Developmental and genetic regulation of the human cortex transcriptome in schizophrenia

Andrew E Jaffe[1,2,3,4,#], Richard E Straub[1], Joo Heon Shin[1], Ran Tao[1], Yuan Gao[1], Leonardo Collado Torres[1,3,4], Tony Kam-Thong[5], Hualin S Xi[6], Jie Quan[6], Qiang Chen[1], Carlo Colantuoni[1,7,8], Bill Ulrich[1], Brady J. Maher[1,9], Amy Deep-Soboslay[1], The BrainSeq Consortium, Alan Cross[10], Nicholas J. Brandon[10], Jeffrey T Leek[3,4], Thomas M. Hyde[1,7,9], Joel E. Kleinman[1,7], Daniel R Weinberger[1,8,9,11,&]

1. Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD 21205, USA
2. Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, 21205, USA
3. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, 21205, USA
4. Center for Computational Biology, Johns Hopkins University, Baltimore MD 21205 USA
5. Roche Pharma Research and Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, 4070 Basel, Switzerland
6. Computational Sciences, Pfizer Inc, Cambridge, MA 02140
7. Department of Neurology, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA
8. Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA
9. Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA
10. AstraZeneca Neuroscience iMed, IMED Biotech Unit, Cambridge, MA, 02139 USA
11. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA

# andrew.jaffe@libd.org (co-corresponding)

& drweinberger@libd.org (co-corresponding)

Summary:

GWAS have identified over 108 loci that confer risk for schizophrenia, but risk mechanisms for individual loci are largely unknown. Using developmental, genetic, and illness-based RNA sequencing expression analysis, we characterized the human brain transcriptome around these loci and found enrichment for developmentally regulated genes with novel examples of shifting isoform usage across pre- and post-natal life. Within patients and controls, we implemented a novel algorithm for RNA quality adjustment, and identified 237 genes significantly associated with diagnosis that replicated in an independent case-control dataset. These genes implicated synaptic processes and were strongly regulated in early development ($p < 10^{-20}$). Lastly, we found 42.5% of risk variants associate with nearby genes and diverse transcript features that converge on developmental regulation and subsequent dysregulation in illness and 34 loci show convergent directionality with illness association implicating specific causative transcripts. These data offer new targets for modeling schizophrenia risk in cellular systems.

Key Words: schizophrenia, functional genomics, RNA sequencing, human postmortem brain, differential expression analysis, RNA degradation

Introduction:

Schizophrenia (SCZD) is a prevalent neuropsychiatric disorder with a combination of genetic and environmental risk factors. Research over the last several decades has suggested that SZCD is a neurodevelopmental disorder arising through altered connectivity and plasticity in relevant neural circuits. However, discovering the causative mechanisms of these putatively developmental deficits has been very challenging[1]. The most consistent evidence of etiologic mechanisms related to SCZD has come from a recent genome-wide association study (GWAS) in which over a hundred independent single nucleotide polymorphisms (SNPs) were identified having an allele frequency difference between patients with schizophrenia and unaffected controls[2]. While these findings have identified regions in the genome harboring genetic risk variants, almost all of the associated SNPs are non-coding, located in intronic or intergenic sequence, and hypothesized to have some role in regulating expression[3]. However, the exact gene(s) and transcript(s) potentially regulated by risk-associated genetic variation are uncertain, as most of these genomic regions contain multiple genes. In principle, the effects of non-coding genetic variation, by whatever mechanisms (e.g. promoter, enhancer, splicing, noncoding RNA, epigenetics, etc), should be observed in the transcriptome. Therefore, to better understand how these regions of genetic risk and their underlying genotypes may confer risk of schizophrenia and to better characterize the molecular biology of the disease state, we sequenced the polyA+ transcriptomes from the prefrontal cortex of 495 individuals with ages across the lifespan, ranging from the second trimester of fetal life to 85 years of age (see Table S1), including 175 patients with schizophrenia (see Figure S1).

Here we identify novel expression associations with genetic risk and with illness state and explore developmentally regulated features, including a subset of genes with previously uncharacterized isoform shifts in expression patterns across the fetal-postnatal developmental transition. We further identify many more expression quantitative loci (eQTLs) in schizophrenia risk regions then previously observed by surveying the full spectrum of associated expression features to generate potential molecular mechanisms underlying genetic risk. We also explore differential gene expression associated with the state of illness in a comparison of the postmortem brains of patients with schizophrenia with non-psychiatric controls. By incorporating a novel, experiment-based algorithm to account for RNA quality differences which have not been adequately controlled in earlier studies, we report a high degree of replication across independent case-control gene expression datasets[4,5]. By combining genetic risk at the population-level with eQTLs and case-control differences, we identify putative human frontal cortex mechanisms underlying risk for schizophrenia and replicable molecular features of the illness state.

Results

We performed deep polyA+ RNA-sequencing of 495 individuals, ranging in age from the second trimester of fetal life to 85 years old (see Table S1), including 175 patients with schizophrenia (see Figure S1). We quantified expression across multiple transcript features, including: annotated 1) genes and 2) exons, 3) annotation-guided transcripts[4] as well as alignment-based 4) exon-exon splice junctions[5] and 5) expressed regions (ERs)[6]. These last two expression features were selected to reduce reliance on the potentially incomplete annotation of the brain transcriptome[7] (Results S1). We find a large number of moderately expressed and previously unannotated splice junctions that tag potential transcripts with alternative exonic boundaries or exon skipping (Figure S2), 95% of which are also found in other large RNA-seq datasets, including a subset that were brain-specific (Table 1). Similarly, we find that only 56.1% of ERs were annotated to strictly exonic sequence – many ERs annotated to strictly intronic (22.3%) or intergenic (8.5%) sequence, or were transcribed beyond existing annotation (e.g. extended UTRs, extended exonic sequence).

Developmental regulation of transcription and shifting isoform usage

Characterizing expression changes in unaffected individuals, particularly across brain development beginning with prenatal life, has previously offered disease-relevant insights into particular genomic loci [8-12]. Specifically, we and others [7,13,14] have shown that genomic risk loci associated with neurodevelopmental disorders including schizophrenia are enriched for transcript features showing differential expression between fetal and postnatal brains. Here too, among the 320 control samples, the strongest component of expression change corresponded to large expression changes in the contrast of pre-natal and early postnatal life, in line with previous data [7] (Figure 1A). We further defined a developmental regulation statistic for each expressed feature using a generalized additive model (see Methods) and found widespread developmental regulation of these expressed features (Results S2, Table S3, Figure S3), including of previously unannotated sequence (Table S4). Motivated in part by previous reports of preferential fetal isoform use among schizophrenia candidate risk genes[10,11] (e.g predominant fetal versus predominant postnatal isoforms), we next formally identified the subset of genes showing alternative isoform expression patterns across fetal and postnatal life using those exons, junctions, transcripts, and ERs that meet the statistical criteria for developmental regulation (, i.e. those genes with at least one developmentally changing feature, see Methods). There were 9,161 Ensembl genes (28% of the set of developmentally regulated genes) with both positive and negative expression features having genome-wide significant correlations to age (each with p_{bonf}<0.05, Figure 1B, Table S5, Figure S4). In other words, these represent alternate transcript isoforms of the same gene that show opposite patterns of expression across the prenatal-postnatal transition.

We performed gene set analyses of those genes with shifting isoform usage compared to the larger set of genes with at least one developmentally regulated feature but without shifting isoform usage to identify more specific biological functions of this unique form of developmental regulation (Table S6). The set of developmentally shifting isoforms was relatively enriched for localization, catalytic activity, signaling-related processes, including synaptic transmission and cell communication, and neuronal development, among many others. Interestingly, genes identified with shifting isoforms across development based exclusively on junction counts were enriched for both dopaminergic (FDR=5.11x10^{-7}) and glutamatergic (FDR=4.02x10^{-4}) synapse KEGG pathways (Figure 1C), the two neurotransmitter systems most prominently implicated in schizophrenia pathogenesis and treatment.

Schizophrenia risk is associated with shifting isoform usage across brain development

Based on the KEGG analysis, we hypothesized that the genes with developmentally regulated isoform shifts may relate to risk for schizophrenia. Indeed, genes within the SZCD GWAS risk loci were more likely to harbor these isoform shifts occurring in the fetal-postnatal developmental transition compared with the rest of the expressed transcriptome (Figure 2D). For example, genes with developmental isoform shifts identified by exon and junction counts were 62% (p=4.1x10^{-5}) and 69% (p=5.6x10^{-7}) more likely to lie within the PGC2 risk regions (with permutation-based p = 0.048 and 0.02 respectively, see Methods) than developmentally regulated genes without isoforms shifts (Table S7). These results further underscore the role of changes in the regulation of transcription and splicing in early brain development in schizophrenia risk.

Expression associations with chronic schizophrenia illness

We next explored the expression landscape of the prefrontal cortex of the schizophrenia illness state and its potential link with developmental regulation and genetic risk. We performed differential expression modeling using 351 higher quality adult samples (196 controls, 155 cases), and found extensive bias by RNA degradation within both univariate analysis (where 12,686 genes were differentially expressed at FDR<5%) and even after adjusting for measured levels of RNA quality typical of all prior studies (Figure S5). We therefore implemented a statistical framework based on an independent molecular degradation experiment (see Methods, Results S3), called “quality surrogate variable analysis” (qSVA, see Methods)[15]. We further utilized replication RNA-seq data from the CommonMind Consortium (CMC) dataset, using a subset of age range-matched 159 schizophrenia patients and 172 controls. Interestingly, adjusting for observed factors related to RNA quality that characterize all earlier studies of gene expression in schizophrenic brain, the proportion of genes with differentially expressed features at genome wide significant FDR < 5% that replicate (with directionality and marginal significance) in the CMC dataset was very small (only 11.0%, 244/2,215, Figure S6). In contrast, using our new statistical qSVA approach, 40.1% of differentially expressed genes at FDR < 5% (N=75/183) replicate in the CMC dataset. At genome-wide significant FDR<10% (see Methods), we identified 237 genes with 556 DE features that replicated in the CMC dataset (33.6% gene-level replication rate, Table S8, Table S9).

The differences in expression levels between cases and controls of these DE features were generally small in both our discovery and the replication datasets (Figure 2A, Figure S7), perhaps a direct result of the clinical and molecular heterogeneity of this disorder [13,16]. Gene ontology analysis implicated transporter- and channel-related signaling as significantly consistently downregulated in patients compared to controls across genes annotated in all three expression summarizations (Figure 2B, Table S10). These results suggested decreased signaling in patients with schizophrenia, but could raise the possibility that these replicated expression differences between patients and controls relate to the epiphenomena of illness, such as treatment with antipsychotics which affect signaling in the brain[14], as the majority of patients were on anti-psychotics at the time of death (64%, Table S1). Only two genes (*KLC1* and *PPP2R3A*) in the 108 significant schizophrenia GWAS loci were significantly differentially expressed. However, in an exploratory analysis, we found that overall the differential expression statistics within the loci were significantly different than those features outside the loci (Results S4, Figure S8, Table S11). We also investigated the relationships between transcription and genomic risk for schizophrenia using genome wide Risk Profile Scores (RPS) from each subject calculated as previously described[2] (see Methods). Using the subset of 209 Caucasian samples, we largely found a lack of association between RPS and expression of individual expression features. We further found a lack of enrichment of RPS on expression comparing the differentially expressed and replicated case-control features to the rest of the transcriptome, as well as lack of directionally consistency between RPS- and diagnosis-associated statistics among expressed features (Table S12).These results further suggest that the significant case-control expression differences show little overlap with genetic risk for the disorder.

In an earlier study of the epigenetic landscape of frontal cortex of patients with schizophrenia, we showed that DNA methylation levels in patients were closer to fetal methylation levels than to those of adult control samples[17]. Here we tested for analogous effects in the RNA-seq data related to the illness state. Every significant gene with differentially expressed features in the adult case-control analysis and replicated in the independent dataset showed evidence for developmental regulation across at least two expression feature types. We further found that the expression features normally more highly expressed in postnatal life tended to be more lowly expressed in patients compared to controls (max: $p=3.24\times10^{-11}$, min: $p=1.05\times10^{-70}$, Figure 2C) and features more highly expressed in fetal life tended to be more highly expressed in patients with schizophrenia compared to controls regardless of summarization type (max: $p=6.86\times10^{-33}$, min: $p < 10^{-100}$, Figure 2D). Analogous analyses for developmental regulation of schizophrenia-associated features without adjusting for the RNA quality qSVs were significant in the opposite directions, namely that schizophrenia-associated changes were further from, rather than closer to, fetal expression levels, which would be predicted as an artifact of residual RNA quality confounding (as the quality of the samples rank as fetal > adult control > adult SZ, see Table S1).These results further converge on a role for genes changing during brain development and maturation in schizophrenia, specifically that both DNA methylation and expression levels in adult patients appear to reflect levels in the developing brain more strongly than do those of unaffected individuals. These results also underscore the risk of spurious findings based on uncorrected RNA quality confounding.

