Enrichment of rare genetic variants in astrocyte gene enriched co-expression modules altered in postmortem brain samples of schizophrenia Neurobiology of Disease. 121. pp. 305 - 314. Elsevier.

FUNDING

This work was partially supported by a National Alliance for Research in Schizophrenia and Depression (NARSAD) Young Investigator Award (to B.X.).



Enrichment of rare genetic variants in astrocyte gene enriched co-expression modules altered in postmortem brain samples of schizophrenia

Javier González-Peñas, Javier Costas, María José Ginzo Villamayor, Bin Xu

PII: S0969-9961(18)30729-0
DOI: doi:10.1016/j.nbd.2018.10.013
Reference: YNBDI 4306

To appear in: *Neurobiology of Disease*

Received date: 23 November 2017
Revised date: 27 September 2018
Accepted date: 17 October 2018

Please cite this article as: Javier González-Peñas, Javier Costas, María José Ginzo Villamayor, Bin Xu , Enrichment of rare genetic variants in astrocyte gene enriched co-expression modules altered in postmortem brain samples of schizophrenia. Ynbdi (2018), doi:10.1016/j.nbd.2018.10.013

TITLE

Enrichment of rare genetic variants in astrocyte gene enriched co-expression modules altered in postmortem brain samples of schizophrenia.

Author names and affiliations

Javier González-Peñas, PhD, Hospital Gregorio Marañón, Madrid, Spain. IiSGM, School of Medicine
Calle Dr Esquerdo, 46
Madrid, 28007
Phone 0034 662051044
Email: Javier.gonzalez@iisgm.com

Javier Costas, PhD, Instituto de Investigación Sanitaria de Santiago, Complexo Hospitalario Universitario de Santiago de Compostela, Servizo Galego de Saúde, Santiago de Compostela, Spain
Grupo de Xenética Psiquiátrica
Santiago de Compostela, A coruña (Spain)

María José Ginzo Villamayor, M.Sc. Departamento de estadística y análisis matemático, Universidad de Santiago de Compostela, Spain.
Santiago de Compostela, A coruña (Spain)

Bin Xu, PhD Department of Psychiatry, Columbia University, New York NY
Department of Psychiatry, Columbia University
1051 Riverside Drive Unit 25
New York, NY 10032
Phone 646 774-5259
Fax 646 774-8438
Email: bx2105@cumc.columbia.edu

Corresponding authors

Javier González-Peñas and Bin Xu.

Author Contributions

J.G. and B.X. designed research, with some contribution from J.C.; J.G. performed research and data analysis with some contribution from M.G.; and J.G. and B.X. wrote the paper.

Keywords

RNA-seq, postmortem brain, schizophrenia, rare mutations, co-expression modules, astrocytes.

ABSTRACT

The transcriptome profiles of the cingulate gyrus region from the postmortem brain tissues of a set of well-characterized patients with schizophrenia (SCZ) and matched controls were investigated using an integrated approach that analyzed both the alterations in transcription expression pattern and rare genetic variants in expressed genes. We demonstrated increased expression of astrocyte-related genes using spatiotemporal co-expression modules that have previously been established for developing human brain, and showed these results are independent of medication dosage. The relationship between genetic variants and expression pattern in the context of neurodevelopment was further investigated, and we identified an enrichment of rare genetic variants in a set of signature genes that were specific to astrocytes and up-regulated in the patients with SCZ. Our result suggested the involvement of astrocyte malfunction in SCZ pathophysiology. In addition, our approach indicated a novel strategy of narrowing down genetic variants that might contribute to the pathophysiology in the patients with SCZ to a subset of genes that are highly expressed in an affected brain region.

INTRODUCTION

Schizophrenia (SCZ) is a complex disease, affecting 1% of the population with devastating consequences for sufferers, their families and public health. Rare genetic variants with high penetrance, such as single nucleotide variants (SNVs), small insertions/deletions (indels) and even larger copy-number variants (CNVs) have been shown to be collectively enriched in SCZ susceptibility by recent human genetic studies (Kirov et al., 2012; Purcell et al., 2014; Rees et al., 2013; Xu et al., 2008; Xu et al., 2012). However, the contribution of individual variants to the pathophysiology of SCZ is hard to evaluate from genomic studies because a large number of neutral variants exist in the human genome, decrease the signal/noise ratio and prevent the true disease causing variants from being identified with sufficient statistical power. In addition, the genetic variation associated with mental disorders so far has largely been derived from genomic DNA using patients' blood cells. Recently, evidence from various studies demonstrate that somatic mutations and mutations in the genes that control epigenetic mechanisms in affected brain regions could have important contribution to the etiology of neuropsychiatric disorders given the exuberant proliferation of cortical precursors during fetal development (Insel, 2014; Takata et al., 2014; Tsankova et al., 2007).

These observations led us to search for more effective approaches to enrich the potential rare disease causing variants. We reasoned that it would be more informative and powerful if we focus on searching for the rare variants in genes that

express in the affected brain regions and have altered expression pattern in SCZ patients. In addition, rare variants that are enriched in certain co-expression modules that contain the defect genes are more likely to contribute to the disease pathophysiology in the context of neurodevelopment. In this regard, recently developed RNA sequencing (RNA-seq) technique provides not only the greater linear range of detection or lower rates of false negatives and false positives than outperforms classical microarray data (Mortazavi et al., 2008; Wilhelm et al., 2008), but more importantly it provides single nucleotide resolution so that alterations in each expressed transcript can be unequivocally determined. Therefore, information carried by RNA-seq data at both genomic and transcriptional levels can be integrated to gain important insight into the contribution of rare protein-coding variants in the genetic architecture and related molecular pathophysiology of SCZ and isolate true causative variants with high penetrance from the irrelevant neutral variants.

In this study, we developed such an integrated approach to analyze the RNA-seq data of the postmortem brain cingulate gyrus tissues from a set of well-characterized SCZ patients and matched controls (CO) and to explore the relationship between rare genetic variants and altered gene expression pattern in the context of neurodevelopment in SCZ patients (**supplementary figure 1)**. The cingulate gyrus is an integral part of the limbic system involved in emotion formation and processing (Hadland et al., 2003) learning and memory (Sutherland et al., 1988). Its importance in psychiatric disorders (Drevets et al., 2008) is highlighted by abnormalities in cingulate gyrus volume (Costain et al., 2010; Takahashi et al., 2003), metabolism (Haznedar et al., 2004), connectivity (Wang et al., 2015) and astrocyte expression (Katsel et al., 2011) that have been repeatedly associated with the pathophysiology of SCZ. In this study, we utilized a gene expression analysis pipeline to determine the genes dysregulated in SCZ, and a variant detection pipeline in parallel to detect rare variation existing in expressed transcripts. We further evaluated the enrichment of rare variants to the genes that have altered expression pattern observed in the patients' transcriptome (**supplementary figure 1**). Our analysis of relationship between variants distribution in different gene set categories pinpointed several candidate gene sets that provides important new insights regarding the underlying disease pathophysiology.

METHODS

Human brain tissue samples

Human cingulate gyrus *post-mortem* samples from 31 SCZ patients and 26 healthy individuals were obtained from Stanley Medical Research Institute's Array Collection database (http://www.stanleyresearch.org/brain-research/array-collection/). Main demographic characteristics are described in **table1**. Extended individual information by sample was also collected (**supplementary table 1).** SCZ Diagnoses were made using DSM-IV criteria. Neither significant difference in age nor in gender proportions were found between cases and controls (**supplementary figure 2**).

	Control (N=26)	Schizophrenia (N=31)	P-value
Age (years) (mean ± SD) [a]	44.6 ± 7.1	42.4 ± 8.8	0.457
Postmortem interval (hours) [a]	31.3 ± 12.7	29.7 ± 15.4	0.794
Brain pH [a]	6.5 ± 0.2	6.6 ± 0.2	**0.014**
Male/female proportions [b]	22:4	23:8	0.516
Duration of Illness	-	21.8 + 10.1	-
Lifetime dose of antipsychotics (g) [c]	-	78.29 + 93.6	-

Table 1. Demographic Characteristics of Control and Schizophrenia brain samples. For p-value calculations, Mann-Whitney (a) or two-tailed Fisher exact test (b) were used. Lifetime dose of antipsychotics are represented as fluphenazine equivalents (c). Bold values indicate significant P-values.

RNA-seq data of postmortem brain samples of SCZ and CO

RNA-seq data (raw *fastq* files) were obtained from the cingulate gyrus postmortem tissues of 31 SCZ patients and 26 healthy individuals. 75 bp pair-end sequencing was performed on Illumina Hiseq platform for both schizophrenic and unaffected sample set; two sequencing orphaned files were obtained for each sample analyzed (by default *_1.fastq* and *_2.fastq*).

