Supplementary Information

Integrative analyses of splicing in the aging brain: role in susceptibility to Alzheimer's Disease

Contents

Note: Supplementary Tables are provided as separate Excel files.

1. Supplementary Notes

1.1. Religious Orders Study and Memory and Aging Project

Gene expression data[1]. Gene expression data were generated using RNA-sequencing from Dorsolateral Prefrontal Cortex (DLPFC) of 540 individuals, at an average sequence depth of 90M reads. Detailed description of data generation and processing was previously described[2] (Mostafavi, Gaiteri et al., under review). Samples were submitted to the Broad Institute's Genomics Platform for transcriptome analysis following the dUTP protocol with Poly(A) selection developed by Levin and colleagues[3]. All samples were chosen to pass two initial quality filters: RNA integrity (RIN) score >5 and quantity threshold of 5 ug (and were selected from a larger set of 724 samples). Sequencing was performed on the Illumina HiSeq with 101bp paired-end reads and achieved coverage of 150M reads of the first 12 samples. These 12 samples will serve as a deep coverage

reference and included 2 males and 2 females of nonimpaired, mild cognitive impaired, and Alzheimer's cases. The remaining samples were sequenced with target coverage of 50M reads; the mean coverage for the samples passing QC is 95 million reads (median 90 million reads). The libraries were constructed and pooled according to the RIN scores such that similar RIN scores would be pooled together. Varying RIN scores results in a larger spread of insert sizes during library construction and leads to uneven coverage distribution throughout the pool. RNA-seq data were processed by our parallelized pipeline. This pipeline included trimming the beginning and ending bases from each read, identifying and trimming adapter sequences from reads, detecting and removing rRNA reads, aligning reads to reference genome (using Bowtie[4]) and quantification of transcript expression levels (using RSEM[5]). Specifically, RNA-Seq reads in FASTQ format were inspected using FASTQC program[6]. Barcode and adapter contamination, low quality regions (8bp at beginning and 7bp at ending of each fastq reads) were trimmed using FASTX-toolkit. To remove rRNA contamination, we aligned trimmed reads to rRNA reference (rRNA genes were downloaded from UCSC genome browser selecting the RepeatMask table) by BWA then extracted only paired unmapped reads for transcriptome alignment. rRNA depleted reads were then mapped to transcriptome reference (gencode v14) using Trinity package with RSEM as output option. Gene expression FPKM values were estimated by “rsem-calculate-expression” from RSEM. Samples from 494 individuals were used in the eQTL analysis, which include those that had QC’d genotype and pass the expression outlier test (a D-statistic below 0.99).

DNA methylation data[7]. DNA methylation data were generated using the 450K Illumina array from DLPFC of 740 individuals. Detailed description of data acquisition and QC are previously published[7]. Briefly, methylation probes that coincided with common polymorphic sites were removed. Initial normalization of CpG probes to account for differences between type I and type II probes, was performed using the BMIQ algorithm from the Watermelon package[8] and beta-

values were extracted for further analysis. The SNM approach[9] was then used to regress out the effects of batch, PMI, sex, age at death, and a previously published estimate of proportion of neurons present in each sample[7]. The samples from 468 individuals were analyzed for which gene expression data was also available[2]. As described below, this decision was made to enable using gene expression data to estimate the proportions of the five major brain cell types. This correction for cell type proportions was done in addition to the regression approach for removing the effect of generic neuronal proportions based on DNAm marks[7].

Histone modification data[1]. Histone modification data were generated using H3K9Ac ChIP-sequencing from DLPFC of 714 individuals. Single-end reads were aligned by the BWA algorithm[10], and peaks were detected in each sample separately using the MACS2 algorithm (using the broad peak option and a q-value cutoff of 0.001). A series of QC steps were employed to identify and remove low quality reads, and samples that did not reach (i) ≥ 15x106 unique reads, (ii) non-redundant fraction ≥ 0.3, (iii) cross correlation ≥ 0.03, (iv) fraction of reads in peaks ≥ 0.05 and (v) ≥ 6000 peaks were removed. Cross correlation was defined as the maximum Pearson's correlation between the read coverage on the negative and positive strand after binning reads into 10bp bins[11]. Cross correlation was calculated after shifting the reads on the negative strand by s base pairs for s = 0, 10, 20, ..., 1000, and the maximum cross correlation was reported. In total, 669 samples passed quality control. H3K9Ac domains were defined by calculating all genomic regions that were detected as a peak in at least 100 of the 669 samples (15%). Regions within 100bp from each other were merged and very small regions of less than 100bp were removed. Reads were then extended towards the 3' end to the fragment size of the respective sample. The fragment size was estimated by the shift smax that maximized the cross correlation (mean smax = 271bp). Finally, the number of extended reads in each H3K9Ac region was determined for each sample. Only uniquely mapped distinct reads were considered. Quantified histone acetylation data were quantile

normalized to account for variability in sequencing depth across individuals. Samples from 433 individuals for which gene expression data were available were used in our analysis.

1.2 Mount Sinai Brain Bank Alzheimer's Disease Data

Brain specimens were obtained from the Mount Sinai/JJ Peters VA Medical Center Brain Bank (MSBB) which holds over 1,700 samples. This cohort was assembled after applying stringent inclusion/exclusion criteria and represents the full spectrum of disease severity. Neuropathological assessments are performed according to the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) protocol[12] and include assessment by hematoxylin and eosin, modified Bielschowski, modified thioflavin S, and anti-β amyloid (4G8), anti-tau (AD2) and anti-ubiquitin (Daka Corp.). Each case is assigned a Braak AD-staging score for progression of neurofibrillary neuropathology Braak et al.[13]. Quantitative data regarding the density of neuritic plaques in the middle frontal gyrus, orbital frontal cortex, superior temporal gyrus, inferior parietal cortex and calcarine cortex are also collected as described[14]. Clinical dementia rating scale (CDR) and mini–mental state examination (MMSE) severity tests are conducted for assessment of dementia and cognitive status. Final diagnoses and CDR scores are conferred by consensus. Based on CDR classification[15], subjects are grouped as no cognitive deficits (CDR = 0), questionable dementia (CDR = 0.5), mild dementia (CDR = 1.0), moderate dementia (CDR = 2.0), and severe to terminal dementia (CDR = 3.0–5.0). This tissue source was used to perform two separate experiments. RNA sequencing data was generated across 3 brain regions - selected based on the pilot experiment - across 196 individuals. Additional data is being generated for the primary study to include additional samples as well as whole exome sequencing data.