Clinical enrichment of eQTL associations for schizophrenia

In order to elucidate the RNA features associated with schizophrenia risk variants themselves, rather than LD regions, we first performed a genome-wide *cis* (<500kb) expression quantitative trait loci (eQTL) analysis within the 412 post-adolescent subjects (see Methods). We then replicated SNP-feature eQTL pairs in two independent datasets (CMC and GTEx) to form a significant and replicated set of high-confidence eQTLs ("core eQTLs") to interrogate for clinical risk. Overall, we confirm widespread genetic control of expression in human brain (Figure 3A, Table S13, Table S14) which can be accessed by the research community via a user friendly browser at eqtl.brainseq.org. The majority of these core eQTLs were largely gene-specific, consistent across both interrogated races (Results S5, Figure 3B), very proximal to the TSS of genes (Figure 3C), and expression of the majority of genes was associated with more than 1 LD-independent SNP. We further found the largest effect sizes on average for the junction and ER eQTLs (Figure 3D), and interestingly a large proportion of these eQTLs corresponded to unannotated transcriptional activity. While the majority of genes with features as eQTLs have multiple possible Ensembl transcript isoforms, 67.0% of exon eQTLs and 31.1% of junction eQTLs were specific to a single Ensembl transcript when multiple transcripts existed for a given gene. Together these eQTL results, based on analyses across multiple feature summarizations from RNA sequencing data, demonstrate more widespread genetic regulation of expression than previously reported, including extensive transcript specificity and regulation of unannotated sequence in the human brain. These observations offer novel insights into cis mechanisms of common variation in brain.

We next explored the landscape of eQTLs associated with genetic risk using the set of core eQTLs among the PGC2-significant risk SNPs. There were 46 (42.5%) risk SNPs significantly associated with expression levels of 764 nearby features at FDR < 1% significance (Table S15, of the 108/128 PGC2 "index" SNPs that were present in our data). This proportion of SNPs was highly significant ($p=4.6x10^{-9}$) relative to the 19.7% of all tested common variants with eQTL signal. These risk eQTL features mapped back to 134 Ensembl genes (and 116 gene symbols). Interestingly, many of these adult-identified eQTLs show similar genetic regulation in fetal brain, analogous to observations made among methylation quantitative trait loci (meQTLs) at these risk SNPs[17,18]. In addition, the majority of these genetic risk-associated expression features were developmentally regulated across the lifespan in our control samples (N=528, 69.1%), suggesting that different developmental time periods may have different magnitudes of eQTL effects among these risk-associated expressed features. Even more strikingly, 57 genes (42.5%) having risk SNP eQTLs had shifting isoform usage across pre- and post-natal life, a highly enriched proportion relative to all expressed genes (9,205/43,046 genes with isoform shifts, OR=2.73, $p=4.21x10^{-9}$), further highlighting developmental relevance for schizophrenia genetic risk.

Identifying risk transcripts through clinical and molecular convergence

In principle, transcripts with convergence of directionality of association with both genetic risk and with clinical illness are especially attractive as putative molecular mechanisms of illness and as candidates for designing pathogenic models. We therefore used allele information and the corresponding effect estimates in the GWAS (odds ratio) and eQTL analyses (change in expression per allele copy) to test which of the expressed features that differentiated patients and controls were directionally consistent with the allelic directionality of genetic risk. In other words, if a given allele at a risk associated SNP is associated with increased expression of a nearby feature via the eQTL analysis, we tested whether expression of the same feature was more likely to be higher in patients compared to controls (Results S6, Figure 4A). Indeed, we found two-fold enrichment of this directional consistency between the GWAS, case-control, and FDR-significant eQTLs ($p=4.2x10^{-6}$). In order to reduce the chances of off-target eQTLs being driven by correlated expression features, we lastly retained SNP-feature associations that were at least half as significant on the $\log_{10}$ scale as the best eQTL for each GWAS index SNP (e.g. if the best eQTL association for a risk SNP was $p < 1x10^{-10}$, then eQTLs with $p > 1x10^{-5}$ were excluded).

This approach resulted in core eQTL signal across 34 GWAS loci with convergence of allele and illness state expression directionality (encompassing 59 genes across 231 features, Table S16). These are transcript features and genes that show consistent directionality of association of genetic risk alleles and of illness state, implicating these as particularly likely pathogenic factors underlying the GWAS loci. In the majority of these loci (N=22, 64.7%) the eQTLs represented a single Ensembl gene, nominating it as the most well supported “risk gene”, based on schizophrenia genetic risk coupled to expression features in the polyA+ fraction of postmortem DLPFC (Figure 4 and Figure S9). Of the remaining 12 GWAS loci that did not associate with features in only a single gene, 6 (17.6%) were similarly associated with the expression of two nearby genes, which may indicate different candidate genes whose expression is driven by different chromosomal haplotypes (Table 2). The effects of the remaining risk SNPs were more difficult to untangle, each associating with similar significance levels with three (N=2), four (N=3) or seven (N=1) genes. Finally, there were transcript-specific signals for many of the loci, including individual isoforms among multi-gene risk eQTLs. For example, there were 40 genes with transcript annotation (excluding gene- and region-level summarizations) and 25 genes contained features tagging single Ensembl transcripts (Table S17). This transcript specificity cautions that many gene count- or microarray-based approaches may miss important risk gene signals in eQTL analyses (see Discussion). These eQTL results highlight significant and independently-replicated risk-associated schizophrenia candidate genes and their specific transcripts that comprise links in the causative chain of schizophrenia in the human brain.

Discussion

We have explored the diverse landscape of expression correlates of schizophrenia risk and illness state in the postmortem human frontal cortex. Using deep RNA sequencing to define convergent measures of gene expression and early brain development, we identified widespread developmental regulation of transcription, including novel discoveries related to preferential isoform usage across brain development. These isoform “shifts” were associated with genetic risk for schizophrenia, and the directionality of dysregulation of developmentally regulated features suggest a more fetal-like expression profile in patients with schizophrenia compared to controls. Our approach to transcript characterization, which included extensive characterization of unannotated sequence, revealed that many more schizophrenia risk associated SNPs are brain eQTLs than previously reported - many risk SNPs only associate with a single gene, or even a single transcript, and many of these adult-identified eQTLs show similar genetic regulation in the fetal brain. Lastly, we identified significant and replicated genes differentially expressed in patients with schizophrenia compared to unaffected controls using a new experiment-based statistical framework to estimate and reduce the effects of latent RNA degradation bias. Without this new approach to RNA quality adjustment, replication across datasets is markedly limited if not negligible, and the directionality of the association with developmental isoform shifts is anomalous. These data suggest a convergence of developmental regulation and genetic risk for schizophrenia that appears relatively stable in patients ascertained at death, following decades of illness after diagnosis. We previously observed analogous stability of epigenetic marks highlighting prenatal life in adult patients with schizophrenia[17], suggesting that both genetic and environmental risk factors implicated in schizophrenia illness involve early developmental events which are still observable in the brain tissue of adult individuals who have been ill for many years.

While our approach utilizing convergent expression features – genes, exons, transcripts, junctions, and expressed regions – results in more complicated data processing and analysis, it can potentially cast a wider net in the search for valid biological signals in RNA sequencing datasets. Using all convergent features overcomes the limitations related to any given feature summarization, including the inability to measure and interrogate unannotated or novel transcribed sequence using gene and exon counts, and the difficulties in full transcript assembly from short sequencing reads [19]. We note that both quantifying and analyzing splice junctions, and also transcript-level data, rely on junction-spanning reads for statistical power. In our data, there were approximately 3 times (IQR: 2.86-3.24) more reads available by gene/exon counting approaches than those that contain splice junctions, likely explaining why gene counts discovered more differentially expressed genes in the schizophrenia diagnosis analyses. Two relatively new approaches utilized here – direct quantification and statistical analyses of splice junction counts and expressed regions – can identify differential expression signal when it is outside of the annotated transcriptome. The junction-level approach can also identify previously uncharacterized novel transcribed sequence, which we replicated in other large publicly available datasets, as well as delineate individual transcripts or classes of transcripts that share a particular splice junction, and as read lengths increase, the proportion of reads containing splice junctions will increase, making junction- and transcript-based approaches even more powerful, including those recently developed to identify splicing QTLs [20].

Our analysis of RNA-seq data identified widespread shifts in preferential isoform use across brain development, which would have been impossible to identify using only gene-level data and incomplete with only exon-level data (Figure 4). The genes with these isoform shifts were significantly enriched for neurodevelopmental and cellular signaling processes, and as well as for genes in regions of genetic risk for schizophrenia. A prevalent hypothesis suggests that schizophrenia is a neurodevelopmental disorder that arises because of altered connectivity and plasticity in the early assembly of relevant neural circuits[1], and the potential convergence of genetic risk with developing signaling processes across human brain development should point to specific candidate molecular disruptions occurring during the wiring of the fetal brain. Indeed, inefficient or disrupted signaling and tuning is thought to underlie the expression of illness in the adult brain [1], and the most successful therapeutics work through improving these processes[14]. Consistent with this hypothesis, we find evidence for differences in the expression of genes coding for subunits of ion channels in the cortices of patients with schizophrenia compared to controls. We observed significant differential expression of both voltage-gated (*KCNA1*, *KCNC3*, KCNK1, *KCNN1*, *SCN9A*) and ligand gated ion channels (*GRIN3A*, *GABRA5*, *GABRB3*), transporters (*SLC16A2*, *ALC25A33*, *SLC26A11*, *SLC35F2*, *SLC7A3*), and ion channel auxiliary subunits (*KCNIP3*, *SCN1B*), supporting other evidence that the clinical phenomenology of schizophrenia is associated with altered neuronal excitability [21]. While these findings implicating basic mechanisms of cortical circuit dynamics may underlie fundamental aspects of the clinical disorder, the possibility that they are driven by the effects of pharmacological treatment and are thus state dependent epiphenomena cannot be excluded.

Our eQTL analyses are among the largest and most comprehensive to date in human brain tissue, requiring stringent genome-wide significance and independent replication, and offer additional insights into the genetic regulation of RNA expression levels. We show that over half of the genes with eQTL signals are associated with three or more nearby LD-independent SNPs, suggesting a potential mechanism to explain how expression levels at the population-level are relatively normally distributed while so many genes show distinct eQTL signals. Our data also suggest more widespread regulation of specific transcript isoforms, which we were able to identify using exon- and junction-level analyses. This transcript-specific genetic regulation was particularly prevalent among schizophrenia risk variants, where 62.5% of loci containing multiple transcripts showed clinically- and molecularly-consistent eQTL signal to a single Ensembl transcript isoform. Overall, we have identified many more eQTLs to genome-wide significant schizophrenia risk variants – 42.5% - than previously reported, experimentally implicating far more potential “risk” genes within these loci than previously characterized.

These eQTL associations within the genome-wide significant schizophrenia loci generate novel putative biological mechanisms underlying risk for the disorder. We highlight 34 GWAS loci that contain significant eQTLs that show convergence of consistent altered expression across both genetic risk and case status in the human brain. These loci often point to individual “risk” genes or even more specific “risk” transcripts that can represent targetable entry points for more focused cellular assays and model organism work to better characterize schizophrenia risk mechanisms. Moreover, observation of convergence in directionality of molecular association with genetic risk and clinical state identifies a compelling strategy and directionality for target rescue, specifically to increase or decrease the function of the target transcript(s) and downstream effectors. Focusing solely on increased or decreased expression in brains of patients compared to controls, without considering genetic risk variants and their regulation of local gene expression, will likely predominantly highlight molecular changes resulting from the schizophrenia illness state, as we suggest with consistent down-regulation of ion channels.

We stress the priority of identifying the most relevant cellular consequences of genetic risk, which we view as production of particular isoforms with predicted directionality, rather than trying to identify “causal” mutations tagged by “marker” risk SNPs from the GWAS. We suggest that identifying convergence between genetic risk and potential molecular consequences of the disorder is likely to result in better, or at least more consistent support for, targets for drug discovery efforts.

Author Contributions

A.E.J – performed primary data processing and analyses, led the writing of the manuscript

R.E.S – contributed to data analysis and writing of the manuscript

J.H.S., R.T. , Y.G. – performed RNA sequencing data generation (RNA extraction, library preparation, and sequencing) and QC analyses

L.C.T.,J.T.L – performed region-level data generation and assisted in data analysis and interpretation

T.K.T.,S.X.,J.Q.,C.C.,B.J.M., A.C.,N.B.,BrainSeq – provided feedback on manuscript and contributed to data analyses and interpretations on eQTL analyses.

B.U. – created user-friendly database of eQTLs

A.D.S., T.M.H.,J.E.K.- collected, consented, characterized, and dissected human brain tissue

D.R.W. – designed and oversaw the research project, wrote the manuscript

Tony Kam-Thong is employed by F. Hoffmann-La Roche

Hualin S Xi and Jie Quan are employees of Pfizer Inc.

Alan Cross, and Nicholas J.Brandon were full time employees and shareholders in AstraZeneca at the time these studies were conducted.

The remaining authors declare no competing financial interests.

Data Availability: sequencing reads and genotype data are available through SRA and dbGaP at accession numbers: [TBD] following publication**.**

Acknowledgements:

We thank Dr. Ronald Zielke, Robert D. Vigorito, and Robert M. Johnson of the National Institute of Child Health and Human Development Brain and Tissue Bank for Developmental Disorders at the University of Maryland for their provision of fetal, child, and adolescent brain specimens; This work was supported by the funding from Lieber Institute for Brain Development and the Maltz Research Laboratories. The work was partially supported by R21MH109956 to A.E.J. L.C.T was supported by Consejo Nacional de Ciencia y Tecnología México 351535.