Differential expression analysis between SCZ and CO

STAR aligner v2.3 (Dobin et al., 2013) was used to map reads from Illumina Genome Analyzer to the Human genome assembly hg19 following STAR 2-pass approach default parameters recommended by the Broad Institute RNA-seq best practices (https://software.broadinstitute.org/gatk/best-practices/). Output SAM files were first sorted with samtools v.1.2 (Li et al., 2009) and then imported into *htseq-count*, a tool developed in HTSeq package v 0.6.1 (Anders at al., 2014), to preprocess RNA-Seq data by counting the overlap of reads with genes. The whole counts vector files obtained for each sample were subsequently merged into a final matrix with the information of all read counts per gene and per sample. Gencode v19 comprehensive annotation file obtained from the UCSC Genome Browser was used to annotate the genes and isoforms. Overlapping isoforms of the same gene or different genes were merged into a unique locus using *cufmerge* tool from the cufflinks package (Trapnell et al., 2012). Throughout this paper, we used the term “genes” instead of “*loci*” being aware that in some cases more than one gene is collapsed into the same unit. A total of 48099 genes were annotated, and gene counts were converted to RPKM values. We performed several filters to remove possible artifacts from our data. Firstly, genes with average reads per kilobase per million mapped reads (RPKM) < 0.3 in the whole expression dataset were removed as it was defined as a reliable cutoff to distinguish expression from noise (Ramsköld et al., 2009). This filter reduced the expressed dataset from 48099 to 22386 genes with detectable expression in our samples. Secondly, to avoid introducing bias due to read depth variability, we search for samples

not reaching RPKM > 0.3 cutoff in at least 90% of genes, and detected 2 SCZ and 9 CO samples with marked low coverage (**supplementary figure 3**). PCA was then performed on the filtered expression dataset (22386 genes – all samples) to ensure low coverage samples were not separated from the rest (**supplementary figure 4**). Low and normal coverage samples are similarly distributed when representing PC1 vs PC2 (both encompassing 84.2% of the variability). Indeed, a cross-validation procedure was also performed. Most variance (71.6%) was explained by first principal component (**supplementary figure 4**). Thus, eigenvalues vector for PC1 was then created excluding one different sample each time, and performing PCA on the remaining samples. Boxplot diagram was constructed to detect any potential outlier within eigenvalues distribution. No outlier was found across our data (**supplementary figure 5**). Therefore, we did not exclude any sample for the gene expression analysis.
The filtered raw gene counts were used by both EdgeR v.3.13.0 (Robinson et al., 2010) and DESeq v.1.22.0 (Anders et al., 2010) R packages to determine the differentially expressed (DE) genes between SCZ and CO. The overlapped genes with adjusted P values < 0.05 from both DESeq and EdgeR were used to generate a list of 1876 DE genes (**supplementary figure 1**). To ensure our method is robust, whole expression analysis was repeated removing the 2 SCZ and 9 CO samples with low coverage, obtaining similar results (**results**).

Correlation analysis of gene expression pattern with lifetime medication exposure in the patient samples

Influence of lifetime medication exposure on gene expression could be a confounding factor for clarifying the relationship between genetic variants and altered expression pattern due to the natural consequence of the disease. As the lifetime medication exposure was publicly available as part of samples information provided by Stanley Foundation, we calculated Pearson correlations between gene expression and lifetime medication for the 31 SCZ samples and 26 CO, for each of the 22386 genes with detectable expression (RPKM > 0.3). P values for each correlation were calculated using 500 bootstrap replicates relying on random sampling with replacement (Efron and Tibshirani, 1993).

We used Benjamini-Yekutieli FDR (BY-FDR, Benjamini et al., 2001) procedure for multiple testing corrections for this correlation analysis, as this estimator is one improved from Benjamini-Hochberg FDR (BH-FDR, Benjamini et al., 1995) under some forms of dependence. We set, for a given gene, BY-FDR p-value > 0.05 as non-correlated with medication. In this study, we considered all genes with detectable expression (N=22386) and the subset of medication-independent genes (2104 genes with BY-FDR p-value > 0.05). Similarly, when only analyzing DE genes, we considered all DE genes (N=1876) and the subset of medication-independent DE genes (N=174).

Gene Set Enrichment Analysis (GSEA)

We performed GSEA (Subramanian et al., 2005) to check the enrichment in collected gene sets of co-expression modules recently published (Hawrylycz et al., 2012; Kang et al., 2011; Miller et al., 2014). Firstly, Kang et al. identified 29 co-expression modules (we named K1 to K29 here in this text) from the study of 57 brain samples across 16

distinct brain regions, mainly neocortical, spanning 15 periods from embryonic development to late adulthood (from 5.7 post-conceptual weeks (pcw) to 82 years old). Secondly, Hawrylycz et al. identified 13 co-expression modules (M1 to M13) from exhaustive transcriptomic study of two adult brains (24 and 39 years old) assaying more than three-hundred distinct structures at least once in both brains (Hawrylycz et al., 2012). Thirdly, Miller et al. identified 42 distinct modules from developmental transcriptome (C1 to C42) of four prenatal brains (two from 15 and 16 pcw and two from 21 pcw) across 25 areas of the developing neocortex. We also check the enrichment in cell-type enriched gene signatures from previous publications (Cahoy et al., 2008; Zhang et al., 2014). Brain cell-type specific signatures were defined by the genes whose expression levels are at least 1.5-fold higher in a specific cell type than the average of all other cell types (Zhang et al., 2014). We used cell type specific signatures derived from microarray data of developing and mature mouse forebrains (Cahoy et al., 2008), in which the significance was determined with a FDR p value < 1%, and, subsequently, we confirmed our findings with the cell-type specific signatures derived from the RNA-seq analysis of mouse cerebral cortex (Zhang et al., 2014). For each test, expression dataset (N=22386) was used, a minimum of 10 genes were required for a given gene set to perform the GSEA analysis, 1000 permutations, *weighted* statistic enrichment and *signal2noise* metric were used for ranking genes. The default FDR adjusted p-values were used to assess enrichment significance.

Cell-type Specific Expression Analysis (CSEA)

We also analyzed cell-specific co-expression enrichment of DE genes with CSEA (Dougherty et al., 2010, Xu et al., 2014) using the available online tool (http://genetics.wustl.edu/jdlab/csea-tool-2/), which tests gene cell-type specific and spatiotemporal enrichment across development with Brainspan data (http://www.brainspan.org/). To test for association in any cell type of neurodevelopmental state, hypergeometric tests are performed, and the default BH-FDR procedure is used to correct for multiple tests (Benjamini et al., 1995). We performed CSEA on DE genes up-regulated and down-regulated in SCZ, separately, as well as using medication-independent DE genes.

RNA-seq variant identification pipeline

Briefly, aligment of reads were carried out with GATK according to the best practices for variant calling on RNAseq, following by filtering of variants with low depth of coverage (DP<10), within unambiguously expressed genes (RPKM > 0.3) and MAF > 0.01 in 1KG, ExAC or ESP.

Specifically, raw alignment files were treated with Picard Tools (http://broadinstitute.github.io/picard/) version 1.109 to add read groups information, sort files, mark duplicated reads and index the output BAM file. These steps were done with both *AddOrReplaceReadGroups.jar* and *MarkDuplicates.jar* arguments. Marked BAM files were then subjected to a Genome Analysis Toolkit (GATK) pipeline (McKenna et al., 2010) according to the instruction of The GATK Best Practices for variant calling on RNAseq (http://gatkforums.broadinstitute.org/discussion/3892/the-gatk-best-practices-for-variant-calling-on-rnaseq-in-full-detail). We applied an additional new GATK tool called *SplitNCigarReads*, specifically developed for RNA-seq analysis,

and necessary to correctly handle splice junctions. Next, split BAM files were subject to local realignment, base-score recalibration, and variant calling with the *IndelRealigner*, *BaseRecalibration*, and *HaplotypeCaller* tools from GATK. dbSNP138 and Mills and 1000G gold standard indels databases were used as "known sites" to recalibrate quality values in every BAM files. *HaplotypeCaller* (*HC*) was used as variant calling algorithm instead of *UnifiedGenotyper*(*UG*), and *recoverDanglingHeads* and *dontUseSoftClippedBases* arguments were used to minimize false positive and false negative calls. The minimum Phred-scaled confidence threshold for calling variants was lowered from the default value 30 to 20, following the Broad Institute criteria for RNA-seq data. VCF files were obtained from the variant calling procedure and subsequently filtered on Fisher Strand values (FS > 30) and Quality by Depth values (QD < 2.0). (**Supplementary figure 6**).

Next, to improve the quality of variant detection, we used a positional cutoff of minimum Depth of coverage (DP) in every sample to exclude the positions with lower depth of coverage, and only select those representing genomic positions expressed in all samples. Thus, we only preserved genome positions with at least ten reads (DP > 10) at that position in all samples. To achieve this goal, we used the *Genomecov* tool from Bedtools v 2.25.0 (Quinlan et al., 2014) to calculate coverage in processed BAM files. Gencode v19 comprehensive annotation merged with *cufmerge* was used, as in the expression analysis. Finally, *IntersectBED* program from Bedtools was used to extract the genes that satisfied our coverage cutoff criteria. Functional annotation of identified variants was performed with ANNOVAR (Wang et al., 2010) and AVIA tool (Vuong et al., 2015). As we focused on rare variants, all variants with a global frequency higher than 0.01 in any of three databases (1000 genomes phase 3 release (1KG), Exome sequencing project (ESP) v2 and Exome Aggregation Consortium v 0.3 (ExAC)) were removed. Moreover, as some of the detected variants could be RNA editing sites (RES) instead of real genomic variants, we interrogated two well-validated and recently published RNA editing databases, REDIportal (Picardi et al., 2016) and SPRINT (Zhang et al., 2017) to additionally filter detected variants. Both databases profiled RNA editing in human tissues, including brain samples.

