Supplementary Information:

Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis

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, William S Ulrich, Brady J. Maher, Amy Deep-Soboslay, The BrainSeq Consortium, Alan J. 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 S2). 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 28,127 unique Ensembl genes (across 19,515 gene symbols), including a core set of 15,334 Ensembl genes with all five features showing convergent expression association with development/aging (Figure S3). There were 64,397 previously unannotated expressed features corresponding to the majority of genes (14,295 gene IDs to 13,085 symbols) with genome-wide significant dynamic expression across brain development in these subjects (Table S4), suggest putative biological importance to at least a subset of this unannotated sequence. The majority of these unannotated features were identified using the exon-exon splice junction counts (N= 26,482, 41.1%) of which 18,350 junctions (69.3%) tagged alternative exonic boundaries and 8,132 junctions (30.7%) corresponding to exonic skipping, while there were 23,094 differentially expressed regions (DERs, 35.9%) that correspond to alternative exonic boundaries and 14,821 transcripts of which 13,862 (93.5%) correspond to exonic skipping novel isoforms.

We further performed sensitivity analyses to assess the effects of RIN on differential expression across the lifespan in our spline modeling. Statistically adjusting the age spline model for RIN changed very little of the inference – at the gene level, there were 21710 genes at Bonferroni significance compared to the 22209 genes reported in the text without the RIN adjustment (similar comparisons were seen at the other four feature summarization levels).

3. eQTL maps of the human frontal cortex

We hypothesized that, in general, analyzing transcript features like exons and junctions would increase statistical power for eQTL discovery if genetic variation regulated the expression levels of specific mRNA transcripts. At the gene-level, which collapses data from all transcripts into a single measure and is the most commonly implemented feature for eQTL discovery, the vast majority of expressed genes were associated with the expression of at least one nearby genetic variant. There were eQTLs to 6748 Ensembl Gene IDs (of which 4955 genes had HGC symbols) at stringent Bonferroni-adjusted significance ($p < 8.41x10^{-9}$, see Methods), and eQTLs to 18,416 Ensembl Gene IDs at more liberal FDR < 1% significance ($p < 1.84x10^{-4}$). However, we found a larger number of genes with eQTLs using exon-level analysis – 48,031 exons mapping to 8386 Ensembl IDs - at Bonferroni significance (“eExons”, $p < 7.64x10^{-10}$). Interestingly, while transcript-specific by nature, we actually found the fewest eQTLs to assembled-and-quantified transcripts (3,263 eTxns at $p < 1.73x10^{-9}$), in line with previous reports highlighting the difficulties in merging assemblies across many samples[19]. Lastly, there were an additional 3,022 eGenes identified with exon-level analysis compared to the 5364 eGenes identified with both summarization levels.

Among the 18908 junctions with eQTL signal at Bonferroni significance (“eJxns”, $p<1.1x10^{-9}$), 21.6% (N=4089) were previously unannotated, including 1312 eJxns to exon-skipping splicing events and 2777 eJxns to shifted exonic boundaries (acceptor or donor splice sites).

The eJxns also highlight a large degree of potential transcript specificity, both in the 4089 unannotated junctions as well as 3388 additional annotated eJxns that delineate individual transcript isoforms (when multiple isoforms are present for the gene). At the expressed region-level, among the 27,643 ERs with eQTL signal at Bonferroni significance ("eERs", $p<1.28x10^{-9}$), 14,890 were either fully or partially unannotated, with partial events including 4521 exon extensions into neighboring intronic sequence and 769 extended untranslated regions (UTRs) and fully unannotated events being strictly intronic (N=6,255) and intergenic (N=3,345) sequences. These two feature classes also had the largest eQTL effect sizes of the tested features, with 41.4% and 29.2% change in expression per allele copy for eJxns and eERs. Lastly, we found that 1,042 Ensembl genes had eQTLs exclusively to unannotated sequence with no corresponding eQTL signal to annotated features in the genes.

4. 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.

5. 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.

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).