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health. Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI\SAIC-Frederick, Inc. (SAIC-F) subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to The Broad Institute, Inc. Biorepository operations were funded through an SAIC-F subcontract to Van Andel Institute (10ST1035). Additional data repository and project management were provided by SAIC-F (HHSN261200800001E). The Brain Bank was supported by a supplements to University of Miami grants DA006227 & DA033684 and to contract N01MH000028. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951, MH090937, MH101820, MH101825), the University of North Carolina - Chapel Hill (MH090936 & MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University St Louis (MH101810), and the University of Pennsylvania (MH101822). The data used for the analyses described in this manuscript were obtained from dbGaP accession number phs000424.v6.p1 on October 6, 2015.

Data were generated as part of the CommonMind Consortium supported by funding from Takeda Pharmaceuticals Company Limited, F. Hoffman-La Roche Ltd and NIH grants R01MH085542, R01MH093725, P50MH066392, P50MH080405, R01MH097276, RO1-MH-075916, P50M096891, P50MH084053S1, R37MH057881 and R37MH057881S1, HHSN271201300031C, AG02219, AG05138 and MH06692. Brain tissue for the study was obtained from the following brain bank collections: the Mount Sinai NIH Brain and Tissue Repository, the University of Pennsylvania Alzheimer's Disease Core Center, the University of Pittsburgh NeuroBioBank and Brain and Tissue Repositories and the NIMH Human Brain Collection Core. CMC Leadership: Pamela Sklar, Joseph Buxbaum (Icahn School of Medicine at Mount Sinai), Bernie Devlin, David Lewis (University of Pittsburgh), Raquel Gur, Chang-Gyu Hahn (University of Pennsylvania), Keisuke Hirai, Hiroyoshi Toyoshiba (Takeda Pharmaceuticals Company Limited), Enrico Domenici, Laurent Essioux (F. Hoffman-La Roche Ltd), Lara Mangravite, Mette Peters (Sage Bionetworks), Thomas Lehner, Barbara Lipska (NIMH).

Members of the BrainSeq consortium include: Christian R. Schubert, Patricio O'Donnell, Jie Quan, Jens R. Wendland, Hualin S. Xi, Ashley R. Winslow, Enrico Domenici, Laurent Essioux, Tony Kam-Thong, David C. Airey, John N. Calley, David A. Collier, Hong Wang, Brian Eastwood, Philip Ebert, Yushi Liu, Laura Nisenbaum, Cara Ruble, James E. Scherschel, Ryan Matthew Smith, Hui-Rong Qian, Kalpana Merchant, Michael Didriksen, Mitsuyuki Matsumoto, Takeshi Saito, Nicholas J. Brandon, Alan J. Cross, Qi Wang, Husseini Manji, Hartmuth Kolb,

Maura Furey, Wayne C. Drevets, Joo Heon Shin, Andrew E. Jaffe, Yankai Jia, Richard E. Straub, Amy Deep-Soboslay, Thomas M. Hyde, Joel E. Kleinman, Daniel R. Weinberger

Figure Legends

Figure 1: Developmental regulation of expression. (A) Principal component #1 of the gene-level expression data versus age; PCW: post-conception weeks, remaining ages are in years. (B) Expression features fall into two main development regulation signatures, increasing in expression from fetal to postnatal life (orange) or decreasing from fetal to postnatal life (blue). Y-axis is Z-scaled expression (to standard normal), dark lines represent median expression levels, and confidence bands represent 25th-75th percentiles of expression levels for each class of features. (C) KEGG pathways enriched for genes with isoform shifts, stratified by which feature type identified the gene as having a switch. Coloring/scaling represents -log10(FDR) for gene set enrichment. Analogous data for GO gene sets (biological processes, BP, and molecular function, MF) are available in Table S6. DER: differentially expressed region. Enrichment analyses for isoform shift genes among PGC2 schizophrenia GWAS risk loci with exon and junction counts using both (D) parametric p-values) and (E) permutation-based p-values. OR: odds ratio.

Figure 2: Differential expression comparing patients with schizophrenia to controls. (A) Histogram of fold changes of the diagnosis effect of those features that were significant and independently replicated, colored by feature type. (B) Gene set analyses of genes with decreased expression in patients compared to controls by feature type. Coloring/scaling represents -log10(FDR) for gene set enrichment. Significant directional effects of developmental regulation among diagnosis-associated features for those features that (C) increased and (D) decreased across development (i.e. those features shown in Figure 1B). P-values provided for Wilcoxon rank sign test for those features developmentally regulated among case-control differences to those not developmentally regulated.

Figure 3: Expression quantitative trait loci (eQTL) in the human brain. (A) Table summarizing eQTL results across the five feature summarization types (gene, exon, transcript, junction and expressed region). Bonf: Bonferroni-adjusted p-value, Ens: Ensembl, Sym: Symbol, IQR: interquartile range, 25th-75th percentiles, bp = base pairs, kb = kilobases. (B) Similar standardized effect sizes (T-statistics) for eQTLs identified in full sample within just Caucasian (CAUC) and African American (AA) computed separately, here using exon-level eQTL results. (C) Histogram of the distance of the most significant SNP to each eQTL expression feature. (D) Boxplots of the effect sizes (absolute fold changes per minor allele copy) stratified by feature type.

Figure 4: Clinical enrichment of schizophrenia risk. (A) Image describing enforcement of directional consistency between GWAS, eQTL, and case-control differences. We highlight example eQTLs to GWAS-positive variants among annotated exons in (B) *SRR* and (C) *STAT6*, as well as unannotated expressed features, for example an intronic region in (D) *HSPD1 /HSPE1*, shown as a star in (E), and a junction linked to (F) only one annotated exon of

IMMP2L, shown as a bar in (G). In panels E and G, dark blue: exon, light blue: intron; coordinates relative to hg19.

Tables

		Transcript Class			
Number	**Datasets**	**In Ensembl**	**Exon Skip**	**Alt Boundary**	**Novel**
251557	Neither	2680(1.3%)	118(1.4%)	315(1.6%)	4613(24.8%)
	Geuvadis	134(0.1%)	3(0%)	14(0.1%)	116(0.6%)
	GTEX	7883(3.8%)	1257(14.8%)	4312(22.3%)	6768(36.3%)
	Both	194342(94.8%)	7136(83.8%)	14730(76%)	7136(38.3%)

Table 1: splice junction annotation and characterization in GTEx and GEUVADIS for moderately expressed junctions (mean reads per 80M mapped reads, RP80M > 1). Each column represents a 2x2 table for presence of identified junctions in 495 DLPFC samples in two independent polyA+ datasets.

GWAS Rank	Risk Direction	Gene	Feature Type	Best p-value
3	UP	*AS3MT*	E	6.73E-40
3	UP	*BORCS7*$^{+}$	G,E,R	9.23E-34
3	UP	*AS3MT**	J	9.57E-25
7	DOWN	*FTSJ2*	E,J,G,R	2.64E-07
8	DOWN	*GNL3LP1*	E	3.11E-10
8	DOWN	*ERCC8*	E,R	6.22E-10
9	DOWN	*ARL6IP4*	J	4.12E-36
16	UP	*IMMP2L*	J,T	6.04E-10
17	UP	*SNX19*	J	9.77E-07
23	UP	*C2orf82*	T	6.33E-06
24	DOWN	*NRGN*	R	3.5E-05
24	DOWN	*VSIG2*	R	8.6E-05
31	UP	*GOLGA2P7*	E,R	5.12E-30
31	UP	*GOLGA6L5P*	R	8.13E-24
33	UP	*HSPD1*	R	3.34E-14
34	UP	*XPNPEP3*	E,J,R	3.63E-08
42	DOWN	*PCCB*	G,E,R	2.95E-11
47	DOWN	*SRR*	E,G,R,J	2.1E-12
51	DOWN	*GATAD2A*	R	2.51E-08
51	DOWN	*NDUFA13*	J	9.36E-06
52	UP	*VPS45*	R	1.88E-16
57	UP	*RERE*	E	7.7E-07
73	UP	*CTNNA1*	E	5.18E-07
84	UP	*GPM6A*	R	3.12E-06
94	UP	*PRMT7*	R,J,E	5.37E-09
94	UP	*TSNAXIP1*	G,R	8.78E-06
96	DOWN	*ATPAF2*	E	4.58E-07
103	UP	*STAT6*	E,J	1.71E-07
106	DOWN	*SOX2-OT*	R	8.18E-12
106	DOWN	*RP11-275H4.1*	T	2.62E-06
117	DOWN	*FANCL*	E,R	2.45E-17
120	DOWN	*BRINP2*	R	6.7E-06
121	UP	*CD46*	E,R,G,J	1.57E-37
125	UP	*IRF3*	J,E	4.51E-10

Table 2: eQTL and clinical directional consistent signal around the GWAS risk variants for schizophrenia. GWAS rank is relative to Supplementary Table 2 in Ripke et al [2]. Direction refers to whether schizophrenia risk represents increased or decreased expression of the particular features. Feature type indicates which of (G)enes, (E)xons, (J)unctions, (T)ranscripts, or expressed (R)egions showed eQTL association to the risk SNP. The best p-value for the risk SNP to features in each gene are recorded in the last column. $^{+}$*C10orf32* in Ensembl v75 but updated in later versions; * unannotated junction found to tag a new transcript of the gene, AS3MT^{d2d3}, that has been more fully characterized by Li et al[22].

References

1 Weinberger, D. R. & Levitt, P. in *Schizophrenia* 393-412 (Wiley-Blackwell, 2011).

2 Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-+, doi:10.1038/nature13595 (2014).

3 Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195, doi:10.1126/science.1222794 (2012).

4 Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **33**, 290-295, doi:10.1038/nbt.3122 (2015).

5 Nellore, A. *et al.* Human splicing diversity across the Sequence Read Archive. *bioRxiv*, doi:10.1101/038224 (2016).

6 Collado Torres, L. *et al.* Flexible expressed region analysis for RNA-seq with derfinder. *Nucleic Acids Research* **In Press.**, doi:10.1101/015370 (2016).

7 Jaffe, A. E. *et al.* Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nat Neurosci* **18**, 154-161, doi:10.1038/nn.3898 (2015).

8 Tan, W. *et al.* Molecular cloning of a brain-specific, developmentally regulated neuregulin 1 (NRG1) isoform and identification of a functional promoter variant associated with schizophrenia. *The Journal of biological chemistry* **282**, 24343-24351, doi:10.1074/jbc.M702953200 (2007).

9 Kao, W. T. *et al.* Common genetic variation in Neuregulin 3 (NRG3) influences risk for schizophrenia and impacts NRG3 expression in human brain. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 15619-15624, doi:10.1073/pnas.1005410107 (2010).

10 Tao, R. *et al.* Expression of ZNF804A in human brain and alterations in schizophrenia, bipolar disorder, and major depressive disorder: a novel transcript fetally regulated by the psychosis risk variant rs1344706. *JAMA psychiatry* **71**, 1112-1120, doi:10.1001/jamapsychiatry.2014.1079 (2014).

11 Hyde, T. M. *et al.* Expression of GABA signaling molecules KCC2, NKCC1, and GAD1 in cortical development and schizophrenia. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **31**, 11088-11095, doi:10.1523/JNEUROSCI.1234-11.2011 (2011).

12 Birnbaum, R., Jaffe, A. E., Hyde, T. M., Kleinman, J. E. & Weinberger, D. R. Prenatal expression patterns of genes associated with neuropsychiatric disorders. *The American journal of psychiatry* **171**, 758-767, doi:10.1176/appi.ajp.2014.13111452 (2014).

13 Buchanan, R. W. & Carpenter, W. T. Domains of psychopathology: an approach to the reduction of heterogeneity in schizophrenia. *The Journal of nervous and mental disease* **182**, 193-204 (1994).

14 Winterer, G. & Weinberger, D. R. Genes, dopamine and cortical signal-to-noise ratio in schizophrenia. *Trends in neurosciences* **27**, 683-690, doi:10.1016/j.tins.2004.08.002 (2004).

15 Jaffe, A. E. *et al.* A framework for RNA quality correction in differential expression analysis. *bioRxiv*, doi:10.1101/074245 (2016).

16 Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427, doi:10.1038/nature13595 (2014).

17 Jaffe, A. E. *et al.* Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nature neuroscience* **19**, 40-47, doi:10.1038/nn.4181 (2016).

18 Hannon, E. *et al.* Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nature neuroscience* **19**, 48-54, doi:10.1038/nn.4182 (2016).

19 Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nature methods* **10**, 1177-1184, doi:10.1038/nmeth.2714 (2013).

20 Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600-604, doi:10.1126/science.aad9417 (2016).

21 Uhlhaas, P. J. & Singer, W. Abnormal neural oscillations and synchrony in schizophrenia. *Nature reviews. Neuroscience* **11**, 100-113, doi:10.1038/nrn2774 (2010).

22 Li, M. *et al.* A human-specific AS3MT isoform and BORCS7 are molecular risk factors in the 10q24.32 schizophrenia-associated locus. *Nature medicine*, doi:10.1038/nm.4096 (2016).