Tissue preparation and RNA isolation. This distribution contain 1030 samples collected from 301 individuals from Brodmann Areas 10, 22, 36 and 40. The

specific brain regions, Brodmann areas, were dissected while frozen from flash frozen never-thawed ~8 mm thick coronal tissue blocks using a dry ice cooled reciprocating saw. The dissected regions were then pulverized to a fine powder consistency in liquid nitrogen cooled mortar and pestle and distributed into 50 mg aliquots. All aliquots were barcoded and stored at -80oC until RNA isolation. The total RNA were isolated using RNeasy Lipid Tissue Mini Kit from Qiagen (cat#74804) according to the manufacturer's protocol (The RNeasy Lipid Tissue Mini Kit Handbook, Qiagen 104945, 02/2009).

MSBB RNA-seq protocol. Preparation of samples for RNA-Seq analysis was performed using the TruSeq RNA Sample Preparation Kit v2 (Illumina, San Diego, CA). Briefly, rRNA was depleted from total RNA using the Ribo-Zero rRNA Removal Kit (Human/Mouse/Rat) (Illumina, San Diego, CA) to enrich for coding RNA and long non-coding RNA. The cDNA was synthesized using random hexamers, end-repaired and ligated with appropriate adaptors for sequencing. The library then underwent size selection and purification using AMPure XP beads (Beckman Coulter, Brea, CA). The appropriate Illumina recommended 6-bp bar-code bases are introduced at one end of the adaptors during PCR amplification step. The size and concentration of the RNAseq libraries was measured by Bioanalyzer (Agilent, Santa Clara, CA) and Qubit fluorometry (Life Technologies, Grand Island, NY) before loading onto the sequencer. The Ribo-Zero libraries were sequenced on the Illumina HiSeq 2500 System with 100 nucleotide single end reads, according to the standard manufacturer’s protocol (Illumina, San Diego, CA).

Alignment and quantification. The raw sequence reads were aligned to human genome hg19 with the star aligner (v2.3.0e). Ensembl gene annotation model version GRCh37.70 was utilized to assist with the mapping of reads onto known human genes. Then the gene and exon level expression (read counts) were quantified by featureCounts[16] (v1.4.4) from the Subread package.

1.3 CommonMind Consortium Dataset

Detailed description of CMC consortium RNA-seq data generation and processing was previously described in Fromer et al.[17].

Post-mortem samples. Data generated for this study came from post-mortem human brain specimens originating from the tissue collections at the three brain banks described below. All samples were shipped to the Icahn School of Medicine at Mount Sinai (ISMMS) for nucleotide isolation and data generation.

Selection criteria. Post-mortem tissue from schizophrenia (SCZ) and bipolar or other affective/mood disorder (AFF) cases were included if they met the appropriate diagnostic DSM-IV criteria, as determined in consensus conferences after review of medical records, direct clinical assessments, and interviews of family members or care providers. Cases were excluded if they had neuropathology related to Alzheimer's disease and/or Parkinson's disease, acute neurological insults (anoxia, strokes and/or traumatic brain injury) immediately before death, or were on ventilators near the time of death. Three case samples (2 with leucotomies, and 1 with a history of a head injury before diagnosis) were included; these were not outliers on any metrics that we used to evaluate our samples.

"**MSSM" sample:** Mount Sinai NIH Brain Bank and Tissue Repository (NBTR) (http://icahn.mssm.edu/research/labs/neuropathology-and-brain-banking). The Mount Sinai Brain Bank was established in 1985. The NBTR obtains brain specimens from the Pilgrim Psychiatric Center, collaborating nursing homes, Veteran Affairs Medical Centers and the Suffolk County Medical Examiner's Office. Diagnoses are made based on DSM-IV criteria and are obtained through direct assessment of subjects using structured interviews and/or through psychological autopsy by extensive review of medical records and informant and caregiver interviews[18,19]. Informed consent is obtained from the next of kin. The brain bank procedures are approved by the ISMMS IRB and exempted from

further IRB review due to the collection and distribution of postmortem specimens. All samples for the study were dissected from the left hemisphere of fresh frozen coronal slabs cut at autopsy from the dorsolateral prefrontal cortex (DLPFC) from Brodmann areas 9/46. Immediately after dissection, samples were cooled to −190 °C and dry homogenized to a fine powder using an L-N2 cooled mortar and pestle. Tissue was transferred on dry ice to ISMMS as a dry powder for DNA and RNA extraction.

"Pitt" sample: The University of Pittsburgh Brain Tissue Donation Program. Brain specimens from the University of Pittsburgh Program are obtained during routine autopsies conducted at the Allegheny County Office of the Medical Examiner (Pittsburgh) following the consent of the next of kin[20]. An independent committee of experienced research clinicians makes consensus DSM-IV diagnoses for all subjects on the basis of medical records and structured diagnostic interviews conducted with the decedent's family member[21]. All procedures for Pitt samples have been approved by the University of Pittsburgh's Committee for the Oversight of Research involving the Dead and Institutional Review Board for Biomedical Research. At autopsy, the right hemisphere of each brain is blocked coronally, immediately frozen, and stored at −80 °C[22]. Samples for this study contained only the gray matter of DLPFC, where Brodmann area 9/46 was cut on a cryostat and collected in tubes appropriate for DNA or RNA extraction. The DNA and RNA tubes were shipped on dry ice to ISMMS as homogenized tissue in Trizol for RNA extraction and thinly sliced tissue for DNA extraction. Specimens from Pitt were provided as matched case/control pairs. These were perfectly matched for sex, and as closely as possible for age (73% of pairs were matched within 5 years, and 95% within 10 years) and race (71% of pairs were matched for race). Members of a pair were always processed together for RNA-seq. Tissue for 10 of the Pitt controls was extracted in duplicate, once as part of a SCZ pair and once as part of a bipolar pair.