Variant-expression analyses

In order to study variants in relation with expression profiles, only variation across genes with detectable expression (RPKM > 0.3) was gathered. As we noticed that 2 SCZ and 9 CO had markedly less depth of coverage and do not pass sample cutoff (**supplementary fig 5**), but expression results were consistent after removing these samples, we studied genetic variants using filtered dataset of 29 SCZ and 17 CO. This way, despite reducing our sample size, we increased number of genomic positions reaching DP > 10 in every sample. We finally tested enrichment of rare singletons across co-expression modules significantly up or down-regulated in SCZ (Hawrylycz et al., 2012; Kang et al., 2011; Miller et al., 2014) (**Supplementary Figure 6**). For variant enrichment analyses, two-tailed Fisher exact test was performed between SCZ and CO in any given gene set.

RESULTS

Differentially expressed genes in the cingulate gyrus region of the post-mortem brain tissues from the patients with SCZ

We analyzed the differentially expressed gene pattern from the RNA-seq data of 57 postmortem cingulate gyrus samples including 31 SCZ and 26 CO generated by the Stanley Research Foundation Array Collection (**supplementary figure 1, see Methods**). No potential sample outliers were found after performing a PCA cross-validation procedure (**Methods, supplementary figure 4**). A total of 22386 Gencode genes ("whole genes") showed detectable expression (RPKM > 0.3). From the whole gene set, 1876 were differentially expressed ("DE genes") (FDR $p < 0.05$, see **Methods**). 1133 out of 1876 (60.4 %) genes had increased expression in patients compared to CO ("up-regulated DE genes"), while the remaining 743 (39.6 %) were down-regulated ("down-regulated DE genes") in SCZ (**supplementary table 2, supplementary figure 7**).

Given the demonstrated influence of medication throughout life in gene expression (Crespo-Facorro et al, 2015) in patients with SCZ, we also established Pearson correlations between gene expression levels and lifelong medication dosages in the patients (http://www.stanleyresearch.org/brain-research/array-collection/). Genes from "whole genes" were then classified as "medication-dependent genes or "medication-independent genes" based on correlation significance values using BY-FDR threshold of 0.05. Of "All genes", there are 2104 "medication-independent genes" and 20282 "medication-dependent genes". Interestingly, 174 of 2104 "medication-independent genes" were DE between SCZ and CO ("medication-independent DE genes"); of which 79 were up-regulated and 95 down-regulated ("medication-independent up-regulated genes" and "medication-independent down-regulated genes", respectively, **supplementary table 3)**. We used independently "whole genes", "medication-independent genes" or "medication-dependent genes" as background in the GSEA analysis, while we focus on analyzing the corresponding DE genes for enrichment analyses henceforth throughout this manuscript**.**

Medication-independent up-regulated genes are enriched in a co-expression module with astrocyte signature

Taking into consideration the neurodevelopmental nature of SCZ (Najas-García et al., 2014, Rapoport et al., 2012), we explored the enrichment of DE genes in various spatiotemporal co-expression modules previously defined in several transcriptome studies on developing human brain (Hawrylycz et al., 2012; Kang et al., 2011; Miller et al., 2014). Gene Set Expression Analysis (GSEA) is a well-established tool that is good at detecting small and cumulative effects in altered expression pattern (Subramanian et al., 2005). A total of 84 co-expression modules that were from three published studies and covered a series of embryonic development to late adulthood time periods and various distinct brain areas (see **Methods**) were used in our analysis. We first tested if there is any enrichment of the DE genes between SCZ and CO within these 84 distinct modules using whole genes (N=22386), medication-dependent genes (N= 20282), and medication-independent genes (N=2104) as background. 77, 68 and 35 modules passed size filters criteria for "whole genes", "medication-dependent genes" and "medication-independent genes", respectively and were used for GSEA analysis (**supplementary table 4**). When "whole genes" were considered, "up-regulated genes"

were significantly enriched in 5 modules including M10, C38, C36, K1 and K18; while "down-regulated genes" were enriched in 2 modules (M1 and K15) (FDR q<0.05, f**igure 1, table 2, supplementary table 5, supplementary figure 8-9**). There are several modules such as the module C38, which was defined as an astrocyte gene set module (Cahoy et al., 2008); M10, with enrichment in astrocyte genes (Miller et al., 2010); C36, with enrichment in astrocyte probable genes (Cahoy et al., 2008); In addition, module M1 is enriched in genes related to parvalbumin interneurons (Miller et al., 2010); module K1 is associated with cell cycle related functionality (Kang et al., 2011), K15 is enriched in synaptic transmission functionality (Kang et al., 2011), and K18 is enriched in cell morphogenesis functionality (Kang et al., 2011). When only "medication-independent genes" background (N=2104) were considered, "medication-independent up-regulated genes" were significantly enriched in only one module, C38 from Miller et al (N_{C38}=37; NES=1.92; BH-FDR-q=0.014, **figure 1-2, table 2, supplementary figure 10**). These results suggested that there is a significant enrichment of astrocyte signature in SCZ, which might be involved in the pathophysiology of the disease.

Module	Module Reference	Brain type	Status in SCZ	Genes in module	Genes in expression dataset	NES	FDR-q	Module Function
(A) All genes expressed (22386 genes)								
K15	Kang et al.	Development	downregulated	309	241	-1.881	0.031	Synaptic transmission
K1	Kang et al.	Development	upregulated	336	192	1.834	0.033	Cell cycle
M10	Hawrylycz et al.	Adult	upregulated	330	244	1.881	0.039	Metallothionein
C36	Miller et al.	Prenatal	upregulated	169	130	1.835	0.04	Lipid metabolism
K18	Kang et al.	Development	upregulated	86	72	1.912	0.042	Cell morphogenesis
M1	Hawrylycz et al.	Adult	downregulated	117	84	-1.904	0.048	Ion transport
C38	Miller et al.	Prenatal	upregulated	628	384	1.914	0.049	Cell cycle
(B) Medication-independent genes expressed 2104 genes)								
C38	Miller et al.	Prenatal	upregulated	628	37	1.922	0.014	Cell cycle

Table 2. Co-expression modules are enriched in the whole expression dataset (A) and the medication-independent expression dataset (B). Modules with significant enrichment after corrections for multiple tests (FDR-q < 0.05) are shown in the table. The original study they come from, brain type analyzed, status (downregulation/upregulation), genes expressed, Normalized Enrichment Score (NES) and main function described on original publications are also described.