Methods

Postmortem brain samples

Post-mortem human brain tissue was obtained by autopsy primarily from the Offices of the Chief Medical Examiner of the District of Columbia, and of the Commonwealth of Virginia, Northern District, all with informed consent from the legal next of kin (protocol 90-M-0142 approved by the NIMH/NIH Institutional Review Board). Additional post-mortem fetal, infant, child and adolescent brain tissue samples were provided by the National Institute of Child Health and Human Development Brain and Tissue Bank for Developmental Disorders (http://www.BTBank.org) under contracts NO1-HD-4-3368 and NO1-HD-4-3383. The Institutional Review Board of the University of Maryland at Baltimore and the State of Maryland approved the protocol, and the tissue was donated to the Lieber Institute for Brain Development under the terms of a Material Transfer Agreement. Clinical characterization, diagnoses, and macro- and microscopic neuropathological examinations were performed on all samples using a standardized paradigm, and subjects with evidence of macro- or microscopic neuropathology were excluded. Details of tissue acquisition, handling, processing, dissection, clinical characterization, diagnoses, neuropathological examinations, RNA extraction and quality control measures were described previously in Lipska, et al. [23]. The Brain and Tissue Bank cases were handled in a similar fashion (http://medschool.umaryland.edu/BTBank/ProtocolMethods.html). Antipsychotic use was measured using toxicology at time of death.

RNA extraction and sequencing

Post-mortem tissue homogenates of dorsolateral prefrontal cortex grey matter (DLPFC) approximating BA46/9 in postnatal samples and the corresponding region of PFC in fetal samples were obtained from all subjects. Total RNA was extracted from ~100 mg of tissue using the RNeasy kit (Qiagen) according to the manufacturer’s protocol. The poly-A containing RNA molecules were purified from 1 µg DNAse treated total RNA and sequencing libraries were constructed using the Illumina TruSeq© RNA Sample Preparation v2 kit. Sequencing indices/barcodes were inserted into Illumina adapters allowing samples to be multiplexed in across lanes in each flow cell. These products were then purified and enriched with PCR to create the final cDNA library for high throughput sequencing using an Illumina HiSeq 2000 with paired end 2x100bp reads.

RNA sequencing data processing

The Illumina Real Time Analysis (RTA) module performed image analysis, base calling, and the BCL Converter (CASAVA v1.8.2), generating FASTQ files containing the sequencing reads. These reads were aligned to the human genome (UCSC hg19 build) using the spliced-read mapper TopHat (v2.0.4) using the reference transcriptome to initially guide alignment, based on known transcripts of the previous Ensembl build GRCh37.67 (the "–G" argument in the software) [24]. We achieved a median of 85.3 million (IQR: 71.7M-111.2M) aligned reads per sample (see Table S1).

We characterized the transcriptomes of these 495 samples using five convergent measurements of expression ("feature summarizations")– (1) gene and (2) exon counts, and (3) transcript-level quantifications that rely on existing gene annotation, and two annotation-agnostic approaches we have developed that are determined solely from the read alignments – (4) read coverage supporting exon-exon splice junctions (e.g. coordinates of potentially intronic sequence that are spliced out of mature transcripts captured by a single read) and (5) read coverage overlapping each base in each sample which we have summarized into contiguous "expressed regions" (ERs, see Methods, Figure S1). These last three measurements generate expression for features of interest that can "tag" elements of transcripts in the data that are not constrained by limitations or incompleteness of existing annotation, and the counts for these features can then be directly used for differential expression analysis.

1. Gene counts were generated using the featureCounts tool[25] based on the more recent Ensembl v75, which was the last stable release for the hg19 genome build, using single end read counting [featureCounts –a $GTF –o $OUT $BAM]. We converted counts to RPKM values using the total number of aligned reads across the autosomal and sex chromosomes (dropping reads mapping to the mitochondria chromosome).
2. Exon counts were also generated using the featureCounts tool[25] based on the more recent Ensembl v75, using single end read counting, and allowing reads to be assigned to multiple exons (e.g. those with splice junctions) [featureCounts –O –f –a $GTF –o $OUT $BAM]. We converted counts to RPKM values using the total number of aligned reads across the autosomal and sex chromosomes (dropping reads mapping to the mitochondria chromosome).
3. Junction counts were generated by first filtering the TopHat BAM file to primary alignments only [samtools view -bh -F 0x100 $BAM > $NEWBAM] and regtools [26] was used to extract analogous junction information (coordinates and number of reads supporting) as the TopHat output. We found that native TopHat output (junctions.bed) was based on both primary and secondary alignments, which could influence the degree of potentially novel splice junctions. We used a modified version of TopHat's "bed_to_juncs" program to retain the number of supporting reads (in addition to returning the coordinates of the spliced sequence, rather

than the maximum fragment range), and used R code (see Supplementary Code) to combine and annotate these junctions across all samples. We annotated identified splice junctions using Ensembl v75 – while the initial alignment was guided by Ensembl v67, novel junctions, by definition, are identified in the second genome alignment, rather than the initial guided transcriptome alignment step. We converted counts to "RP80M" values, or "reads per 80 million mapped" using the total number of aligned reads across the autosomal and sex chromosomes (dropping reads mapping to the mitochondria chromosome), which can be interpreted as the number of reads supporting the junction in an average library size (we were targeting 80M reads in the sequencing). Most junctions were lowly expressed in our homogenate tissue, with fewer than 1 average normalized supporting read (N=3,330,642; 92.98%) including approximately half unique to a single individual (N= 1,779,241, 49.67%).

4. Transcripts were assembled using StringTie[4] (version 1.1.2) guided by Ensembl v75 annotation within each sample [stringtie $BAM –o $OUT –G $GTF]. We then used "CuffMerge" [27] to merge all assembled transcriptomes across all samples, and then re-quantified the expression of each transcript isoform in each sample again using StringTie to this global set of transcripts [stringtie $BAM –B –e –o $OUT –G $GTF_ALL] to have expression measurements on the same transcripts across all samples. We then used the "ballgown" tool[28] to merge all assembled and quantified transcripts across all samples (N= 733,339), and used liberal filtering to remove lowly or uniquely expressed transcripts (mean FPKM > 0.025), resulting in 188,578 transcripts across the 495 samples.
5. Expressed regions (ERs) were calculated using the "derfinder" R Bioconductor package[6] using a cutoff of 5 normalized (to 80M reads) read coverage, which identified 389,797 ERs. We retained the 275,885 ERs that were at least 12 basepairs, and annotated the ERs to Ensembl v75.

Genotype data processing

SNP genotyping with HumanHap650Y_V3 (N=135), Human 1M-Duo_V3 (N=357), and Omni5 (N=3) BeadChips (Illumina, San Diego, CA) was carried out according to the manufacturer's instructions with DNA extracted from cerebellar tissue. Genotype data were processed and normalized with the crlmm R/Bioconductor package[29] separately by platform. Genotype imputation was performed on high-quality observed genotypes (removing low quality and rare variants) using the prephasing/imputation stepwise approach implemented in IMPUTE2[30] and Shape-IT[31], with the imputation reference set from the full 1000 Human Genomes Project Phase 3 data set, separately by platform. We retained common SNPs (MAF > 5%) that were present in the majority of samples (missingness < 10%) that were in Hardy Weinberg equilibrium (at $p > 1x10^{-6}$) using the Plink[32] version 1.9 tool kit [`plink --bfile $BFILE --geno 0.1 --maf 0.05  --hwe 0.000001`]. We then identified linkage disequilibrium (LD)-independent SNPs to use in genome-wide clustering of samples and in the number of independent eQTL tests performed [`plink –bfile $BFILE --indep 100 10 1.25`]. Multidimensional scaling (MDS) was performed on the autosomal LD-independent construct genomic ancestry components on each sample, which can be interpreted as quantitative levels of ethnicity – the first component separated the Caucasian and African American samples. This processing and quality control steps resulted in 7,421,423 common variants in this dataset of 495 subjects.

Risk polygene score (RPS) analysis

We computed risk polygene scores (RPS) using the allelic dosage files following imputation described above and the SNPs from provided by the PGC to the Lieber Institute that did not contain completely different clinical subjects used in the GWAS[2]. We considered expression associations at the gene, exon and junction-level to the RPS scores from the first 5 clinical SNP sets, corresponding to GWAS p-value thresholds of $p < 5e-8$ (s1), $p < 1e-6$ (s2), $p < 1e-4$ (s3), $p < 0.001$ (s4), and $p < 0.01$ (s5) – subsequent SNP sets were ignored due to clinical risk plateauing at s5. We also focused only on Caucasian individuals (96 cases, 113 controls), as the s5 RPS was increased in patients relative to controls in this sample ($p=3.2x10^{-5}$), but did not differ among African Americans (p=0.9). Within each expression feature type, we modeled expression levels as a function of each RPS set (s1-s5), adjusting for 3 MDS components of the genotype data, sex, and the first *K* principal components (PCs) of the normalized expression features, where *K* was calculated using the Buja and Eyuboglu permutation-based algorithm[33] in the "sva" Bioconductor package[34]. The resulting p-values of RPS on expression, adjusting for the above factors, were subject to false discovery rate (FDR) control to account for multiple testing.

Public data processing

GTEx: Raw RNA-seq reads from all brain samples with corresponding genotype data were downloaded from SRA and aligned to the genome using TopHat2 [24] (version 2.0.14) using the iGenomes transcriptome and genome annotations based on hg19. As above, featureCounts [25] was used to quantify expression of genes and exons relative to Ensembl v75, and junctions were quantified with regtools[26] as above. We used StringTie with the assembled merged GTF from the LIBD DLPFC samples on the GTEx BAM files to quantify the same transcripts, and used bwtool to quantify the coverage of the same expressed regions from the GTEx brain samples. Genotype data from the two platforms (Illumina Omni 5M and 2.5M) were imputed separately as described above and merged into a single plink[32] set.

GEUVADIS: Raw RNA-seq reads from all LCL samples were downloaded from SRA and aligned to the genome using TopHat2 [24] (version 2.0.9) using the iGenomes transcriptome and genome annotations based on hg19. As above, featureCounts[25] was used to quantify expression of genes and exons relative to Ensembl v75, and junctions were quantified with regtools[26] as above. We used StringTie with the assembled merged GTF from the LIBD DLPFC samples on the GEUVADIS BAM files to quantify the same transcripts, and used bwtool to quantify the coverage of the same expressed regions from the GEUVADIS LCL samples.

CommonMind Consortium (CMC): 547 BAM files were downloaded from Synapse, which were aligned with TopHat2 (version 2.0.9) using Ensembl v70 transcriptome annotation and the hg19 genome. As above, featureCounts [25] was used to quantify expression of genes and exons relative to Ensembl v75, and junctions were quantified with regtools [26] as above. We used StringTie with the assembled merged GTF from the LIBD DLPFC samples on the CMC BAM files to quantify the same transcripts, and used bwtool to quantify the coverage of the same expressed regions from the CMC brain samples. Genotypes were converted to plink file sets from GEN files obtained from Synapse using posterior probabilities > 90%, resulting in genotype data across 9,506,038 SNPs and 547 samples.

Differential expression across brain development

We modeled differential expression across age at each of the five feature summarizations (gene, exon, junction, transcript, and ER) in the 320 control subjects across the lifespan. We modeled expression, after transforming with log2 with an offset of 1, as a function of age after creating using linear splines with breakpoints at ages: birth (0), 1, 10, 20, and 50, further adjusting for sex and ancestry/ethnicity (first 3 MDS components). F-statistics were computed comparing the model containing age (including the linear splines), sex, and ethnicity, to a statistical model with just sex and ethnicity, with corresponding p-values calculated based on an F-distribution with 11 and 308 degrees of freedom, and Bonferroni adjustment within each feature type was performed using the number of features with non-zero expression (gene RPKM > 0.01, exon RPKM > 0,1, and junction RP80M > 0.2 with non-novel annotation) across all samples as the number of tests (which varied by feature type). We also computed post-hoc statistics on the data, including the Pearson correlation between “cleaned” expression (after regressing out the effects of sex and ethnicity, holding the age effects constant), and age to determine if the expression of the fetal rose or fell across the lifespan, and also measured the fetal versus postnatal $\log_2$ fold changes.

Preferential isoform usage across aging was determined by identifying the subset of genes (by Ensembl ID) that contained at least one Bonferroni-significant feature that had positive correlation with age and another Bonferroni-significant feature that had negative correlation with age. We also computed the difference in positive and negative correlations as a measure of the magnitude of the preferential isoform use. Gene set analyses using pre-defined gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) sets were performed using the clusterProfiler R/Bioconductor package[35], here using the genes (mapping from Ensembl to Entrez ID) that had such preferential isoform use to those that were developmentally regulated (having at least one feature that was associate with age at Bonferroni significance). Enrichments with the PGC2 schizophrenia risk loci – defined by the chr:start-end roughly corresponding to linkage disequilibrium blocks in the published manuscript - were performed both parametrically, by overlapping the genomic coordinates of the 108 risk regions with those genes that had preferential isoform usage, compared to a background of all genes with each set of expressed features, as well as by permuting the locations of the 108 regions across the genome 10,000 times and each time, re-computing the overlap within these null regions – see additional details in Jaffe et al 2015[7]. Empirical p-values were calculated by counting the number of the odds ratios across the 10,000 null permutations to each observed odds ratio.