“Penn” sample: University of Pennsylvania Brain Bank of Psychiatric illnesses and Alzheimer's Disease Core Center (http://www.med.upenn.edu/cndr/biosamples-brainbank.shtml). Brain specimens are obtained from the Penn prospective collection. Disease diagnoses were made based on DSM-IV criteria and obtained through a clinical interview by psychiatrist and review of medical records. All procedures for Penn are approved by the Committee on Studies Involving Human Beings of the University of Pennsylvania, and the use of control postmortem tissues was considered exempted research in accordance with CFR 46.101 (b), item 65 of Federal regulations and University policy. At autopsy, the right or left hemisphere of each brain is blocked into coronal slabs, which are immediately frozen and stored at −80 °C. For this study, Brodmann areas 9/46 were dissected from either the left or right hemisphere and pulverized in liquid nitrogen. The tissue was shipped in tubes appropriate for DNA or RNA extraction to ISMMS as homogenized tissue in Trizol for RNA extraction and as dry pulverized tissue for DNA extraction.

Tissue, RNA and DNA preparation. Total RNA was isolated from approximately 50 mg homogenized tissue in Trizol using the RNeasy kit according to manufacturer protocol. Samples were processed in batches of 12, and the Pitt matched case/control pairs were always processed in the same batch. The order of extraction for SCZ-affected and control samples was assigned randomly with respect to brain bank, diagnosis, and all other sample characteristics. Because the affective disorder cases (AFF) and matched controls from Pitt were not available until after the processing of the SCZ and controls was underway, these samples were randomized among the remaining 132 SCZ and control samples still queued for extraction at that time. The mean total RNA yield was 15.3 μg (± 5.7). The RNA Integrity Number (RIN) was determined by fractionating RNA samples on the 6000 Nano chip (Agilent Technologies) on the Agilent 2100 Bioanalyzer. 51 samples with RIN < 5.5 were excluded from the study. Among the remaining samples, the mean RIN was 7.7 (± 0.9), and the mean ratio of 260/280 was 2.0 (± 0.02).

DNA was isolated from approximately 10 mg dry homogenized tissue from specimens coming from the MSSM and Penn brain banks. The thinly sliced tissue from Pitt was homogenized before DNA isolation. All DNA isolation was performed using the Qiagen DNeasy Blood and Tissue Kit according to the manufacturer's protocol. DNA yield was quantified using Thermo Scientific's NanoDrop. The mean yield was 12.6 µg (± 4.6), the mean ratio of 260/280 was 2.0 (± 0.1), and the mean ratio of 260/230 was 1.8 (± 0.6).

RNA library preparation and sequencing. Processing order was re-randomized before ribosomal RNA (rRNA) depletion, and samples were processed in batches of 8. To expedite sequencing, processing began before extraction was complete and randomization occurred among all available extracted samples in sets of 120 to 226. Briefly, rRNA was depleted from about 1 µg of total RNA using Ribo-Zero Magnetic Gold kit (Illumina/Epicenter Cat # MRZG12324) to enrich for polyadenylated coding RNA and noncoding RNA. The Pitt case/control pairs were batched together in each processing step, including Ribo-Zero depletion, sequence library preparation, and sequencing lane. Ten of the Pitt controls were extracted and sequenced as independent duplicates, once as part of a SCZ pair and once as part of a bipolar pair. The sequencing library was prepared using the TruSeq RNA Sample Preparation Kit v2 (RS-122-2001-48 reactions) in batches of 24 samples. The insert size and DNA concentration of the sequencing library was determined on Agilent Bioanalyzer and Qubit, respectively. A pool of 10 barcoded libraries were layered on a random selection of two of the eight lanes of the Illumina flow cell bridge amplified to ~250 million raw clusters. One hundred base pair paired-end reads were obtained on a HiSeq 2500. The sequence data were processed for primary analysis to generate QC values (reads were mapped to the human reference genome using TopHat. Samples with a minimum of 50 million mapped reads (~25 million paired-end reads) and less than 5% rRNA-aligned reads were retained for downstream analysis. We attempted a single round of resequencing for samples that failed

these QC criteria. In the end, a total of 15 samples did not meet these sequencing criteria and were discarded.

DNA genotyping, QC, ancestral evaluation and polygenic scoring. Genotyping was performed on the Illumina Infinium HumanOmniExpressExome 8 v 1.1b chip (Catalog #: WG-351-2301) using the manufacturer's protocol. Samples for genotyping were aliquoted onto 96 well plates, where each plate had an internal control from the HapMap project (NA12878, Coriell Institute) in two unique locations. Initial QC was performed using PLINK57 to remove markers with: zero alternate alleles, genotyping call rate ≤ 0.98, Hardy-Weinberg P value < 5 × 10−5, and individuals with genotyping call rate < 0.90. This removed 2 samples from the analysis. After QC, 668 individuals genotyped at 767,368 markers were used for imputation. Phasing was performed on each chromosome using ShapeIt v2.r790[23], and variants were imputed in 5 Mb segments by Impute v2.3.1[24] with the 1,000 Genomes Phase 1 integrated reference panel11 excluding singleton variants. Note that, in addition to the 22 autosomes, we also included chromosome X, split out into pseudoautosomal (PAR) and non-PAR genomic regions to properly handle male haploidy in the non-PAR regions.