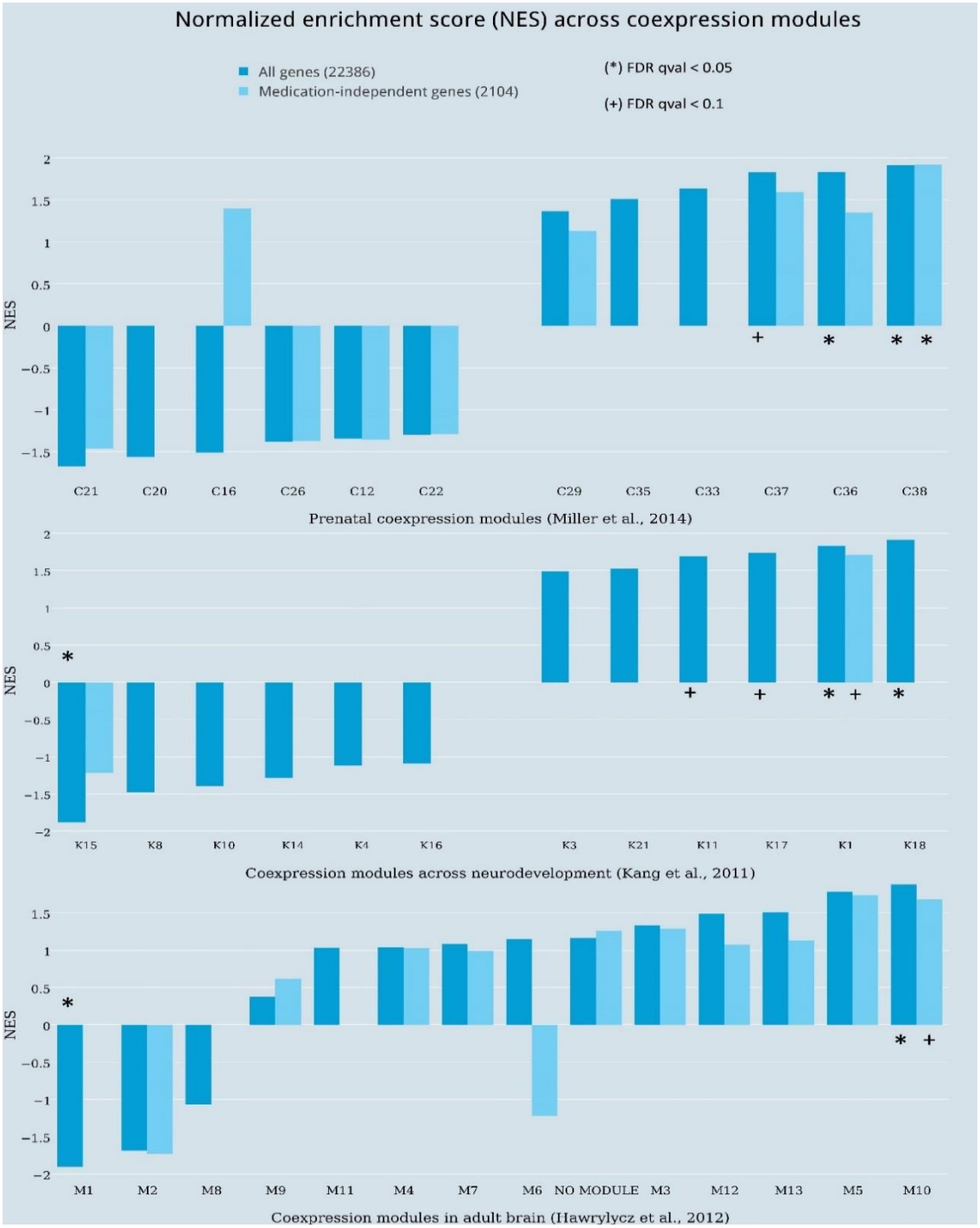


Figure 1. GSEA expression enrichment across spatiotemporal co-expression modules (Hawrylycz et al., 2012; Kang et al., 2011; Miller et al., 2014) within all genes with detectable expression (N = 22386), and medication-independent genes (N= 2104). Normalized enrichment score (NES) is represented in the figure, with negative NES values describing down-regulated modules in SCZ and positive NES values describing up-regulated modules in SCZ. Modules significantly enriched are marked in the figure (When hypergeometric test is significant FDRq-val < 0.05 is marked with "*", when FDRq-val < 0.1 is marked with "+"). All enrichment values, P values and numbers of genes are described in supplementary table 5A. Modules with less than 10 genes were filtered out and not analyzed (methods, **supplementary table 4**).

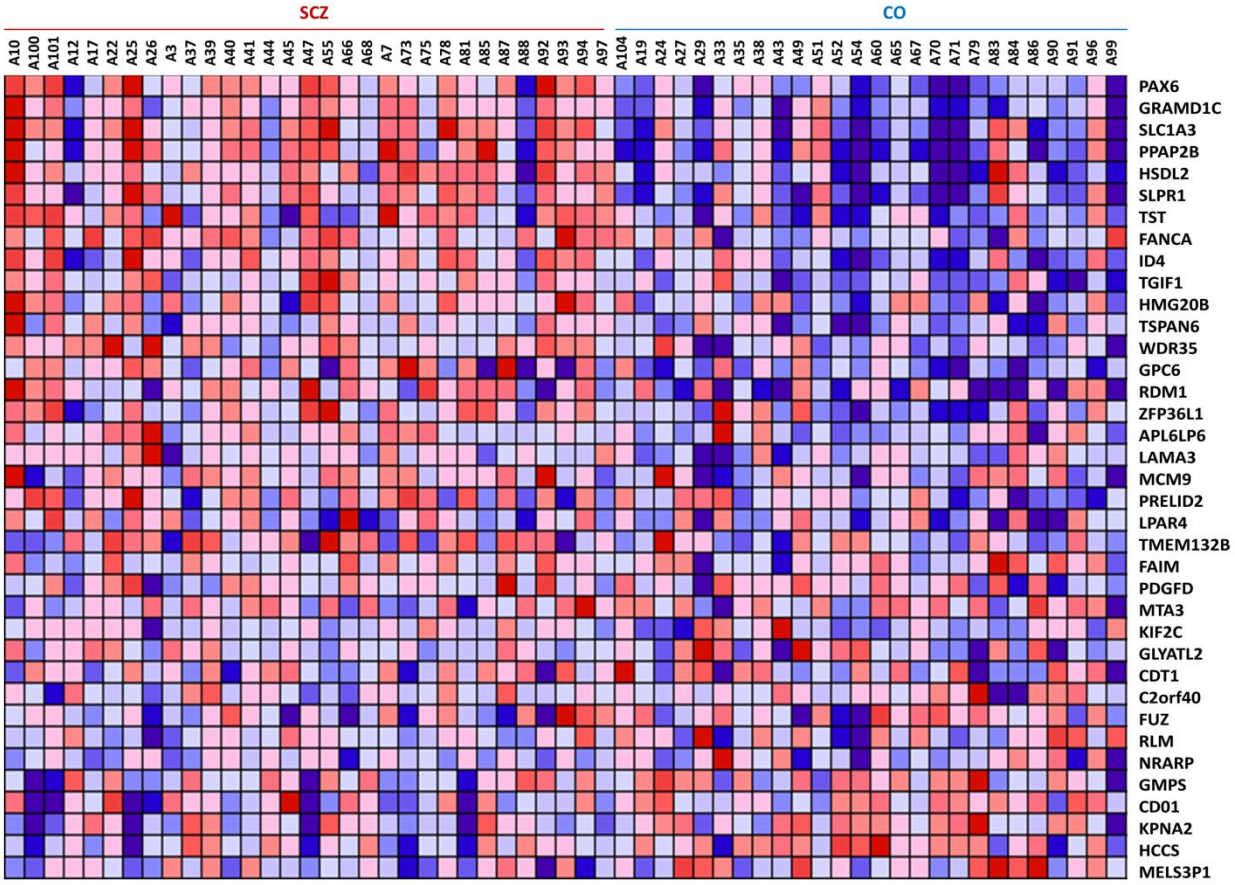


Figure 2. Heatmap of expression measures (RPKM) for the 37 genes that overlapped between module C38 (total # of genes = 628) and medication-independent gene set (total # of genes =2104). Expression measures from DESeq we used to compare RPKM values between 31 SCZ samples (left-grey colored) and 26 CO samples (right-yelow colored).

In addition to PCA analysis and gene filtering (RPKM > 0.3) steps, to assess the robustness of our developed expression analysis methodology, we removed 9 CO and 2 SCZ samples with low DP and repeated GSEA analyses to test enrichment of DE genes within co-expression modules using both whole genes (N=22386) and medication-independent genes (N=2104) as background. In both cases, we observed significant results similar to what happened in expression pipeline using all samples(**supplementary table 6**). Interestingly, when only medication-independent genes were considered, medication-independent up-regulated genes were again significantly enriched in C38 module from Miller et al (N_{C38}=37; NES=1.80; BH-FDR-q=0.038**, supplementary table 6**). Thus, we verified robustness of our methodology, and confirmed expression results do not depend on sample coverage differences.

Cell-type Specific Expression Analysis shows alterations in cortical astrocyte signature in SCZ

The finding that the “medication-independent genes” were enriched in an astrocyte related module promoted us to further investigate cell-type specific changes in our general GSEA analysis; we further performed cell-type signature enrichment analysis

with GSEA tool using brain cell type signatures defined by previous experiments. We first analyzed enrichment using cell-type specific signature from one microarray data study (Cahoy et al., 2008). The cell-type specific GSEA analysis confirmed the "medication-independent up-regulated genes" were enriched in astrocyte signature defined by the microarray study (N=156 genes; NES=1.71; FDR q-val=0.037). Independently, we used the cell type signatures derived from an RNA-seq analysis of mouse cerebral cortex (Zhang et al., 2014) and further confirmed the enrichment astrocyte signature defined by the RNA-seq study under "medication-independent genes" background (N=279 genes; NES=1.8; FDR q-val=0.025; **supplementary table 7, supplementary figures 11-14**).

We confirmed this finding in a more systematic approach by utilizing the Cell-type Specific Expression Analysis (CSEA) tool, which are derived cell type signature using TRAP data (Xu et al., 2014). CSEA is an independent tool that based on experimentally affinity purification of the complete suite of translating mRNA from genetically labeled cell populations and a statistical integration of multiple specificity index probability (pSI) thresholds to determine overrepresentation of a putative list of genes in any given cell type and developmental state in central nervous system (Dougherty et al., 2010). As expected, the "medication-independent up-regulated genes" were significantly enriched in cortical astrocytes across several pSI thresholds (the lowest FDR q-val=8.01 x 10^{-4} at pSI threshold 0.01), and, less significantly in, cerebellum astrocytes (the lowest FDR q-val=0.035 at pSI threshold 0.01) and other glial cell types (the lowest FDR q-val=0.007at pSI threshold 0.05) (**figure 3, supplementary table 8**). This highly significant enrichment confirmed previous GSEA findings and suggested that alterations in astrocyte related genes were associated with the disease condition. In contrast, the "medication-independent down-regulated genes" were not found to be enriched in any specific cell types (**supplementary figure 12**, **supplementary table 8**).