Schizophrenia differential expression analyses

Discovery dataset analysis: we first filtered the subjects with RNA-seq to retain a more stringent set of 155 SCZD cases and 196 controls (criteria: ages between 17-80, gene assignment rate > 0.5, mapping rate > 0.7, RIN > 6, not outlying on 2nd ancestry PC, only self-reported Caucasians and African Americans). We fit three statistical models across each of the expression summarizations, modeling $\log_2$ transformed expression (with an offset of 1) as a function of:

(1) Adjusted ("_adj" suffix in supplementary tables): SCZD diagnosis, adjusting for age, sex, ancestry (SNP PCs 1, 5, 6, 9, 10, which were at least marginally associated with diagnosis), and then observed measures related to RNA quality: RIN, mitochondrial mapping rate, and gene assignment rate.

(2) Adjusted + Quality Surrogate Variables ("_qsva" suffix in supplementary tables): SCZD diagnosis adjusting for "Adjusted" model as well as the first 12 PCs from the degradation matrix (see below) based on polyA+ libraries (selected using to using the BE algorithm [33] in the sva Bioconductor package[34] while providing the adjusted model as input).

(3) Adjusted + Principal Components ("_pca" suffix in supplementary tables): SCZD diagnosis adjusting for "Adjusted" model as well as the first *k* PCs from the expressed features (using the 50000 most variable features) depending on the feature type (gene: 23 PCs, exon: 20 PCs, transcript: 26 PCs, junction: 26 PCs, ERs: 23 PCs).

We used the `lmTest` and `ebayes` functions in the limma Bioconductor package [36] to fit all of the statistical models to estimate $\log_2$ fold changes, moderated T-statistics, and corresponding p-values. Multiple testing correction via the false discovery rate (FDR) was applied using the set of expressed features in this sample set for each summarization type: 24,122 genes (mean RPKM > 0.1), 420,022 exons (mean RPKM > 0.2), 61,950 transcripts (mean FPKM > 0.2), 229,846 junctions (mean RP80M > 1), and the 275,885 ERs.

RNA quality correction: We summarize the RNA quality correction approach here – for more detail, see the companion paper by Jaffe et al 2016 [15]. Briefly, the quality surrogate variable analysis (qSVA) uses RNA sequencing data generated from five DLPFC tissue samples left unfrozen for 0, 15, 30 and 60 minutes, resulting in 20 RNA samples. These samples were sequenced with both polyA+ and RiboZero library preparations, and gene, exon and junction counts were derived as above. We utilized the gene-level effects of degradation in these data in Figure S5 to demonstrate residual confounding by RNA quality, which we call the “DEQual Plot”.

For a given preparation type, we identified the genomic regions most susceptible to degradation by correlating coverage at expressed regions [6] to degradation time, adjusting for donor. This statistical modeling identified 515 regions significantly susceptible to degradation (at Bonferroni significance) in the RiboZero libraries and the top 1000 regions most susceptible to degradation (among the 35,287 at Bonferroni significance) in the polyA+ libraries – the BED files for these degradation-susceptible regions are available at: https://github.com/nellore/region_matrix.

The algorithm then involves selecting the set of regions for a particular library type and calculating total coverage within each region in the new user-provided samples (e.g. the 495 DLPFC RNA-seq polyA+ samples) to form the degradation matrix (which is either 515 or 1000 rows by N samples). Then PCA is performed on the log2 transformed degradation matrix (with an offset of 1) and the top *K* PCs are selected, for example using the BE algorithm [33], and extracted – the set of these PCs are referred to as quality surrogate variables (qSVs), and are included as adjustment variables in subsequent differential expression analyses.

Replication dataset analysis: we performed analogous sample selection procedures as in the discovery dataset to select 159 patients and 172 controls (total gene assignment rate > 0.3, alignment rate > 0.8, RIN > 6, ages between 18-80, non-outlying on genetic ancestry PCs 3 and 5 and keeping only reported Caucasians and African Americans). We similarly fit the three sets of statistical models to all five feature summarizations, with the following differences compared to the discovery analysis:

(1) Adjusted model: the model here was diagnosis adjusting for age, sex, race, brain bank, RIN, gene assignment rate, alignment rate.

(2) qSVA model: the degradation matrix was constructed using the 515 regions based on the RiboZero libraries in the degradation experiment.

(3) PC adjustment: for each feature summarization type, we included: 27 PCs for genes, 29 PCs for exons, 39 PCs for transcripts, 39 PCs for junctions, and 33 PCs for ERs.

In these replication data we did not perform FDR correction. We were using the study for replication, not discovery, and therefore only used the features that were expressed in our data regardless of the expression levels in CMC. We considered features independently replicated if they had the same directionality for the SCZD versus control $\log_2$ fold change and were marginally significant (at $p < 0.05$) in the CMC dataset.

Gene set analyses on replicated differentially expressed features and genes were performed with clusterProfiler[35] as described above. Set-level analyses on features in the GWAS risk regions were conducted by assigning each expressed feature a binary variable for whether it was in the risk regions or not. Then we fit a linear regression model of the t-statistics for diagnosis, adjusted by the qSVA approach, as a function as whether the feature was in the risk region, adjusting for its average expression level. This analysis was conducted across and then within each of the five feature summarization types.

eQTL discovery analyses

We performed eQTL analyses separately by feature type (gene, exon, junction, transcript, and ER) allowing for a 500kb window around each of the 7,421,423 common SNPs in the 412 age > 13 samples, adjusting for ancestry (first three MDS components from the genotype data), sex, diagnosis, and the first *K* principal components (PCs) of the normalized expression features, where *K* was calculated separately by feature type using the Buja and Eyuboglu permutation-based algorithm[33] in the "sva" Bioconductor package[34] (gene: 22 PCs, exon: 19 PCs, junction: 26 PCs, transcript: 25 PCs, expressed regions: 20 PCs). The eQTL analyses were run using the MatrixEQTL R package[37], which returned the $\log_2$ fold change per allele copy, and corresponding T-statistic, p-value, and FDR for each SNP-feature pair. We further used the LD-independent SNPs to estimate the effective number of tests (by counting the number of features within a 500kb window around each LD independent SNP) for a more conservative Bonferroni adjustment. For all five feature types, we retained all eQTLs with FDR < 1%. We retained the best SNP for each expression feature for more detailed annotation and we calculated post-hoc statistics separately by reported race (Caucasian or African American) using corresponding subsets of the same design matrix.

eQTL replication analyses

We sought to replicate the best SNP-Feature pair for each eQTL in two independent datasets: CommonMind Consortium and the GTEx project. As not every single SNP imputed in the discovery dataset was present in each of CMC and GTEx, we retained the most highly ranked (by eQTL p-value) SNP that was present in each dataset, and tested these SNP-Feature pairs for replication.

CommonMind Consortium: for the ranked eQTL lists for each feature type (gene, exon, junction, transcript, and ER), we retained the best discovery SNP present in the replication dataset. The majority of these tested SNP-Feature pairs used the SNP with the best discovery p-value. For example, among 1,812,008 FDR-significant SNP-Gene pairs, 1,539,218 pairs were present in CMC (84.9%), and among the 18,394 genes with at least one FDR-significant SNP, 14,928 (81.1%) of the top ranked SNPs were present, and therefore used for replication. Another 2,279 genes had FDR-significant SNPs more lowly ranked in the discovery set that were tested for replication, and 1,187 genes did not have any FDR-significant SNPs present in CMC. The proportions of present top-ranked SNPs retained for replication were similar across the different feature summarizations: gene=81.1%, exon=82.1%, junction=81.5%, ER=80.6%, transcript=79.4%. The number of tested SNP-feature pairs across all expression summarizations is provided in Table S9.

We calculated the corresponding principal components (PCs) across all expression features separately in the 547 samples, as in the discovery data, and in our tests for replication, modeled the log2-transformed expression data as a function of additive genotype effect, controlling for the top 10 expression PCs per summary type and the first 3 ancestry MDS components from the genetic data. We lastly used the allele information to ensure additive effects across study were relative to the same allele, flipping the directionality of such effects where the major allele in one study corresponded to the minor allele in the other study. The eQTLs that were directionality consistent (same allelic directionality of the additive effect on expression) and had marginal significance (p<0.01) were considered replicated and retained for summarization.

GTEx: we performed analogous replication analyses as described above across the 1184 GTEx brain samples that had genotype and RNA-seq data. While we calculated eQTL replication statistics separately by brain region, we adjusted for the first 10 expression PCs from the entire sample (across all brain regions), as well as MDS ancestry components from across the genome. We only considered eQTLs from the frontal cortex brain region for replication, enforcing directional consistency as the data set was much smaller than both our discovery and first replication dataset (N=99).

After computing replication statistics in the CMC and GTEx frontal cortex datasets, the eQTLs that were directionality consistent (same allelic directionality of the additive effect on expression) in both datasets and had marginal significance (p<0.01) in CommonMind were considered replicated and retained for summarization (see Table S9). Factors positively influencing replication were, as expected, the statistical magnitude of the eQTL effect (absolute T-statistic), the minor allele frequency, and the mean expression level.

Clinical enrichment analyses:

We first calculated the "risk" effects of the schizophrenia GWAS variants using the set of all GWAS results from: https://www.med.unc.edu/pgc/files/resultfiles/scz2.snp.results.txt.gz and identified which allele at each SNP corresponded to increased risk, e.g. a positive odds ratio. We then used these risk alleles to convert the directionality of our eQTL effects to reflect the reference allele being the risk allele (rather than the minor allele). Lastly, we incorporated the qSV-adjusted case-control fold changes to identify the genetic and clinical directional consistent features. Enrichment analyses for fetal effects, isoform shift genes, and directional consistency was performed using Chi-squared tests with 1 df using the `chisq.test()` in R.

References

2 Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-+, doi:10.1038/nature13595 (2014).

4 Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **33**, 290-295, doi:10.1038/nbt.3122 (2015).

6 Collado Torres, L. *et al.* Flexible expressed region analysis for RNA-seq with derfinder. *Nucleic Acids Research* **In Press.**, doi:10.1101/015370 (2016).

7 Jaffe, A. E. *et al.* Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nat Neurosci* **18**, 154-161, doi:10.1038/nn.3898 (2015).

15 Jaffe, A. E. *et al.* A framework for RNA quality correction in differential expression analysis. *bioRxiv*, doi:10.1101/074245 (2016).

23 Lipska, B. K. *et al.* Critical factors in gene expression in postmortem human brain: Focus on studies in schizophrenia. *Biol Psychiatry* **60**, 650-658, doi:10.1016/j.biopsych.2006.06.019 (2006).

24 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36, doi:10.1186/gb-2013-14-4-r36 (2013).

25 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2014).

26 regtools v. 0.1.0 (2016).

27 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**, 511-515, doi:10.1038/nbt.1621 (2010).

28 Frazee, A. C. *et al.* Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nature biotechnology* **33**, 243-246, doi:10.1038/nbt.3172 (2015).

29 Scharpf, R. B., Irizarry, R. A., Ritchie, M. E., Carvalho, B. & Ruczinski, I. Using the R Package crlmm for Genotyping and Copy Number Estimation. *J Stat Softw* **40**, 1-32 (2011).

30 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5**, e1000529, doi:10.1371/journal.pgen.1000529 (2009).

31 Delaneau, O., Coulonges, C. & Zagury, J. F. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC bioinformatics* **9**, 540, doi:10.1186/1471-2105-9-540 (2008).

32 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559-575, doi:10.1086/519795 (2007).

33 Buja, A. & Eyuboglu, N. Remarks on Parallel Analysis. *Multivariate Behavioral Research* **27**, 509-540, doi:10.1207/s15327906mbr2704_2 (1992).

34 Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882-883, doi:10.1093/bioinformatics/bts034 (2012).

35 Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology* **16**, 284-287, doi:10.1089/omi.2011.0118 (2012).

36 Smyth, G. K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* **3**, Article 3 (2004).

37 Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-1358, doi:10.1093/bioinformatics/bts163 (2012).

Author information: sequencing reads and genotype data are available through SRA and dbGaP at accession numbers: [**TBD].** Reprints and permissions information is available at www.nature.com/reprints**.** Correspondence and requests for materials should be addressed to Andrew Jaffe (andrew.jaffe@libd.org). The following authors have competing interests:

Tony Kam-Thong is employed by F. Hoffmann-La Roche

Hualin S Xi and Jie Quan are employees of Pfizer Inc.

Alan Cross, and Nicholas J.Brandon were full time employees and shareholders in AstraZeneca at the time these studies were conducted.

The remaining authors declare no competing financial interests.

Supplementary Information

Supplementary Figure Legends:

Figure S1: Study overview and cartoon describing quantifying the five different expression summarizations.

Figure S2: Cartoon describing the four different splice junction annotation classes, relative to annotated exons (dark blue rectangles). (A) Annotated splice junctions map between two exons in a known transcript. (B) Exon-skipping splice junctions map to two annotated exons in different transcripts. (C) Alternative start/exon junctions map to only one annotated exon on either the 5' or 3' end. (D) Completely novel junction do not map to any known exon.

Figure S3: Venn diagram of developmentally regulated features mapped back to Ensembl Gene IDs by the five feature summarization methods. DER: differentially expressed region.

Figure S4: Venn diagram of Ensembl Gene IDs that contain significant isoform shifts by the four feature summarization methods that allow for multiple features per gene. DER: differentially expressed region.