1.4 Data Availability

Data from the ROSMAP study were used in this work, and are available at: RADC Research Resource Sharing Hub at www.radc.rush.edu, and https://www.synapse.org/#!Synapse:syn3219045. The details of the xQTL Association Analysis are described in Ng et al. [2]. The xQTL results and analysis scripts can be accessed through online portal, xQTL Serve, at http://mostafavilab.stat.ubc.ca/xQTLServe.

The MSBB RNASeq data is available via the AMP-AD Knowledge Portal at https://www.synapse.org/#!Synapse:syn2580853/wiki/409840.

The data and analysis pipeline is described in Fromer et al. [17]. The CommonMind RNASeq data is available via the CommonMind Consortium Knowledge Portal at https://www.synapse.org/#!Synapse:syn2759792/wiki/69613.

2. Supplementary Tables

See separate Excel spreadsheets for Tables.

Supplementary Table 1. Demographic characteristics of ROS and MAP cohort.

Supplementary Table 2. A list of significantly differentially spliced introns associated with neuropathologies.

Supplementary Table 3. A list of significantly differential spliced introns associated with clinical AD status.

Supplementary Table 4. List of differential spliced introns associated with clinical AD status in ROSMAP that replicates in the MSBB dataset.

Supplementary Table 5. A list of splicing QTLs at FDR 0.05 identified in ROSMAP dataset.

Supplementary Table 6. Significant TWAS genes with association to IGAP AD GWAS.

Supplementary Table 7. Significant TWAS genes with association to meta-analysis of IGAP and UKBB AD GWAS.

3.Supplementary Figures

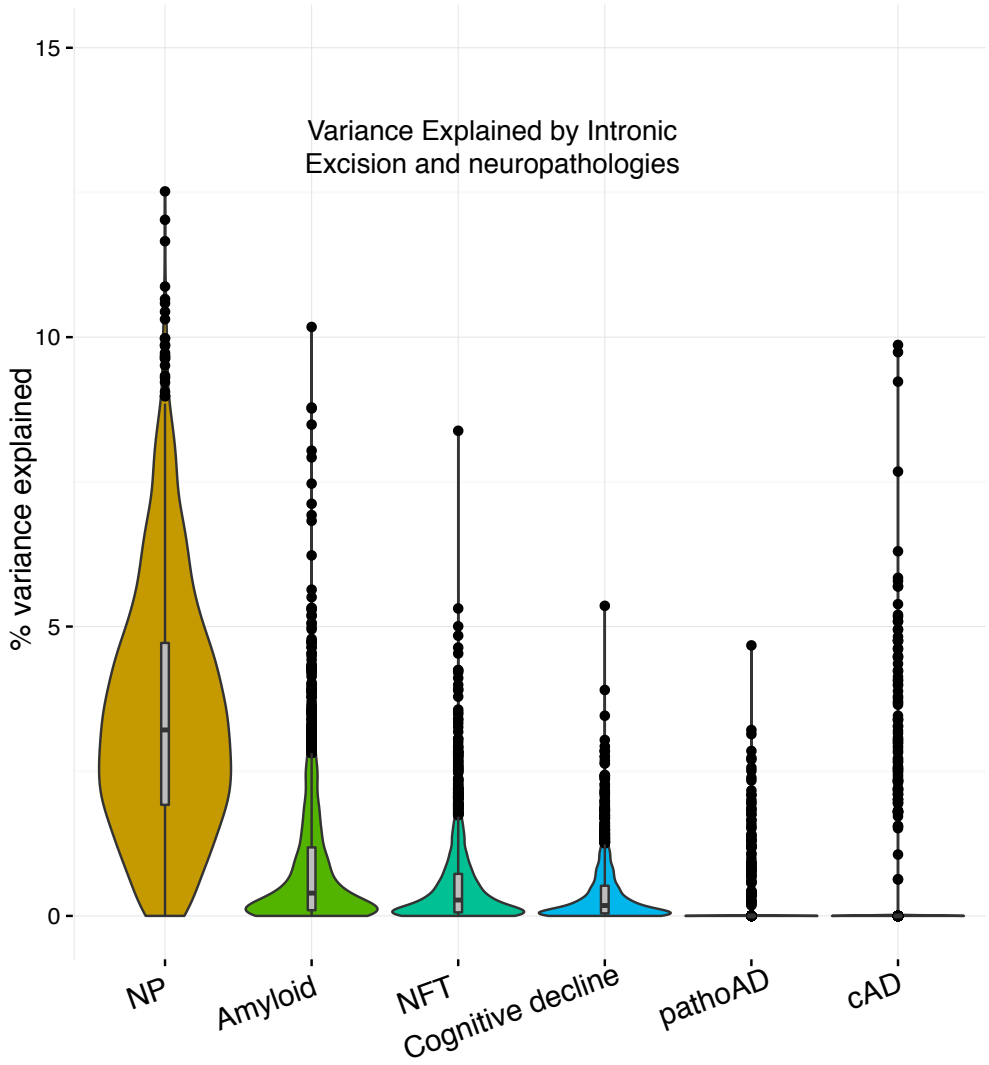

Figure S1. Violin and box plots of percent variation in intronic usage explained by each neuropathology measure including neuritic plaques (NP), amyloid burden, Neurofibrillary tangles (NFTs), global cognitive decline, pathological diagnosis of AD (pathoAD) and clinical diagnosis of AD (cAD). This plot is generated after accounting for technical and biological covariates (batch, PMI, RIN, Ribosomal basses, number of aligned reads, sex, and age of death). Each dot represents an intronic usage cluster. The figure was generated using variancePartition[25].