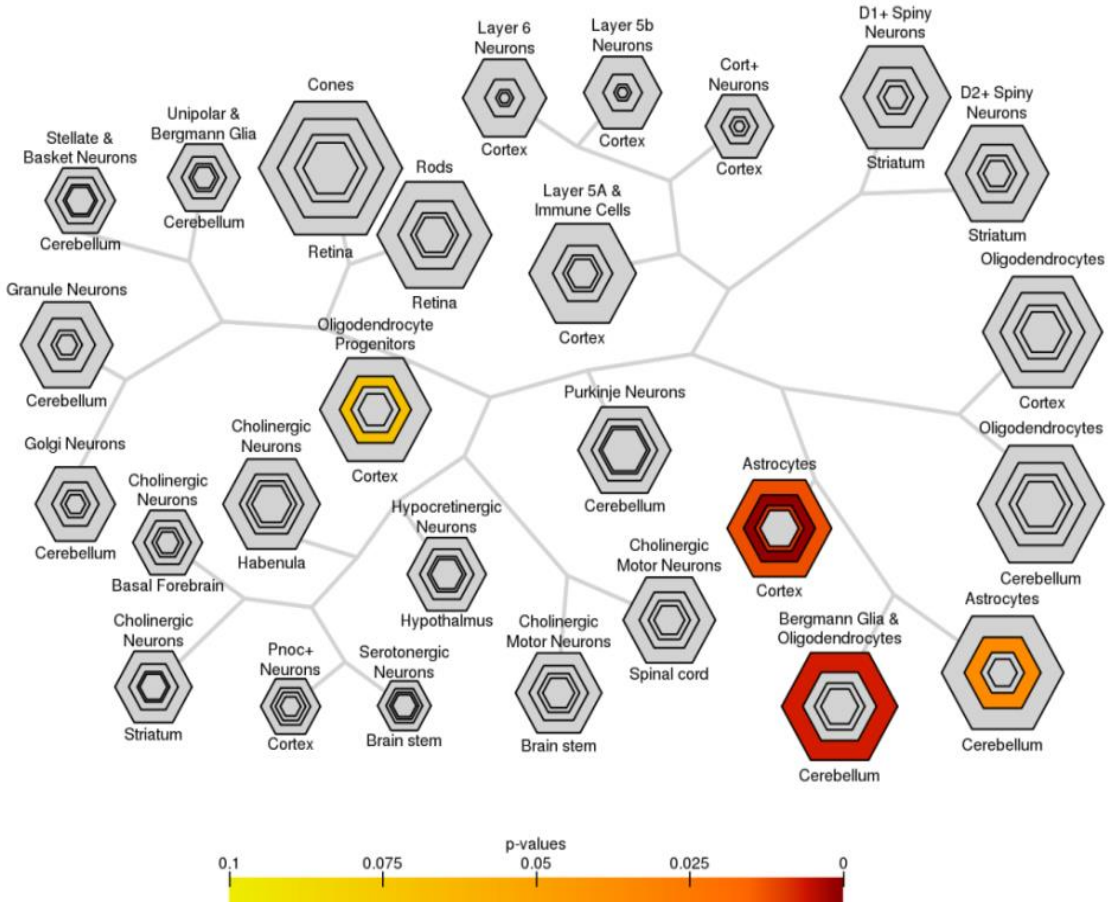


Figure 3. CSEA analysis reveals cortical astrocyte genes up-regulation beyond medication effects. The figure shows enrichments of up-regulated and medication-independent DE genes (N = 95) in specific cell types across brain using CSEA tool (BrainSpan data). Colored hexagons represent significant enrichments after FDR correction for hipergeometric test in different pSI thresholds. Enrichment values for every cell type are described in **supplementary table 8**.

Genetic rare variation is overrepresented in genes from up-regulated astrocytic modules in SCZ

We hypothesized that rare genetic variants (defined as allele frequency < 1% in general population) that are either within the genes with altered expression in affected brain regions of the patients or within the related co-expression networks are more likely to be enriched and the ones that contribute to the disease pathogenesis with high penetrance. Rare variants can be reliably identified from RNA-seq data for those genes with sufficient expression level (see **Methods**). Therefore, we conducted variant detection in the genes that are consistently expressed across all samples (DP>10 and RPKM > 0.3 in all samples, covering 1.35 Mb, which represents approximately 1.85% of GENCODE v19 exome). To ensure high quality in variant detection, we removed 2 SCZ and 9 CO samples that had relatively low depth of coverage and did not passed sample filtering cutoff (**supplementary figure 5**). Importantly, when rerunning all gene expression analysis described above, significant findings hold even if these samples were removed (**supplementary table 6).**

We also analyzed two well-validated databases that profiled RNA editing in human tissues including brain samples, REDIportal (Picardi et al., 2016) and SPRINT (Zhang et al., 2017), that comprised more than 4.5 million positions, to filter out variants described as potential RNA edits (RES). 6 and 2 variants affecting SCZ and CO samples, respectively, were removed based on this filter (**supplementary table 9**). Finally, variants with MAF > 0.01 in 1KG, ExAC or ESP were excluded. In total, we identified 796 rare SNVs and 131 indels in the transcriptomes of 29 SCZ and 17 CO (**supplementary figure 6; Methods**). These variants were within 623 genes. As expected, no global enrichment of rare variants was detected in the SCZ cases as compared to CO ($SCZ_{Var/Ind}$ (CI 95%) = 20.00 ± 1.69, $CO_{Var/Ind}$ (CI 95%) = 20.41 ± 2.49; t-test P-val = 0.771 **supplementary table 10**) due to existence of a large number of neutral variants.

We further analyzed the rare variants within all genes in the altered modules identified in our expression analysis as they are more likely to be involved in the disease and therefore enriched. In our main expression analysis, we demonstrated that genes up-regulated in SCZ are enriched in modules M10, C38, C36, K1 and K18, while genes down-regulated in SCZ are enriched in modules M1 and K15. Thus, we first analyzed variant enrichment in all genes within either the “up-regulated” or the “down-regulated” modules. Interestingly, we observed that rare singleton variants were more enriched (28 of 580 in SCZ versus 4 of 347 in CO; p = 0.0025, OR = 4.35, 95% CI = 1.50 – 17.19; **table 3**) in genes from the upregulated modules including M10, C38, C36, K1 and K18 (**table 3**). There are 28 rare singletons in cases, affecting 16 genes (N =29, 0.96 variants/individual), and 4 singletons affecting 4 genes in controls (N= 17, 0.24 variants/individual), reflecting a significant over-representation in SCZ versus CO (**figure 4**; Cochran-Armitage p = 0.0021; chi-sq = 9.48, df=1). No any enrichment was observed of these variants in genes from downregulated modules M1 and K15 (14 of 586 in SCZ versus 12 of 349 in CO; p = 0.411, OR = 0.69, 95% CI = 0.29– 1.65; **table 3**). We further analyzed rare variants in genes from medication-independent up-regulated modules. There is only a module, C38 that contains significantly up-regulated and medication-independent genes. Interestingly, we detected 8 mutations across three genes (5 mutations in *PCBD2*, two in *CAMTA1* and one in *BCL2L14*) in 29 SCZ and no mutations in 17 CO. The rare variants in this collection were significantly enriched in SCZ (p = 0.028, Fisher exact test; **table 3, supplementary table 11**).

	Expression analysis		**Variant enrichment**				
Gene sets	**Expression analysis**	**Cell type enriched**	**# genes mutated**	**# variants SCZ**	**# variants CO**	**ratio SCZ/CO**	**P-val**
All genes	-	-	623	580	347	0.98	-
Genes overlap with Modules M1 – K15	Down-regulated	Neuronal	20	14	12	0.69	0.412
Genes overlap with Modules C38-C36-K1-K18-M10	Up-regulated	Astrocyte	16	28	4	4.35	**0.0025**
Medicate-independent genes overlap with Module C38	Up-regulated	Astrocyte	3	8	0	-	**0.028**

Table 3. Variant enrichment in genes that belong to significantly up-regulated (C38, C36, K1, K18 and M10) and down-regulated modules (M1 - K15) from expression analysis across 29 schizophrenic and 17 control samples. Enrichment analysis was performed grouping downregulated modules and upregulated modules. Only genes with detectable expression (RPKM > 0.3) and genomic positions ubiquitously expressed (DP > 10) were considered. Among the genes mutated, *ATP1A2* is present in M10 and C36 (1 variant), and *LRIG1* is present in M10 and K18 (1 variant). The remaining genes did not overlap in any of the modules here analyzed (N=4 for M1; N=16 for K15; N=3 for C38; N=4 for C36; N=2 for K1; N=1 for K18 and N=8 for M10). Two-tailed Fisher tests were used for variant enrichment in both cases.