Figure S5: Degradation quality (DEqual) plots for diagnosis-associated differential expression analyses at the gene level. Y-axis: T-statistic for the effect of RNA degradation on each gene, X-axis: diagnosis T-statistic, adjusting for (left) nothing/univariate effect, (middle) observed clinical and technical confounders and (right) the inclusion of the quality surrogate variables.

Figure S6: Replication rates for three differential expression statistical models. X-axis: discovery p-value, Y-axis: replication rate, e.g. the proportion of features that have directional consistency and are marginally significant (at p<0.05) in independent data. Point size is proportion to the number of features called significant at the indicated p-value threshold.

Figure S7: Scatter plot of effect sizes (fold changes) in discovery and replication datasets for those features significant and replicated. Colors have the same legend as Figure 3A.

Figure S8: GWAS loci set-level analysis for (A) all features together and then stratified by only (B) genes, (C) exons, (D) junctions, (E) transcripts and (F) expressed regions. P-values were based on the Wilcoxon rank sign test.

Figure S9: Boxplots of single gene eQTL effects for GWAS associated SNPs for the associations not featured in Figure 4.

Supplementary Table Legends:

Table S1: Demographic information for subjects in the present study, stratified by age and diagnosis group. Dx: diagnosis, N: sample size, F: Female, Cauc: Caucasian, SD: standard deviation, PCW: post-conception weeks. Antipsychotic use was measured using toxicology at

time of death. P-values for diagnosis differences in continuous variables are based on linear regression and P-values for categorical variables are based on chi-squared tests.

Table S2: Splice junction annotation and characterization in GTEx and GEUVADIS for any junction or highly expressed junctions (mean reads per 80M mapped reads, RP80M > 0 and 5). Each column represents a 2x2 table for presence of identified junctions in 495 DLPFC samples in two independent polyA+ datasets. This table is analogous to Table 1.

Table S3: Summary statistics for those features significantly developmentally regulated in the control-only analyses across the lifespan.

Table S4: Significant developmentally regulated features collapsed to Ensembl Gene ID, used to make Figure S3

Table S5: Isoform shifts by Ensembl Gene ID and feature summarization type.

Table S6: Gene set analyses for those genes with significant isoform shift, stratified by feature summarization type. Q-values, which control the false discovery rate, FDR, are shown.

Table S7: Genes within the PGC schizophrenia GWAS risk regions that contain isoform shifts by feature summarization type. 21.8% of PGC2 genes had developmental isoform shifts using exon counts (N=96/440) and 31.9% showed this isoform shift association based on junction counts (N=137/430)

Table S8: eQTL replication metrics for CommonMind Consortium, GTEx frontal cortex, and both among all significant eQTLs in the discovery adult sample.

Table S9: eQTL statistics for the best SNP-feature pair in the discovery sample that replicated in CMC and GTEx frontal cortex, by feature type.

Table S10: Differential expression statistics for those features that were significant and replicated in case-control comparisons.

Table S11: Genes consistently differentially expressed by case-control analysis for the different feature summarizations.

Table S12: Associations between diagnosis, RPS and expression at gene and exon levels. First two columns for each feature: p-values for gene set tests for the significant case-control features among statistics capturing the effect of RPS on expression. Second two columns for each feature: directionality between RPS on expression associations and diagnosis on expression associations.

Table S13: Gene set analysis for genes with features differentially expressed by case-control status, stratified by directionality and feature summarization type.

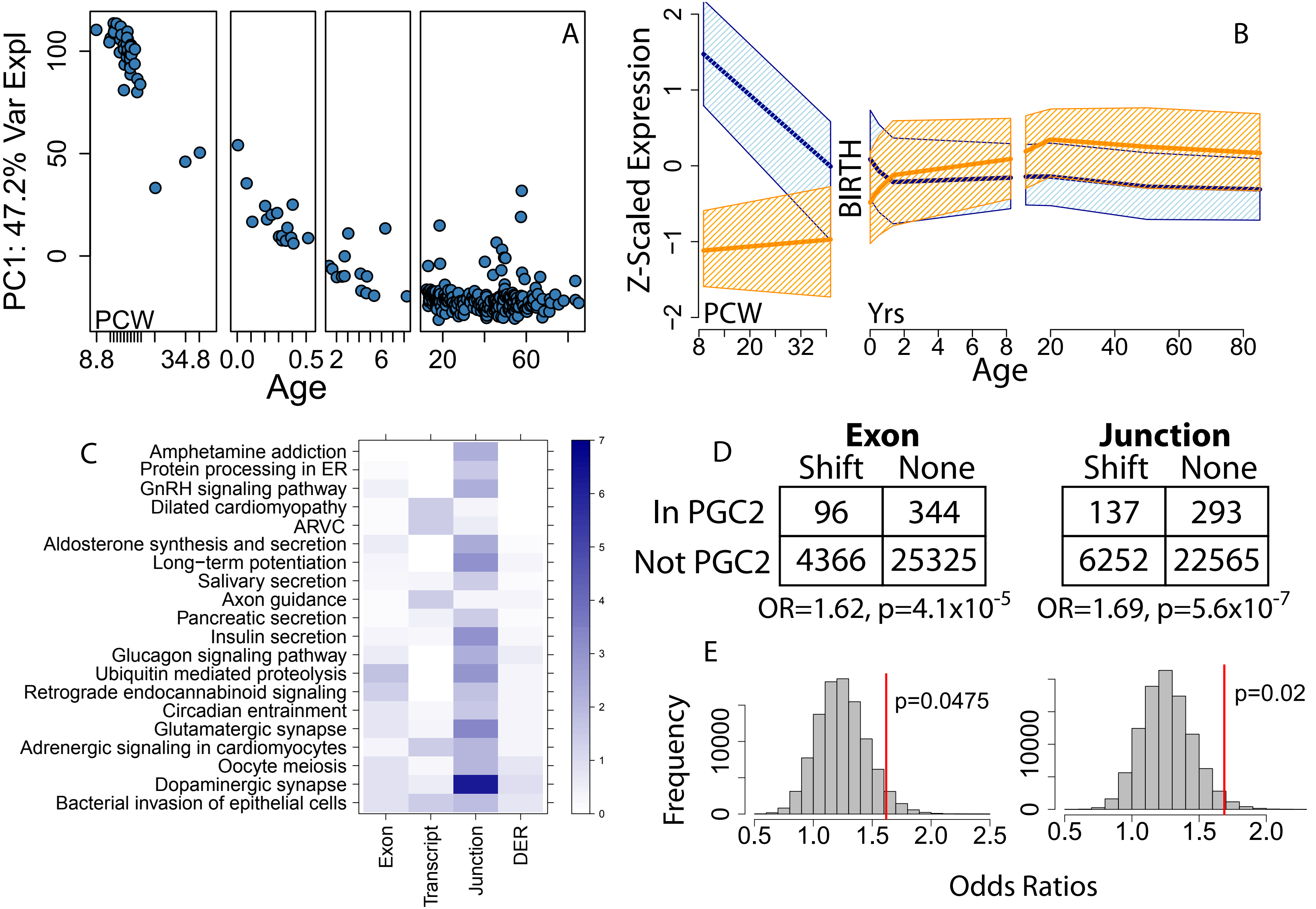
Table S14: GWAS region set-level analyses for diagnosis-associated differentially expressed features, testing whether features in the PGC risk loci were more or less expressed as a set in

cases compared to controls. Qual: qSVA adjusted analysis, Adj: observed covariate adjusted analysis.

Table S15: eQTL statistics for all PGC SCZD GWAS variants, including the risk direction (eQTL and GWAS) and clinical direction (case-control fold change)

Table S16: eQTL statistics filtered to those that were directionally consistent between risk and clinical directionalities

Table S17: Analogous to Table 2 with Ensembl transcript annotation data to highlight transcript specificity.



Exon

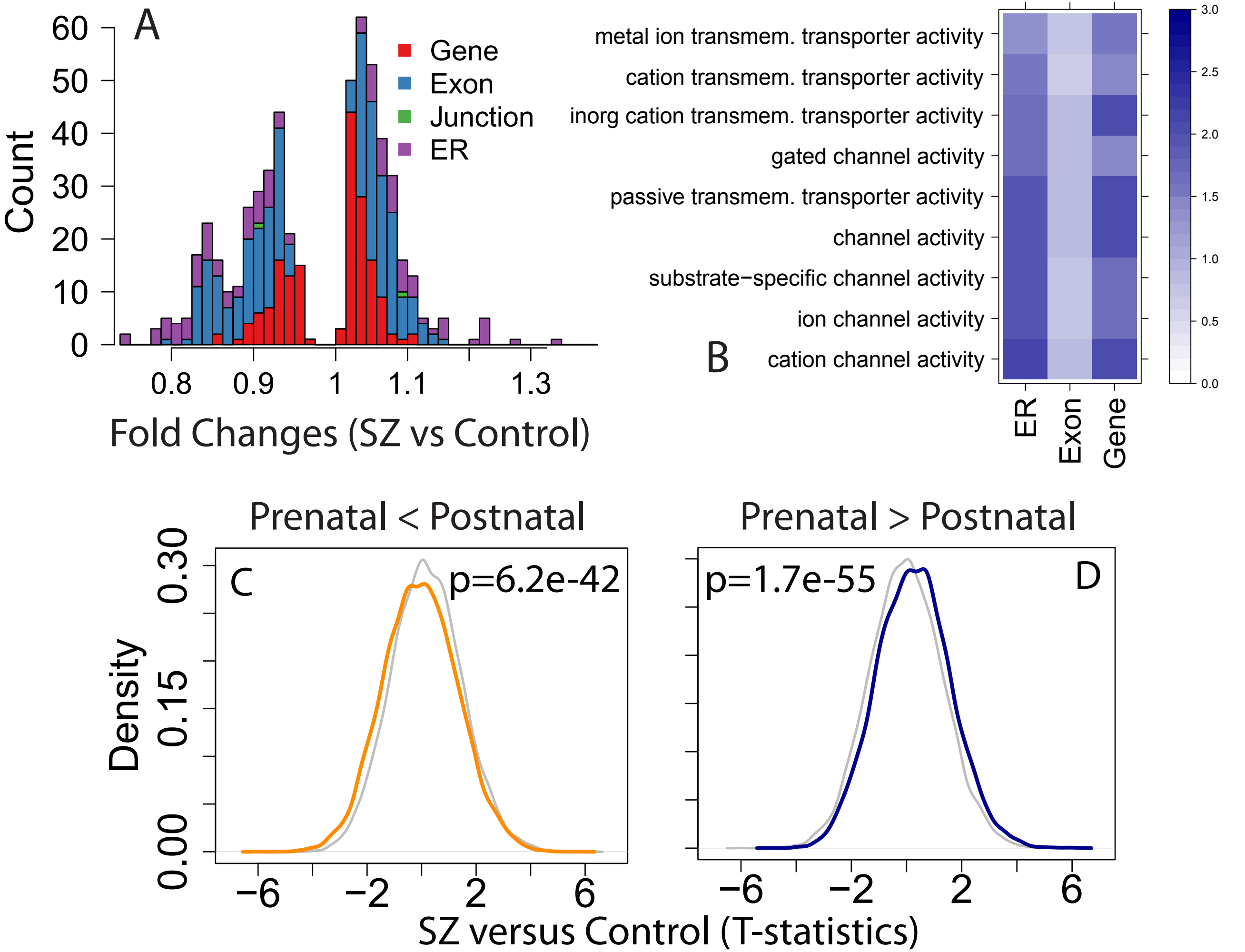
	Shift	None
In PGC2	96	344
Not PGC2	4366	25325