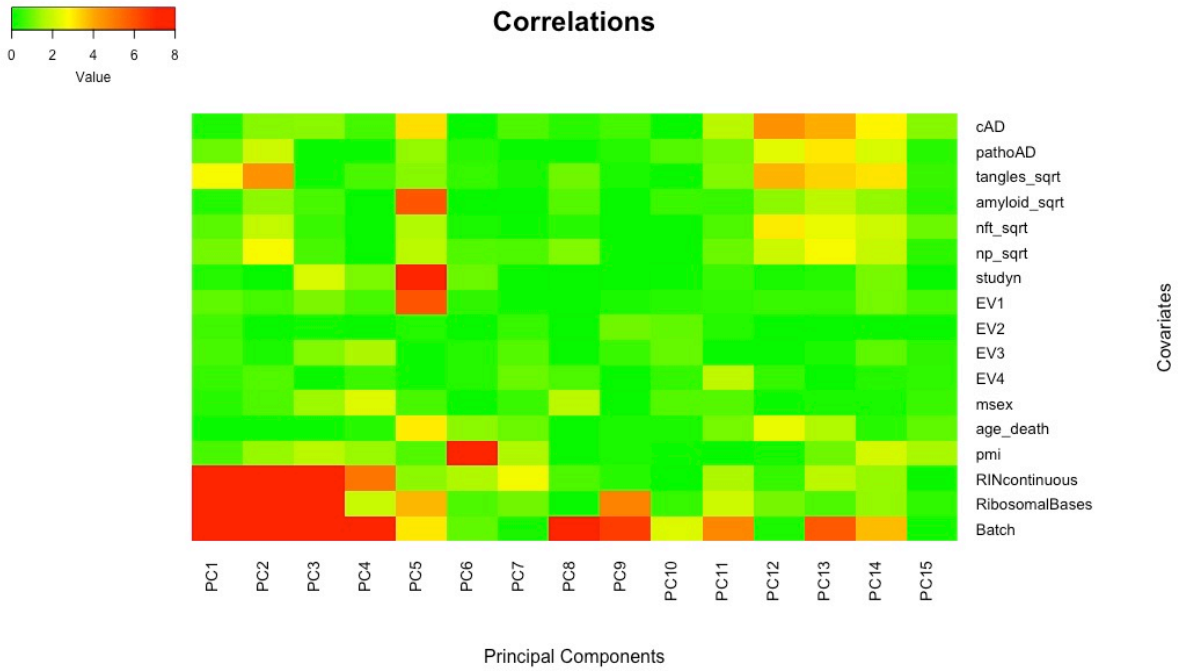

Figure S2. Figure shows the strength of the association between top intronic usage PCs and technical and biological confounding factors. Here batch refers to the date of RNA preparation. Genotype PCs were computed as the top 4 PCs (EV1-EV4) of genotype data. Study index refers to RUSH vs MAP samples. The heatmap depicts the $-\log_{10}$(P-value) for correlation coefficient. PMI= postmortem interval.

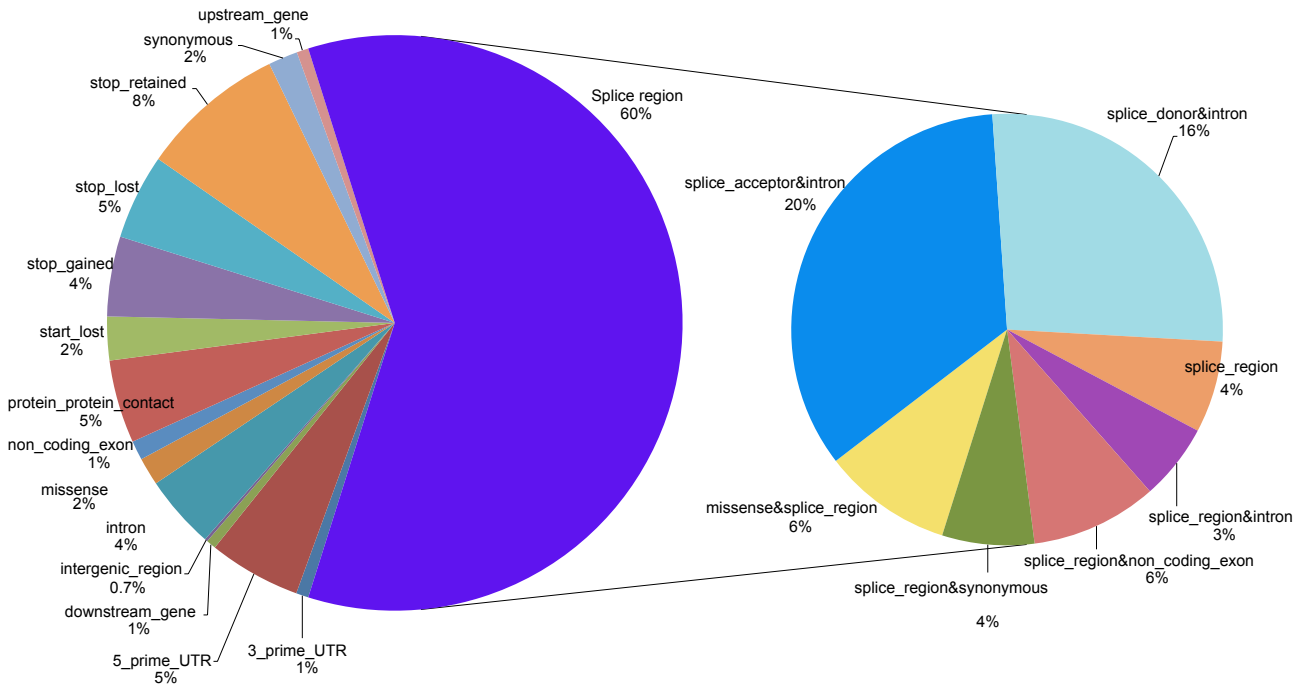

Figure S3. Proportion of sQTLs mapping to different types of variants in relation to a transcript. The location and consequence to transcripts for SNPs were downloaded from Ensembl Variant Effect Predictor database.