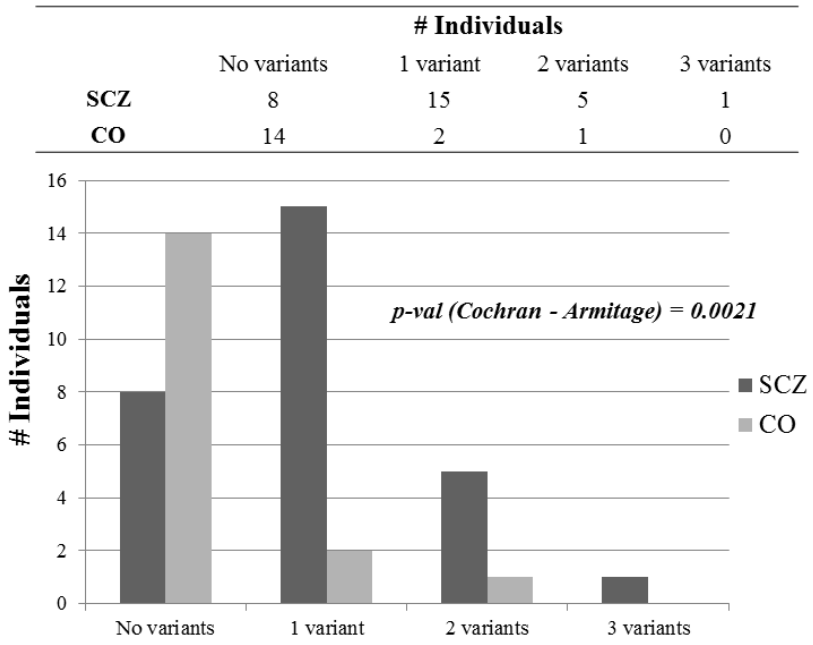


Figure 4. **Trend test result for variant accumulation in SCZ against CO**. In the table and the lower histogram the number of mutations per individual is described, from the absence of them up to 3 mutations in a schizophrenic individual. Cochran-Armitage was performed for trend test (P-val = 2.1×10^{-3}).

We further investigated the potential connection between rare variant enrichment in cell type specific signature using CSEA. Interestingly, we found that astrocyte signature genes that are up-regulated in SCZ carry more variants in SCZ than in CO (lowest FDR q-val=2×10^{-10} at pSI threshold 0.05 for Cerebellum astrocytes, **supplementary figure 15-16**, **supplementary table 12**). These results further support a potential link between alterations in astrocyte related genes and higher occurrence of rare variants in SCZ patients.

DISCUSSION

In this study, we analyzed the genetic makeup of the cingulate gyrus region from the postmortem brain tissues of a set of well-characterized patients with SCZ and matched

CO at both transcriptome and genome levels. We demonstrated some interesting links between gene expression and genetic variants in an affected brain region. First of all, we found a total of 1876 DE genes in SCZ, of which 174 are DE independently of medication. Secondly, dissecting the properties of these DE genes using well-defined gene expression signature modules from various independent studies, we observed a significant enrichment of the DE genes in several co-expression modules previously identified in developing human brain tissues (Hawrylycz et al., 2012; Kang et al., 2011; Miller et al., 2014). Our further analysis on the cell type specific gene sets suggested that there is a consistent enrichment among the up-regulated gene set related to astrocytes and independent of medication administrated through lifetime. Of all the 84 modules analyzed here by GSEA, prenatal co-expression module C38 (Miller et al., 2014), highly enriched in genes with decreased expression with age and related with replication, cell proliferation and chromatin assembly, appeared to be significantly up-regulated in SCZ in medication-independent expression dataset. Coherently with cell-type analyses, module C38 is enriched in genes expressed mainly in astrocytes during prenatal development (supplementary information in Miller et al., 2014), highlighting the importance of this cell type and potential contribution to the pathophysiology of SCZ. During expression analysis, several quality filters regarding outlier detection, gene expression cutoff and sample coverage were applied, demonstrating high robustness of our results.

Astrocytes are starting to be accepted to be equal important as neurons for normal neurodevelopment and function and malfunction of astrocytes can have a big impact on neuronal and higher cognitive functions observed in psychiatric disorders (Moraga-Amaro et al., 2014). In fact, abnormalities in astrocytes have been repeatedly reported in psychiatric diseases. For example, altered gene expression in astrocytes was described in SCZ (Catts et al., 2104; Chandley et al., 2013; Bernstein et al., 2009; Bernstein et al., 2015). In addition, recent studies revealed strong abnormalities in astrocytes in SCZ, which in turn lead to a compensatory effect by which neurons overexpress excitatory amino-acid transporters (EAATs) due to a loss of expression of these transporters in astrocytes (McCullumsmith et al., 2015). In our study, astrocyte related genes appear to be up-regulated even in the medical-independent gene set, pointing out to a possible mechanism underlying the pathophysiology of SCZ instead of the medication effects, which could be possibly extensive to other neuropsychiatric conditions, as major depression (Rajkowska et al., 2013) or Alzheimer disease (Sekar et al., 2015). It also raises the possibilities that abnormalities in astrocytes could be a disease mechanism that is insensitive to medication interference.

Furthermore, by investigating the rare variants within these altered co-expression modules and cell type specific gene sets, we found a significantly increase in the occurrence of rare variants in genes that belong to up-regulated and astrocyte-signature modules. This observation agrees with our hypothesis that rare variants in the genes that are highly expressed in the affected brain regions are more likely to be associated with disease diagnosis. It is worth noting that although our sample size is quite small and has limited power, our result indicated the strategy to focus only in a subset of transcriptome regions covering highly and ubiquitously expressed genes could be a powerful way to pinpoint genetic variation contributing to expression variance and pathophysiology of the disease. However, variant enrichment results may

be taken cautiously and suggestive, as larger sample sizes are needed to confirm these results.

In addition, our study suggested several interesting candidate genes for further investigation: For example, INA, a gene that belongs to the third most associated locus in the most recent SCZ GWAS to date (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) and clearly down-regulated in the patients with SCZ in this study (Fold change = 1.40; P-adj = 9.31×10^{-5}), is the third most significant DE gene among the medication-independent dataset (**supplementary table 3**). PCBD2, a gene that control BH4 metabolism and carries 4 rare mutations in SCZ but none in CO, also deserves further investigation.

4 of the 5 mutations within *PCBD2* are FMRP target sites for the most common form *FMRPiso7* (Ascano et al., 2012), which is interesting regarding the reported implication of this target regions in SCZ (Fromer et al., 2014; Purcell et al., 2014, Kirov et al., 2012) and in autism (De Rubeis et al., 2014). Although protein binding gene *PCBD2* has not been previously reported as a risk factor for neuropsychiatric conditions to our knowledge, this gene is involved in tetrahydrobiopterin (BH4) biosynthesis. BH4 is involved in the production of various neurotransmitters including serotonin, dopamine, epinephrine, and norepinephrine and is vital in nitric oxide (NO) production (Richardson et al., 2007; Werner et al., 2011). Deficiency in BH4 has been associated to SCZ (Richardson et al., 2005, Okusaga et al., 2014, Milstien and Katusic, 1999). In the likely context of astrocyte activation during early neurodevelopmental, post-transcriptionally regulatory mutations along *PCBD2* could have an important role in BH4 synthesis and regeneration pathway, causing damaging disturbances in the mentioned processes. Further study of this pathway in the context of SCZ development is necessary to understand the biological insights of *PCBD2* and involvement of BH4 pathway in SCZ.

In relation to mutated genes from other co-expression modules, heterozygous intragenic rearrangements in *CAMTA1*, calmodulin-binding transcription activator 1, were associated with intellectual disability (ID) and cerebellar ataxia (Mikhail et al., 2011; Thevenon et al., 2012) and deletion in this gene was also associated with autism (Pinto et al., 2010). Within genes from up-regulated module M10, glial fibrillary acidic protein *GFAP*, whose increased expression has been previously reported in SCZ (Catts et al., 2014) is mutated in 3 cases (1 exonic and 2 3'UTR mutations) but not in any CO. One of the UTR singleton variants, rs567636153, is a brain miRNA target site previously reported (Boudreau et al., 2014) and the individual with this variant has the higher expression value of the 46 individuals whose variants are studied, suggesting a likely genetic contribution of this UTR variant to the posttranscriptional regulation of this gene in a schizophrenic patient. Among the others genes mutated, other case has a variant (rs769452) in *APOE* rated as "probably damaging" in PolyPhen-2. This gene has been previously implicated in brain activity (Filippini et al., 2009) and extensively studied in SCZ with contradictory results (Martorell et al., 2008; Tovilla-Zarate et al., 2009). Altered gene expression of PMP2, a peripheral myelin protein from the same family of PMP22, whose deletions at locus *17p12* are associated with SCZ (Rees et al., 2013), has been previously reported in anterior cingulate cortex (Dracheva et al., 2006). The neurotrophic tyrosine kinase receptor type 2 *NTRK2* carries a mutation in cases. This gene has been associated with mood disorders and SCZ, by interacting

with other risk genes (Lin et al., 2013) and related with miRNA target binding sites (Warnica et al., 2015).