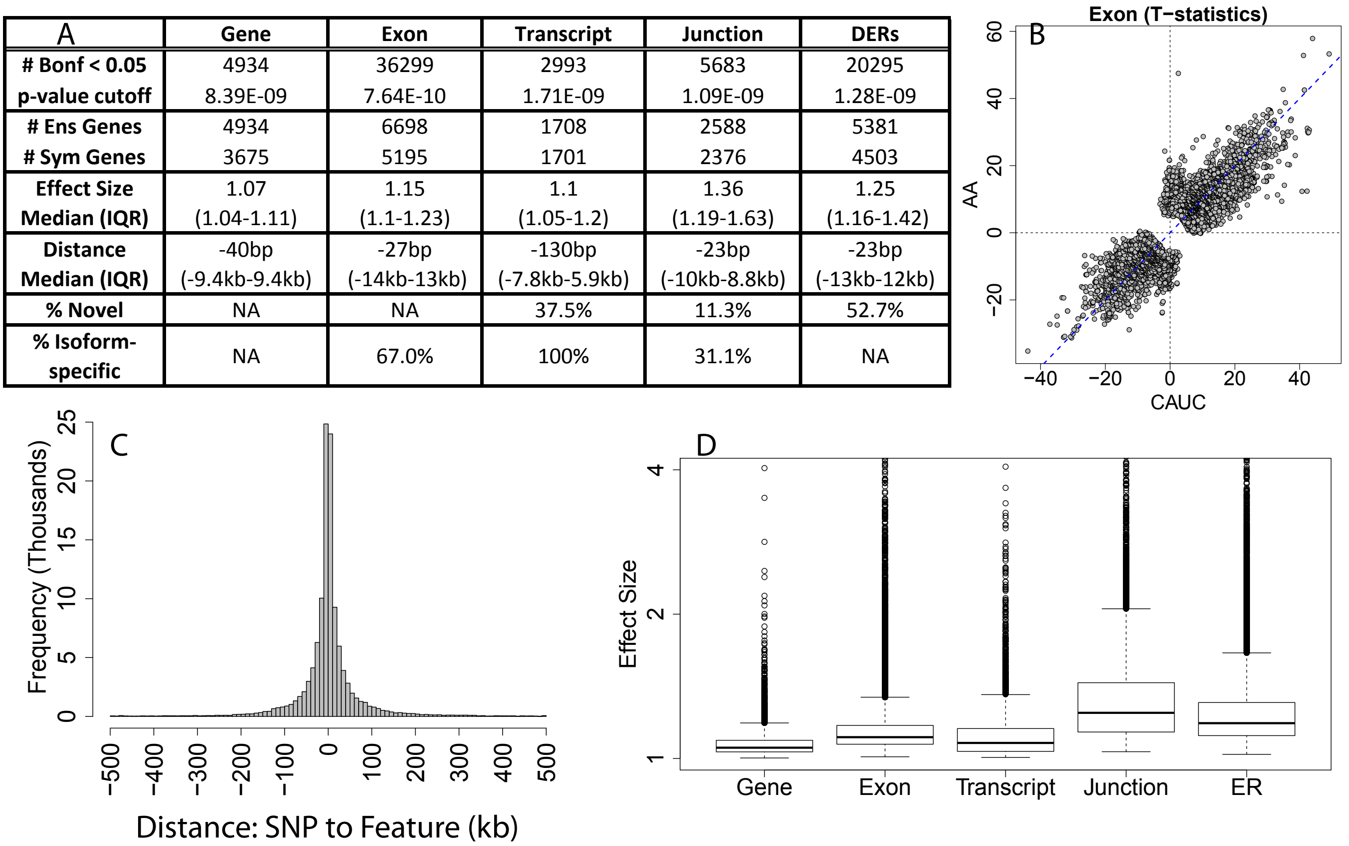
OR=1.62, $p=4.1\text{x}10^{-5}$

Junction

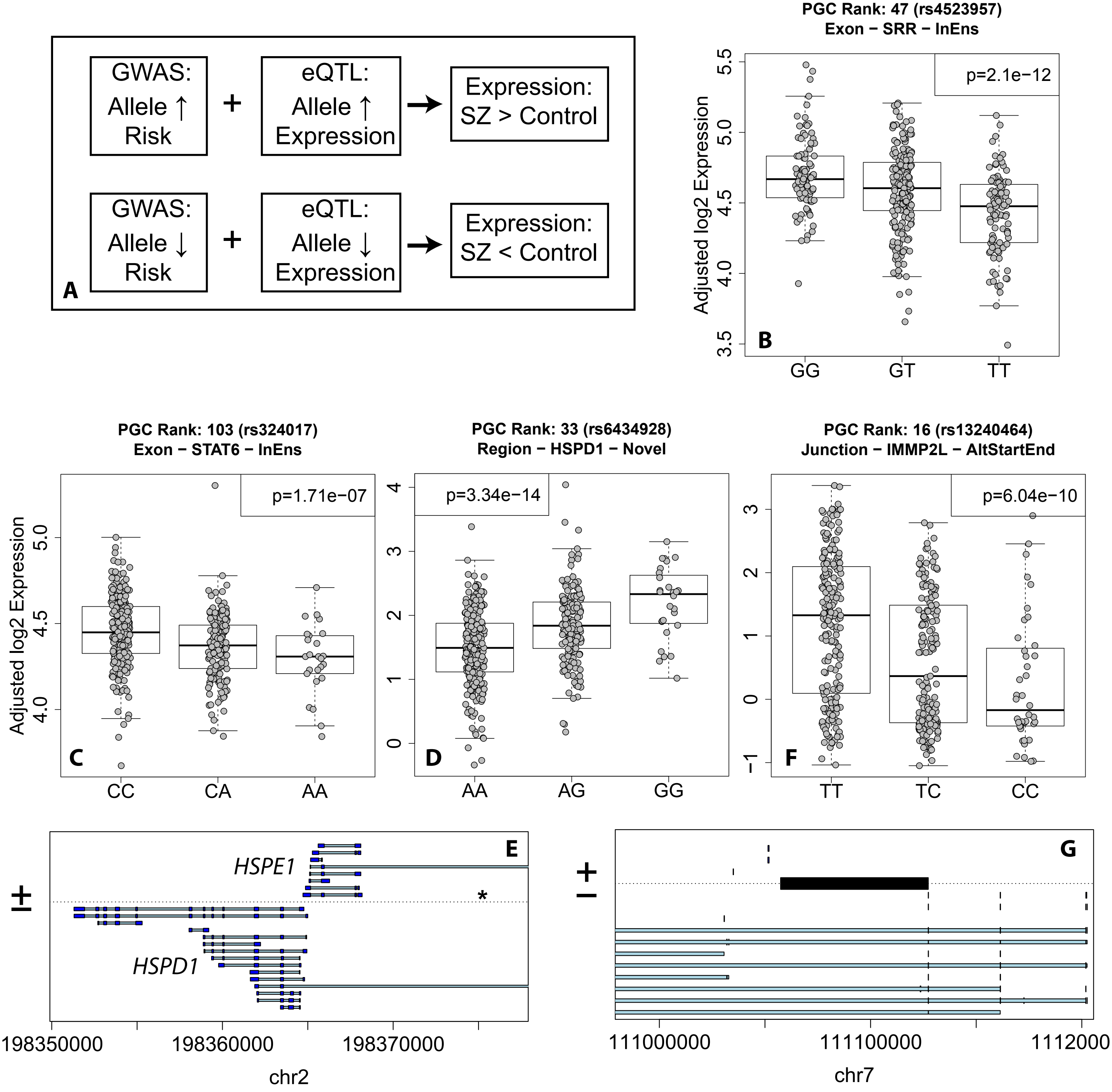
	Shift	None
In PGC2	137	293
Not PGC2	6252	22565

OR=1.69, $p=5.6\text{x}10^{-7}$


A
Count
60
50
40
30
20
10
0
Gene
Exon
Junction
ER
0.8
0.9
1
1.1
1.3
Fold Changes (SZ vs Control)
metal ion transmem. transporter activity
cation transmem. transporter activity
inorg cation transmem. transporter activity
gated channel activity
passive transmem. transporter activity
channel activity
substrate−specific channel activity
ion channel activity
cation channel activity
B
ER
Exon
Gene
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Prenatal < Postnatal
Prenatal > Postnatal
C
p=6.2e-42
p=1.7e-55
D
Density
0.00
0.15
0.30
−6
−2
2
6
−6
−2
2
6
SZ versus Control (T-statistics)



A	Gene	Exon	Transcript	Junction	DERs
# Bonf < 0.05	4934	36299	2993	5683	20295
p-value cutoff	8.39E-09	7.64E-10	1.71E-09	1.09E-09	1.28E-09
# Ens Genes	4934	6698	1708	2588	5381
# Sym Genes	3675	5195	1701	2376	4503
Effect Size	1.07	1.15	1.1	1.36	1.25
Median (IQR)	(1.04-1.11)	(1.1-1.23)	(1.05-1.2)	(1.19-1.63)	(1.16-1.42)
Distance	-40bp	-27bp	-130bp	-23bp	-23bp
Median (IQR)	(-9.4kb-9.4kb)	(-14kb-13kb)	(-7.8kb-5.9kb)	(-10kb-8.8kb)	(-13kb-12kb)
% Novel	NA	NA	37.5%	11.3%	52.7%
% Isoform-specific	NA	67.0%	100%	31.1%	NA


GWAS:
Allele ↑
Risk
+
eQTL:
Allele ↑
Expression
Expression:
SZ > Control
GWAS:
Allele ↓
Risk
+
eQTL:
Allele ↓
Expression
Expression:
SZ < Control
A
PGC Rank: 47 (rs4523957)
Exon − SRR − InEns
p=2.1e−12
Adjusted log2 Expression
5.5
5.0
4.5
4.0
3.5
B
GG
GT
TT
PGC Rank: 103 (rs324017)
Exon − STAT6 − InEns
p=1.71e−07
Adjusted log2 Expression
5.0
4.5
4.0
C
CC
CA
AA
PGC Rank: 33 (rs6434928)
Region − HSPD1 − Novel
p=3.34e−14
4
3
2
1
0
D
AA
AG
GG
PGC Rank: 16 (rs13240464)
Junction − IMMP2L − AltStartEnd
p=6.04e−10
3
2
1
0
−1
F
TT
TC
CC
E
HSPE1
*
HSPD1
+
−
198350000
198360000
198370000
chr2
G
+
−
111000000
111100000
1112000
chr7

Supplementary Results for: Developmental and genetic regulation of the human frontal cortex transcriptome in schizophrenia

Andrew E Jaffe, Richard E Straub, Joo Heon Shin, Ran Tao, Yuan Gao, Leonardo Collado Torres, Tony Kam-Thong, Hualin S Xi, Jie Quan, Carlo Colantuoni, Qiang Chen, Bill Ulrich, Brady J. Maher, Amy Deep-Soboslay, The BrainSeq Consortium, Alan Cross, Nicholas J. Brandon, Jeffrey T Leek, Thomas M. Hyde, Joel E. Kleinman, Daniel R Weinberger

1. Deep characterization of the human cortex transcriptome

We used Ensembl v75 annotation to quantify the normalized expression (via reads per kilobase per million reads mapped, RPKM) of 615,410 exons across 63,677 genes, and guided transcript assembly within and across samples to identify and quantify 188,578 transcripts with at least modest expression across all of the samples[1]. Only approximately half of the assembled transcripts were found in Ensembl annotation (N=100,932, 53.5%), while 41.8% (N=78,751) consisted of transcripts with novel splicing patterns. We also more directly interrogated the spliced alignments, which supported 3,582,199 unique exon-exon splice junctions across the 495 samples. We defined four classes of splice junctions based on the Ensembl gene/transcript annotation – present/existing, exon skipping, alternative exonic boundary, and completely novel (Figure S2, Table S2), and found moderate expression of many potentially unannotated transcripts in brain. To assess the replication and brain specificity of our findings, we identified junctions in both the brain samples from the GTEx project (N=1,393 samples from 201 individuals) and lymphoblastoid cell lines (LCLs) from the GEUVADIS project (N=666, see Methods). Overall, we find extensive replication in human brain GTEx samples of annotated junctions (95.0%), and high replication of potentially novel transcripts that capture exon skipping (75.6%) or alternative exonic boundaries (65.8%), with extremely high replication (>95%, Table 1). Much of this novel transcriptional activity appeared relatively-brain specific, as few junctions identified in DLPFC were present only in GEUVADIS LCLs and not GTEx brain samples further suggesting that this degree of junction discovery was not largely driven by alignment software. We lastly defined “expressed regions” (ERs) of contiguously expressed sequence – these ERs and junctions together can “tag” elements of transcripts in the data that are not constrained by limitations or incompleteness of existing annotation.

2. Developmental regulation of human brain transcription

We sought in this study to more fully characterize developmental regulation of transcription across human brain development and aging, and modeled dynamic expression in 320 control samples using flexible linear splines in each of the five expression summarizations (see Methods). We found widespread developmental regulation across all five features(Table S3) corresponding to 30,490 unique Ensembl genes (across 20,596 gene symbols), including a core set of 15,603 genes with all five features showing convergent expression association with development/aging (Figure S3). There were 89,865 previously unannotated expressed features corresponding to the majority of genes (17,374 IDs and 15,261 symbols) with genome-wide significant dynamic expression across brain development in these unaffected subjects (Table S4). The majority of these unannotated features were identified using the splice junction counts (N=43,512, 48.4%) of which 31,183 junctions (71.7%) tagged alternative exonic boundaries and 12,329 junctions (28.3%) corresponding to exonic skipping, while there were 26,010 differentially expressed regions (DERs, 28.9%) that all correspond to alternative exonic boundaries and 20,343 transcripts of which 19,138 (94.1%) correspond to exonic skipping.

3. Expression associations with chronic schizophrenia illness

Given the apparent lower RNA quality of patient samples in virtually all schizophrenia brain case control studies including our own (see Table S1), we filtered the 384 samples above age 17 (209 controls and 175 patients) to a subset of 351 higher quality samples (196 controls and 155 cases) using metrics of age, ancestry, RNA quality, and cryptic ancestry (see Methods). Even after filtering, using gene-level expression, univariate comparisons between SCZD patients and controls identified 12,686 genes differentially expressed (DE) at a false discovery rate (FDR) < 5%, suggesting a large degree of bias. Regression modeling, adjusting for age, sex, ancestry, and observed RNA quality (see Methods) reduced the extent of confounding, resulting in 1,988 genes DE between patients and controls at FDR < 5%. We identified the largest number of significant and replicated DE features using exon counts (N=274) followed by gene counts (N=170) and expressed region counts (N=110). Very few DE junctions (N=2) and no DE transcripts were identified and independently replicated, which likely highlight the decreased statistical power using these approaches when annotated features are differentially expressed (see Discussion). Approximately one half of the genes annotated by these DE features were identified using gene-level summarization only (86/170, 50.6%), which utilize the largest number of reads collapsed across all possible transcript isoforms (Table S11). The majority of genes with DE signal were only supported by a single summarization type across both Ensembl IDs (142/237, 59.9%) and gene symbols (119/209 56.9%, Table S11).