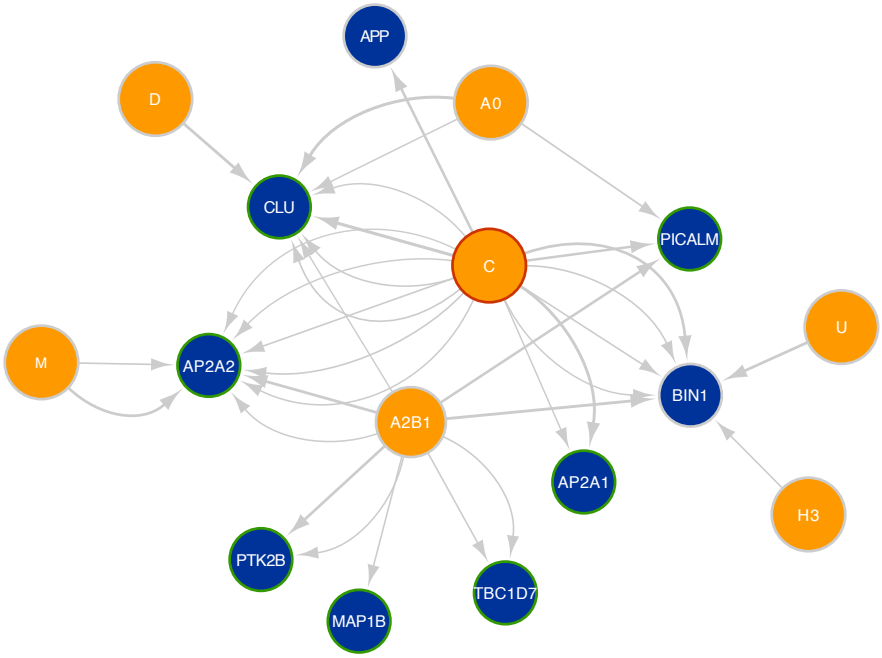

Figure S4. hnRNP splicing factors (in yellow) are correlated with intronic excision levels of hundreds of genes, many of which are in AD susceptibility loci including *BIN1, PICALM, APP, AP2A2, PTK2B, MAP1B, TBC1D7*, and *CLU (in blue).*

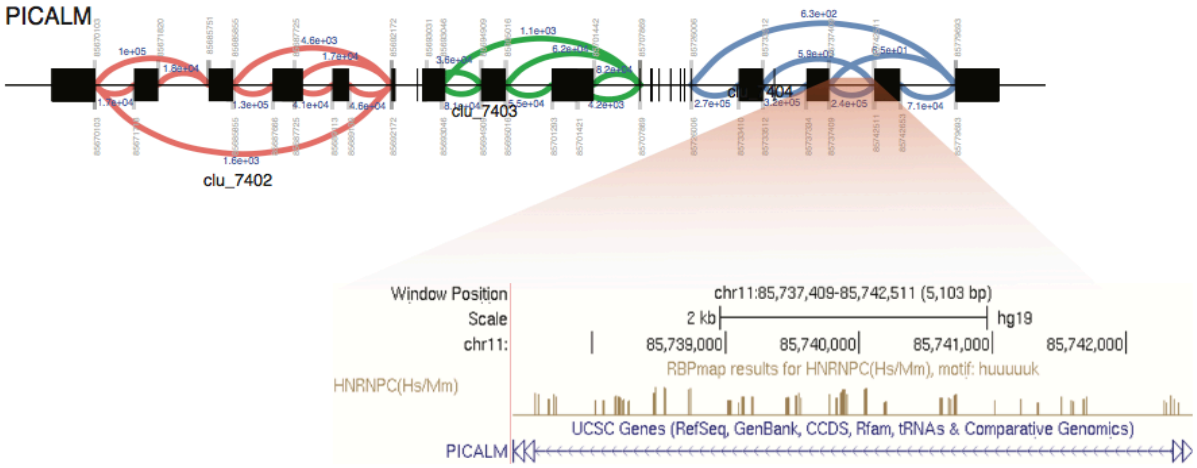

Figure S5. Intronic usage plot for *PICALM* and *HNRNPC* CLIP binding sites. Shown here are three intronic usage clusters (red, green and blue). Zoomed panel shows a region of the intronic excision (Chr11: 85737409-85742511) event with *HNRNPC* CLIP binding sites.