CONCLUSIONS

In summary, we have analyzed the transcriptome of patients with SCZ at both transcription and sequence level for the potential connection between altered gene expression pattern along the neurodevelopment and rare genetic variants within the expressed genes in an affected brain region. We consolidated the validity of neurodevelopmental model of SCZ making use of previously described co-expression modules and also in the context of rare genetic mutations, as had been indicated before (Gulsuner et al., 2013; Xu et al., 2011). Also, we found an interesting connection between up-regulated astrocytic gene sets and the enrichment of rare genetic variants that may potential contribute to differential gene expression in corresponding genes in the patients with SCZ.

CONFLICTS OF INTEREST

The authors declaim no conflicts of interest.

ACKNOWLEDGMENTS

RNA-seq data from postmortem brain tissues was provided by The Stanley Medical Research Institute's brain collection courtesy of Drs Michael B Knable, E Fuller Torrey, Maree J Webster, Serge Weis and Robert H Yolken. We thank Drs Joseph Gogos and Atsushi Takata for their valuable discussion of the analysis results. This work was partially supported by a National Alliance for Research in Schizophrenia and Depression (NARSAD) Young Investigator Award (to B.X.).

ABBREVIATIONS

1KG: 1000 genomes phase 3 release

BH-FDR: Benjamini-Hochberg False Discovery Rate

BY-FDR: Benjamini-Yekutieli False Discovery Rate

CNVs: Copy-number variants

CO: Controls

DE: Differentially expressed

DP: Depth of coverage

ESP: Exome sequencing project

ExAC: Exome Aggregation Consortium

GSEA: Gene Set Enrichment Analysis

NES: Normalized Enrichment Score

pcw: Post-conceptual weeks

RES: RNA editing site

RNA-seq: RNA sequencing

RPKM: Reads per kilobase per million mapped reads

SCZ: Schizophrenia

SNVs: Single nucleotide variant

BIBLIOGRAPHY

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biol*, *11*(10), R106.

Anders, S., Pyl, P. T., & Huber, W. (2014). HTSeq–A Python framework to work with high-throughput sequencing data. *Bioinformatics*, btu638.

Ascano, M., Mukherjee, N., Bandaru, P., Miller, J. B., Nusbaum, J. D., Corcoran, D. L., ...&Tuschl, T. (2012). FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature*, *492*(7429), 382-386.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165-1188.

Bernstein, H. G., Steiner, J., & Bogerts, B. (2009). Glial cells in SCZ: pathophysiological significance and possible consequences for therapy.

Bernstein, H. G., Steiner, J., Guest, P. C., Dobrowolny, H., & Bogerts, B. (2015). Glial cells as key players in SCZ pathology: recent insights and concepts of therapy. *Schizophrenia research*, *161*(1), 4-18.

Boudreau, R. L., Jiang, P., Gilmore, B. L., Spengler, R. M., Tirabassi, R., Nelson, J. A., ...& Davidson, B. L. (2014). Transcriptome-wide discovery of microRNA binding sites in human brain.*Neuron*, *81*(2), 294-305.

Cahoy, J. D., Emery, B., Kaushal, A., Foo, L. C., Zamanian, J. L., Christopherson, K. S., ... & Barres, B. A. (2008). A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *The Journal of Neuroscience*, *28*(1), 264-278.

Catts, V. S., Wong, J., Fillman, S. G., Fung, S. J., &Weickert, C. S. (2014). Increased expression of astrocyte markers in SCZ: association with neuroinflammation. *Australian and New Zealand Journal of Psychiatry*, *48*(8), 722-734.

Chandley, M. J., Szebeni, K., Szebeni, A., Crawford, J., Stockmeier, C. A., Turecki, G., ...& Ordway, G. A. (2013). Gene expression deficits in pontine locus coeruleus astrocytes in men with major depressive disorder. *J Psychiatry Neurosci*, *38*(4), 276-84.

Costain, G., Ho, A., Crawley, A. P., Mikulis, D. J., Brzustowicz, L. M., Chow, E. W., & Bassett, A. S. (2010). Reduced gray matter in the anterior cingulate gyrus in familial SCZ: a preliminary report. *Schizophrenia research*, *122*(1), 81-84.

Crespo-Facorro, B., Prieto, C., & Sainz, J. (2015). SCZ gene expression profile reverted to normal levels by antipsychotics. *International Journal of Neuropsychopharmacology*, *18*(4), pyu066.

De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Cicek, A. E., ... & Singh, T. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, *515*(7526), 209-215.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... &Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15-21.

Dougherty, J. D., Schmidt, E. F., Nakajima, M., & Heintz, N. (2010). Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. *Nucleic acids research*, *38*(13), 4218-4230.

Dracheva, S., Davis, K. L., Chin, B., Woo, D. A., Schmeidler, J., &Haroutunian, V. (2006). Myelin-associated mRNA and protein expression deficits in the anterior cingulate cortex and hippocampus in elderly SCZ patients. *Neurobiology of disease*, *21*(3), 531-540.

Drevets, W. C., Savitz, J., & Trimble, M. (2008). The subgenual anterior cingulate cortex in mood disorders. *CNS spectrums*, *13*(8), 663.

Efron, B. and Tibshirani, J. (1993).An introduction to the bootstrap. Chapman & Hall

Filippini, N., MacIntosh, B. J., Hough, M. G., Goodwin, G. M., Frisoni, G. B., Smith, S. M., ...& Mackay, C. E. (2009). Distinct patterns of brain activity in young carriers of the APOE-ε4 allele. *Proceedings of the National Academy of Sciences*, *106*(17), 7209-7214.

Fromer, M., Pocklington, A. J., Kavanagh, D. H., Williams, H. J., Dwyer, S., Gormley, P., ...& O'Donovan, M. C. (2014). De novo mutations in SCZ implicate synaptic networks. *Nature*, *506*(7487), 179-184.

Gulsuner, S., Walsh, T., Watts, A. C., Lee, M. K., Thornton, A. M., Casadei, S., ...& PAARTNERS Study Group. (2013). Spatial and temporal mapping of de novo mutations in SCZ to a fetal prefrontal cortical network. *Cell*,*154*(3), 518-529.

Hadland, K. A., Rushworth, M. F., Gaffan, D., & Passingham, R. E. (2003). The effect of cingulate lesions on social behaviour and emotion. *Neuropsychologia*, *41*(8), 919-931.

Hawrylycz, M. J., Lein, E. S., Guillozet-Bongaarts, A. L., Shen, E. H., Ng, L., Miller, J. A., ... & Royall, J. J. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, *489*(7416), 391-399.

Haznedar, M. M., Buchsbaum, M. S., Hazlett, E. A., Shihabuddin, L., New, A., & Siever, L. J. (2004). Cingulate gyrus volume and metabolism in the SCZ spectrum. *Schizophrenia research*, *71*(2), 249-262.

Herwig, R., Hardt, C., Lienhard, M., & Kamburov, A. (2016). Analyzing and interpreting genome data at the network level with ConsensusPathDB. *Nature protocols*, *11*(10), 1889-1907.

Insel, T. R. (2014). Brain somatic mutations: the dark matter of psychiatric genetics? *Molecular psychiatry*, *19*(2), 156-158.

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., ... & Guennel, T. (2011). Spatio-temporal transcriptome of the human brain. *Nature*, *478*(7370), 483-489.

Katsel, P., Byne, W., Roussos, P., Tan, W., Siever, L., & Haroutunian, V. (2011). Astrocyte and glutamate markers in the superficial, deep, and white matter layers of the anterior cingulate gyrus in SCZ. *Neuropsychopharmacology*, *36*(6), 1171-1177.

Kirov, G., Pocklington, A. J., Holmans, P., Ivanov, D., Ikeda, M., Ruderfer, D., ...& Owen, M. J. (2012). De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of SCZ. *Molecular psychiatry*, *17*(2), 142-153.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ...& Durbin, R. (2009). The sequence alignment/map format and SAMtools.*Bioinformatics*, *25*(16), 2078-2079.

Lin, Z., Su, Y., Zhang, C., Xing, M., Ding, W., Liao, L., ...& Cui, D. (2013). The interaction of BDNF and NTRK2 gene increases the susceptibility of paranoid SCZ. *PloS one*, *8*(9), e74264.

Martorell, L., Costas, J., Valero, J., Gutierrez-Zotes, A., Phillips, C., Torres, M., ...&Vallès, V. (2008). Analyses of variants located in estrogen metabolism genes (ESR1, ESR2, COMT and APOE) and SCZ.*Schizophrenia research*, *100*(1), 308-315.

McCullumsmith, R. E., O'Donovan, S. M., Drummond, J. B., Benesh, F. S., Simmons, M., Roberts, R., ...& Meador-Woodruff, J. H. (2015). Cell-specific abnormalities of glutamate transporters in SCZ: sick astrocytes and compensating relay neurons&quest. *Molecular psychiatry*.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, *20*(9), 1297-1303.