Interestingly, analogous analyses for developmental regulation of schizophrenia-associated features without adjusting for the RNA quality qSVs were significant in the opposite directions, namely that schizophrenia-associated changes were further from, rather than closer to, fetal expression levels, which would be predicted as an effect of residual RNA quality confounding (as the quality of the samples were ranked fetal > adult control > adult SZ, see Table S1). This directional difference in RNA quality between fetal and postnatal samples is also typical of earlier studies (e.g. Colantuoni, Geschwind, etc).

4. GWAS implications of illness-associated expression differences

Only two genes in the 108 significant schizophrenia GWAS loci (*KLC1* and *PPP2R3A*) had expression features significantly differentially expressed and replicated. To potentially identify more subtle associations within these GWAS risk regions, we performed an exploratory “feature set” analysis (analogous to traditional gene set analyses) comparing the schizophrenia differential expression statistics of all expressed features within the loci to those outside the loci. We found decreased expression of the 15,050 expressed features within the risk loci in patients compared to controls, relative to the 996,775 expressed features outside of the loci, adjusting for mean expression level (Figure S8A, p =3.32×10^{-36}, Table S13). While the absolute set-level effect sizes were small, these GWAS region-level associations were largely consistent when stratified by summarization type, with the exception of transcript counts which showed no association. Of particular importance is the observation that these findings also were only significant after quality adjustment in the differential expression analysis (Figures S8B-F) – analyses adjusting for only the usual observed confounders (e.g. clinical covariates and observed quality variable) showed no enrichment among PGC risk regions (p=0.2, Table S13). Therefore, while many of the case-control expression differences that meet both genome-wide statistical significance and replication criteria may be related to schizophrenia treatment and epiphenomena and depleted for GWAS regions, some differences in expression in some subsets of patient populations might be related to genetic risk for the disorder, though the biological interpretation of the feature distribution differences is not clear. Larger studies can likely improve power to detect expression changes within the GWAS risk regions amidst the clinical and molecular heterogeneity of schizophrenia.

5. Genetic control of RNA expression levels in the cortex

In order to elucidate the RNA features associated with schizophrenia risk variants themselves, rather than LD regions, we first performed a *cis* (<500kb) genome-wide expression quantitative trait loci (eQTL) analysis across the five expression feature summarizations. These analyses were performed within the post-adolescent subjects, including both cases and controls (age > 13, N=412) across densely observed and imputed common and high-quality SNP genotypes (MAF > 5%, N=7,421,423), adjusting for components related to ancestry and expression heterogeneity in each feature type (see Methods). Given the large number of tests, but also the need for comparability across the five expression summarizations, we adjusted for multiple testing using both the false discovery rate (FDR) with a cutoff of 1%, in line with the GTEX project [2] and also a more stringent Bonferroni-based cutoff adjusting for the number of LD-independent SNP-feature pairs. These multiple testing thresholds corresponded to $p<1.0 \times 10^{-4}$ and $p<2.6 \times 10^{-9}$, respectively, across the 5 feature summarizations.

We examined only eQTL associations of 682,828 unique genetic variants to a total of 70,204 nearby expression features, mapping back to 9,271 Ensembl IDs (of which 7,057 had HGNC symbols) that were at Bonferroni significance in our dataset and that replicated in the two independent datasets (Figure 3A, Table S10). We highlight the best SNP for each feature in Table S11, and have a complete database, including statistics and visualizations, of all significant and replicated eQTLs online (as well as significant control-only eQTLs) via a user friendly browser at http://eqtl.brainseq.org.The majority of eQTLs (N=417,276, 61.1%) associated with features in a single Ensembl gene, and another large subset associated with features in two nearby genes (N=133,286, 19.5%). Conversely, the majority of genes were associated with more than 1 LD-independent SNP (76.3%) across the feature summarizations, including approximately half (49.0%) of genes having features associated with more than 3 LD-independent SNPs, suggesting that multiple variants can associate with expression independently of each other, and potentially result in more normally-distributed expression levels at the population-level. Furthermore, the eQTL effects (discovered in multiple/mixed ancestries) were largely consistent between Caucasian and African American subjects (Figure 2B). The most significant SNP for each eQTL feature was extremely proximal to the feature (based on nearest coordinate), with a median < 50bp (IQR less than -15kb to 15kb, Figure 2C). Lastly, we found the largest effect sizes on average for the junction and ER eQTLs (see Supplementary Results, Figure 2D).

While the majority of genes with features as eQTLs have multiple possible Ensembl isoforms, 67.0% of exon eQTLs and 31.1% of junction eQTLs were specific to a single Ensembl transcript when multiple transcripts existed for a given gene. Furthermore, 640 (11.2%), 1,123 (37.5%) and 10,698 (52.7%) of replicated junction, transcript, and expressed region eQTLs respectively, corresponded to novel transcriptional activity, including 387 alternative exonic boundary and 253 exon-skipping junction eQTLs and 4,036 alternative exonic boundary (crossing an intron and exon, and extended UTRs) and 6,662 intronic and intergenic ERs. These results indicate that performing both exon- and junction-level eQTL analyses together provide an encompassing view of genetic regulation of expression, as exon-level analyses identified eQTLs for the most genes as well as transcript-specific events, and junction-level analysis provided the largest biological effect sizes and the ability to identify potentially novel transcript isoforms. Together these eQTL results, based on analyses across multiple feature summarizations from RNA sequencing data, demonstrate more widespread genetic regulation of expression than previously reported, including extensive transcript specificity and regulation of unannotated sequence in the human brain.

There was moderate concordance across the different feature summarizations. Exon-level summarizations resulted in the largest number of unique genes as eQTLs (N=6,698) and transcript-level summarizations resulted in the fewest (N=1,708). While the junction eQTLs were less likely to appear in the GTEX project likely because of the lower coverage, and therefore replicate, those that did replicate tended to have larger effect sizes than other summarizations, with a median 36% (IQR: 19-63%) change in expression per allele copy (Figure 2D). Expressed region-level analysis identified the second most eQTLs to unique genes (N=5,381) and effect sizes (median: 25%, IQR: 16-42%). Gene-level eQTL analysis, which is the de-facto approach in RNA-seq analysis, resulted in the smallest eQTL effects, with a 7% (IQR: 4-11%) change in expression per allele copy, and identified fewer unique genes than exon and expressed region analysis (N=4,934). These data suggest that most eQTLs involve specific transcript isoforms whose signals may be diluted when examining only gene-level expression data. If this dilution when using gene counts results from averaging over potentially transcript-specific eQTLs, then mapping all expressed features (sans expressed regions) to all known Ensembl transcripts, and tabulating the proportion of eQTLs that appear relatively transcript specific should confirm this.

In fact, there was extensive evidence of transcript-specificity of replicated eQTLs across exons, junctions, and transcripts.

6. Characterization of schizophrenia risk-associated expressed features

We found significant enrichment of this directional consistency between the GWAS, case-control, and FDR-significant (34 loci across 444 eQTL features and 91 Ensembl genes, OR=2.00, $p=4.2x10^{-6}$) as well as with the more conservative Bonferroni-significant (19 loci across 161 features and 44 Ensembl genes, OR=4.67, $p=7.97x10^{-8}$) features

The majority of our eQTLs (SNP-feature pairs) had directionally consistent allelic effects in the 50 fetal samples (N=628, 82.2%, $p < 10^{-20}$) and a subset were marginally significant (at $p < 0.05$) even in this small sample size (N=162, 21.2%). As further evidence for similar genetic regulation, every eQTL that was marginally significant in the fetal samples was directionally consistent with the eQTL effect in the adult samples. For example, we find the most significant and replicated association with clinical directional consistency between rs4523957 (PGC rank 47) and exons (as well as other feature types, see Table S15) in Serine Racemase (*SRR*, Figure 4B), an NMDAR modulator previously associated with schizophrenia[3]. While the statistical significance of exon associations were only slightly better than at the gene level ($2.1x10^{-12}$ versus $5.44x10^{-11}$), the effect size was almost three times larger (-0.16 versus -0.06). Similarly we identified significant eQTL association between rs324017 (PGC rank 103), and exon- and junction-level expression of *STAT6* (Figure 4C), a transcription factor involved in interleukin signaling[4] whereas there were no significant and replicated eQTLs at the gene-level.

The approach that was unconstrained by annotation also identified novel associations to schizophrenia risk, for example, the eQTL of rs6434928 (PGC rank 33) with intronic sequence in the *HSPD1* and *HSPE1* heat shock protein genes (Figure 4D,E), which are co-chaperonins linked by a common bidirectional promoter[5]. While both genes were recently highlighted in the subset of 20 PGC GWAS genes that are targets of currently approved drugs[6], the strictly intronic annotation of the eQTL association may suggest a different mechanism of risk from this locus in schizophrenia, and thus a different requirement for a therapeutic agent. Finally, we highlight the eQTL association between rs13240464 (PGC rank 16) and the expression level of a novel splice junction in *IMMP2L* (Figure 4F,G), a gene important in directing mitochondrial proteins to the mitochondria[7] and that has been previously associated with autism and Tourette Syndrome[8].

Supplementary Discussion

This proportion of genetic risk variants that are brain eQTLs – 42.5% – far exceeds previous reports using microarray and gene level (count-based) RNA-seq approaches. For example, of the 18 index SNPs labeled as eQTLs in Supplementary Table 4 In the PGC publication [9], only 8 risk loci had variants moderately linked (at $R^2 > 0.6$) to significant (at $p < 10^{-4}$) brain eQTLs. Our results highlight the enhanced detection of eQTL signals from GWAS-positive risk variants using RNA-seq to quantify convergent measures of expression beyond the usual gene-level quantification in large collections of postmortem human brains.

In the original GWAS report, while there were only 18 reported index SNPs with brain eQTLs to a total of 30 genes, many of the reported eQTLs were themselves not GWAS-significant variants, being only weakly linked (at $R^2 < 0.6$) to the best GWAS SNP. There were only three of these 18 GWAS- and eQTL-significant index SNPs (putatively associating with *C10orf32 (BORCS7), AS3MT*, *OGFOD2*, *NAGA*, and *CENPM*) that were present in our larger RNA-seq data from the three independent studies that we surveyed. Similarly, a recent GWAS and eQTL integration analysis based on peripheral blood eQTLs identified very few overlapping genes[10]. These genes included *PCCB*, *SNX19*, *GATAD2A*, and *IRF3* as directionally-consistent eQTLs (with regard to the GWAS, eQTL, and case-control differences described above), and *SF3B1* and *PSMA4* as non-directionally consistent eQTLs*,* suggesting the need to query the disease-relevant tissue, i.e. human brain. Interrogating annotation-agnostic expression features added a considerable number of loci with eQTL signal further highlighting the value of RNA sequencing-based datasets. For example, we have recently dissected the GWAS signal in the 10q24.32 region in a more focused analysis based on the expression features highlighted in the current report. There we identified based on junction-level analysis a novel risk-associated isoform of AS3MT that is missing exons 2 and 3 [11]. This analysis showed that the previously referenced genetic association with this locus to AS3MT was inaccurate, as no association was found to the full length AS3MT [11], and importantly, the novel isoform does not have arsenic methyltransferase activity. Furthermore, eight additional loci, with clinical risk variants associating with expression of *SNX19*, *NRGN/VSIG2*, *HSPD1*, *GATAD2A/NDUGA13*, *VPS45*, *GPM6A*, *SOX2-O*T, and *BRINP2*, were identified here (and subsequently replicated) only via junction- and expressed region-level analysis.

References

1 Frazee, A. C. *et al.* Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nature biotechnology* **33**, 243-246, doi:10.1038/nbt.3172 (2015).

2 Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660, doi:10.1126/science.1262110 (2015).

3 Labrie, V. *et al.* Serine racemase is associated with schizophrenia susceptibility in humans and in a mouse model. *Human molecular genetics* **18**, 3227-3243, doi:10.1093/hmg/ddp261 (2009).

4 Shimoda, K. *et al.* Lack of IL-4-induced Th2 response and IgE class switching in mice with disrupted Stat6 gene. *Nature* **380**, 630-633, doi:10.1038/380630a0 (1996).

5 Ryan, M. T. *et al.* The genes encoding mammalian chaperonin 60 and chaperonin 10 are linked head-to-head and share a bidirectional promoter. *Gene* **196**, 9-17 (1997).

6 Lencz, T. & Malhotra, A. K. Targeting the schizophrenia genome: a fast track strategy from GWAS to clinic. *Molecular psychiatry* **20**, 820-826, doi:10.1038/mp.2015.28 (2015).

7 Nunnari, J., Fox, T. D. & Walter, P. A mitochondrial protease with two catalytic subunits of nonoverlapping specificities. *Science* **262**, 1997-2004 (1993).

8 Petek, E. *et al.* Molecular and genomic studies of IMMP2L and mutation screening in autism and Tourette syndrome. *Molecular genetics and genomics : MGG* **277**, 71-81, doi:10.1007/s00438-006-0173-1 (2007).

9 Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-+, doi:10.1038/nature13595 (2014).

10 Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature genetics* **48**, 481-487, doi:10.1038/ng.3538 (2016).

11 Li, M. *et al.* A human-specific AS3MT isoform and BORCS7 are molecular risk factors in the 10q24.32 schizophrenia-associated locus. *Nature medicine*, doi:10.1038/nm.4096 (2016).