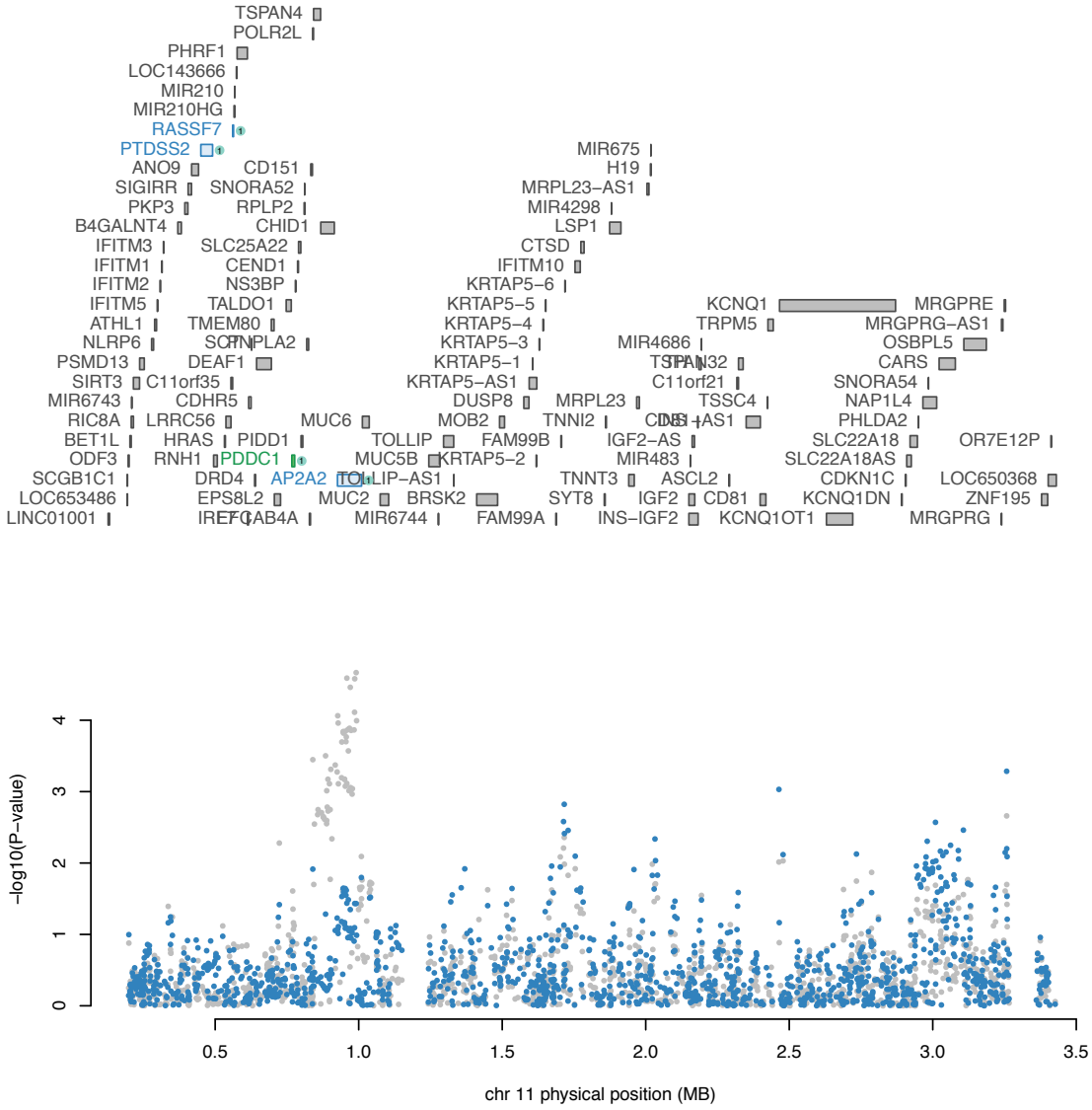

Figure S6. Conditional analysis of IGAP AD GWAS results for splicing effects for *AP2A2*. The AD GWAS effect is explained by splicing effect at *AP2A2* (chr11:946963:959437:clu_5363). Shown here are AD GWAS P-value (in blue) and conditioned P-value (in grey). The GWAS P-value at *AP2A2* is suggestive in the original IGAP study ($p<10^{-5}$).

Figure S7. Conditional analysis of IGAP AD GWAS results for splicing effects for *RABEP1.* The AD GWAS effect is explained by splicing effect at *RABEP1* (chr17:5257785:5264503:clu_10832). Shown here are AD GWAS P-value (in blue) and conditioned P-value (in grey). The GWAS P-value at *RABEP1* is suggestive in the original IGAP study ($p<10^{-5}$).

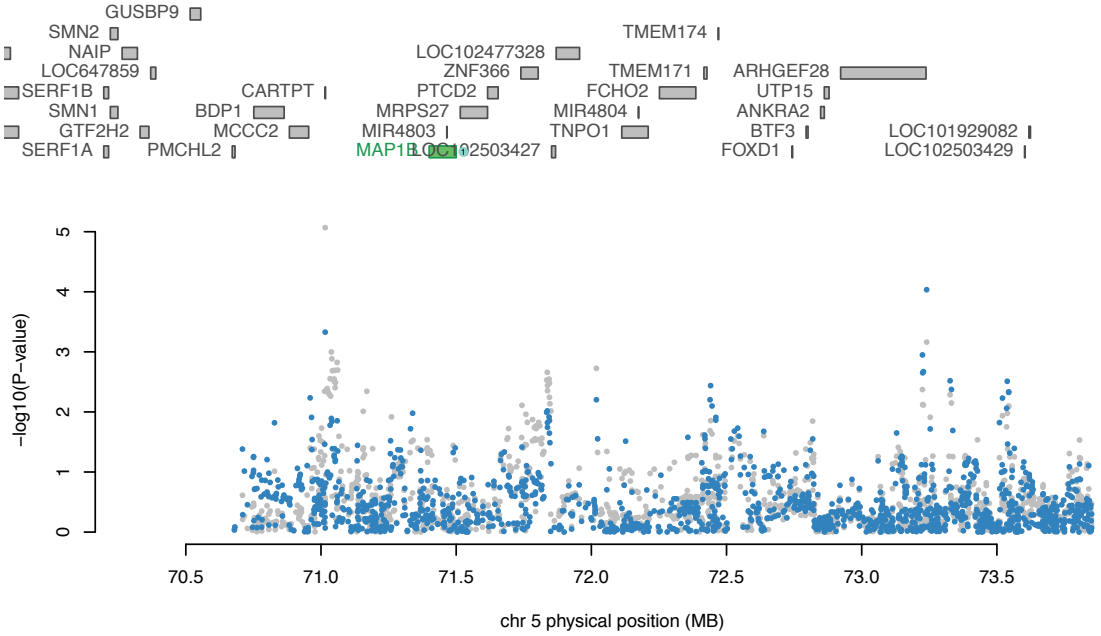

Figure S8. Conditional analysis of IGAP AD GWAS results for splicing effects for *MAP1B.* The AD GWAS effect is mostly explained by splicing effect at *MAP1B* (chr5:71404388:71411525:clu_33875). There seems to be a secondary effect in this locus. Shown here are AD GWAS P-value (in blue) and conditioned P-value (in grey). The GWAS P-value at *MAP1B* is suggestive in the original IGAP study ($p<10^{-5}$).