Mikhail, F. M., Lose, E. J., Robin, N. H., Descartes, M. D., Rutledge, K. D., Rutledge, S. L., ... & Carroll, A. J. (2011). Clinically relevant single gene or intragenic deletions encompassing critical neurodevelopmental genes in patients with developmental delay,

mental retardation, and/or autism spectrum disorders. *American Journal of medical Genetics Part A*, *155*(10), 2386-2396.

Miller, J. A., Horvath, S., & Geschwind, D. H. (2010). Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proceedings of the National Academy of Sciences*, *107*(28), 12698-12703.

Miller, J. A., Ding, S. L., Sunkin, S. M., Smith, K. A., Ng, L., Szafer, A., ... &Pletikos, M. (2014). Transcriptional landscape of the prenatal human brain.*Nature*, *508*(7495), 199-206.

Milstien, S., & Katusic, Z. (1999). Oxidation of tetrahydrobiopterin by peroxynitrite: implications for vascular endothelial function. *Biochemical and biophysical research communications*, *263*(3), 681-684.

Moraga-Amaro, R., Jerez-Baraona, J. M., Simon, F., & Stehberg, J. (2014). Role of astrocytes in memory and psychiatric disorders. *Journal of Physiology-Paris*, *108*(4), 240-251.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, *5*(7), 621-628.

Najas-García, A., Rufián, S., & Rojo, E. (2014). Neurodevelopment or neurodegeneration: review of theories of SCZ. *ActasEsp Psiquiatr*,*42*(4), 185-95.

Okusaga, O., Muravitskaja, O., Fuchs, D., Ashraf, A., Hinman, S., Giegling, I., ... & Hong, E. (2014). Elevated levels of plasma phenylalanine in SCZ: a guanosine triphosphate cyclohydrolase-1 metabolic pathway abnormality?. *PloS one*, *9*(1), e85945.

Pers, T. H., Timshel, P., Ripke, S., Lent, S., Sullivan, P. F., O'donovan, M. C., ... & Hirschhorn, J. N. (2016). Comprehensive analysis of SCZ-associated loci highlights ion channel pathways and biologically plausible candidate causal genes. *Human molecular genetics*, *25*(6), 1247-1254.

Picardi, E., D'Erchia, A. M., Lo Giudice, C., & Pesole, G. (2016). REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic acids research*, *45*(D1), D750-D757.

Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., ... & Almeida, J. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, *466*(7304), 368-372.

Purcell, S. M., Moran, J. L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., ...&Sklar, P. (2014). A polygenic burden of rare disruptive mutations in SCZ. *Nature*, *506*(7487), 185-190.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841-842.

Rajkowska, G., &Stockmeier, C. A. (2013). Astrocyte pathology in major depressive disorder: insights from human postmortem brain tissue. *Current drug targets*, *14*(11), 1225.

Ramsköld, D., Wang, E. T., Burge, C. B., & Sandberg, R. (2009).An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoSComputBiol*, *5*(12), e1000598.

Rapoport, J. L., Giedd, J. N., &Gogtay, N. (2012). Neurodevelopmental model of SCZ: update 2012. *Molecular psychiatry*, *17*(12), 1228-1238.

Rees, E., Walters, J. T., Georgieva, L., Isles, A. R., Chambert, K. D., Richards, A. L., ... & O'Donovan, M. C. (2013). Analysis of copy number variations at 15 SCZ-associated loci. *The British Journal of Psychiatry*, bjp-bp.

Richardson, M. A., Read, L. L., Taylor Clelland, C. L., Reilly, M. A., Chao, H. M., Guynn, R. W., ... & Clelland, J. D. (2005). Evidence for a tetrahydrobiopterin deficit in SCZ. *Neuropsychobiology*, *52*(4), 190-201.

Richardson, M. A., Read, L. L., Reilly, M. A., Clelland, J. D., & Clelland, C. L. T. (2007). Analysis of plasma biopterin levels in psychiatric disorders suggests a common BH4 deficit in SCZ and schizoaffective disorder. *Neurochemical research*, *32*(1), 107-113.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010).edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139-140.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 SCZ-associated genetic loci. *Nature*, *511*(7510), 421.

Sekar, S., McDonald, J., Cuyugan, L., Aldrich, J., Kurdoglu, A., Adkins, J., ...&Liang, W. S. (2015). Alzheimer's disease is associated with altered expression of genes involved in immune response and mitochondrial processes in astrocytes. *Neurobiology of aging*, *36*(2), 583-591.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ...&Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545-15550.

Sutherland, R. J., Whishaw, I. Q., & Kolb, B. (1988). Contributions of cingulate cortex to two forms of spatial learning and memory. *The Journal of neuroscience*, *8*(6), 1863-1872.

Takahashi, T., Suzuki, M., Kawasaki, Y., Hagino, H., Yamashita, I., Nohara, S. & Kurachi, M. (2003). Perigenual cingulate gyrus volume in patients with SCZ: a magnetic resonance imaging study. *Biological psychiatry*, *53*(7), 593-600.

Takata, A., Xu, B., Ionita-Laza, I., Roos, J. L., Gogos, J. A., & Karayiorgou, M. (2014). Loss-of-function variants in SCZ risk and SETD1A as a candidate susceptibility gene. *Neuron*, *82*(4), 773-780.

Thevenon, J., Lopez, E., Keren, B., Heron, D., Mignot, C., Altuzarra, C., ...& Minot, D. (2012). Intragenic CAMTA1 rearrangements cause non-progressive congenital ataxia with or without intellectual disability. *Journal of medical genetics*, *49*(6), 400-408.

Tovilla-Zarate, C., Medellin, B. C., Fresan, A., Apiquian, R., Dassori, A., Rolando, M., ...&Nicolini, H. (2009). APOE-ε3 and APOE-219G Haplotypes Increase the Risk for SCZ in Sibling Pairs. *The Journal of neuropsychiatry and clinical neurosciences*.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., ...&Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, *7*(3), 562-578.

Tsankova, N., Renthal, W., Kumar, A., & Nestler, E. J. (2007). Epigenetic regulation in psychiatric disorders. *Nature reviews. Neuroscience*, *8*(5), 355.

Vuong, H., Che, A., Ravichandran, S., Luke, B. T., Collins, J. R., & Mudunuri, U. S. (2015). AVIA v2. 0: annotation, visualization and impact analysis of genomic variants and genes. *Bioinformatics*, btv200.

Wang, D., Zhou, Y., Zhuo, C., Qin, W., Zhu, J., Liu, H. .& Yu, C. (2015). Altered functional connectivity of the cingulate subregions in SCZ. *Translational psychiatry*, *5*(6), e575.

Wang, K., Li, M., &Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, *38*(16), e164-e164.

Warnica, W., Merico, D., Costain, G., Alfred, S. E., Wei, J., Marshall, C. R., ...& Bassett, A. S. (2015). Copy number variable microRNAs in SCZ and their neurodevelopmental gene targets. *Biological psychiatry*, *77*(2), 158-166.

Werner, E. R., Blau, N., & Thöny, B. (2011). Tetrahydrobiopterin: biochemistry and pathophysiology. *Biochemical Journal*, *438*(3), 397-414.

Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I. & Bähler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, *453*(7199), 1239-1243.

Xu, B., Ionita-Laza, I., Roos, J. L., Boone, B., Woodrick, S., Sun, Y., ... & Karayiorgou, M. (2012). De novo gene mutations highlight patterns of genetic and neural complexity in SCZ. *Nature genetics*, *44*(12), 1365-1369.

Xu, B., Roos, J. L., Dexheimer, P., Boone, B., Plummer, B., Levy, S., ... & Karayiorgou, M. (2011). Exome sequencing supports a de novo mutational paradigm for SCZ. *Nature genetics*, *43*(9), 864-868.

Xu, B., Roos, J. L., Levy, S., Van Rensburg, E. J., Gogos, J. A., & Karayiorgou, M. (2008). Strong association of de novo copy number mutations with sporadic SCZ. *Nature genetics*, *40*(7), 880-885.

Xu, X., Wells, A. B., O'Brien, D. R., Nehorai, A., & Dougherty, J. D. (2014). Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. *The Journal of Neuroscience*, *34*(4), 1420-1431.

Zhang, F., Lu, Y., Yan, S., Xing, Q., & Tian, W. (2017). SPRINT: an SNP-free toolkit for identifying RNA editing sites. *Bioinformatics*, *33* (22), 3538-3548.

Zhang, Y., Chen, K., Sloan, S. A., Bennett, M. L., Scholze, A. R., O'Keeffe, S., ... & Wu, J. Q. (2014). An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *The Journal of Neuroscience*, *34*(36), 11929-11947.

Highlights

- Co-expression modules with astrocyte signature are up-regulated in cingulate gyrus in schizophrenia.
- Differences in gene expression in schizophrenia exist beyond influence of medication administrated.
- Rare somatic variation affecting genes within co-expression modules with astrocyte signature are enriched in schizophrenia.