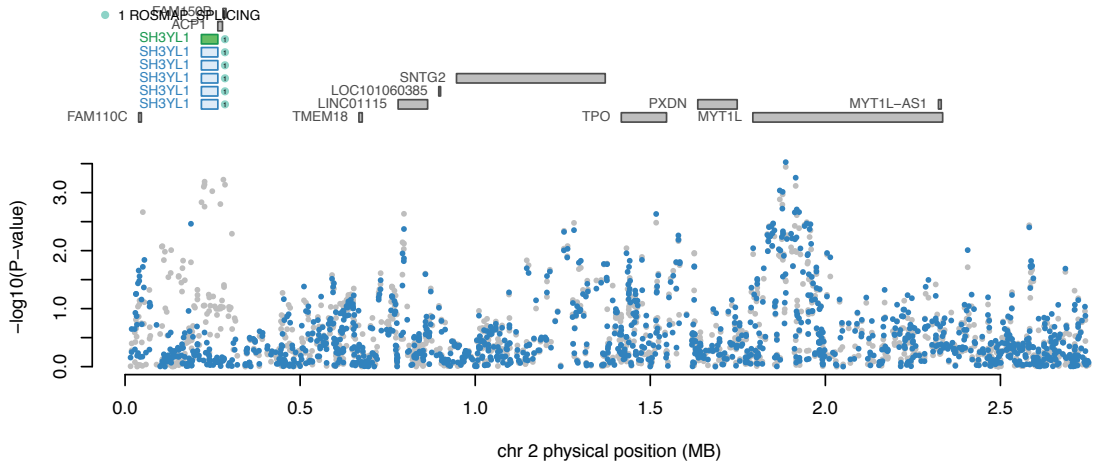

Figure S9. Conditional analysis of IGAP AD GWAS results for splicing effects for *SH3YL1*. The AD GWAS effect is mostly explained by splicing effect at *SH3YL1* (chr2:243562:247538:clu_40805). Shown here are AD GWAS P-value (in blue) and conditioned P-value (in grey).

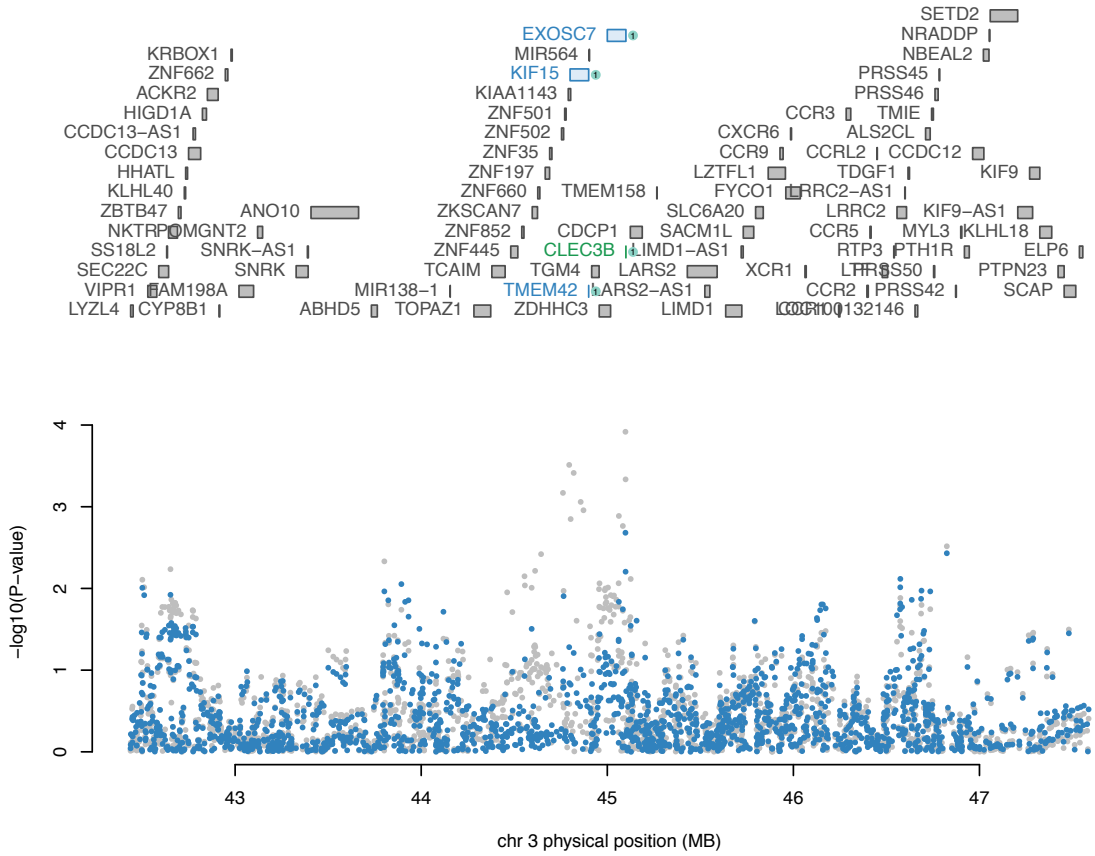

Figure S10. Conditional analysis of IGAP AD GWAS results for splicing effects for *CLEC3B.* The AD GWAS effect is mostly explained by splicing effect at *CLEC3B* (chr3:45067963:45072319:clu_38309). Shown here are AD GWAS P-value (in blue) and conditioned P-value (in grey). The GWAS P-value at *CLEC3B* is suggestive in the original IGAP study ($p<10^{-5}$).

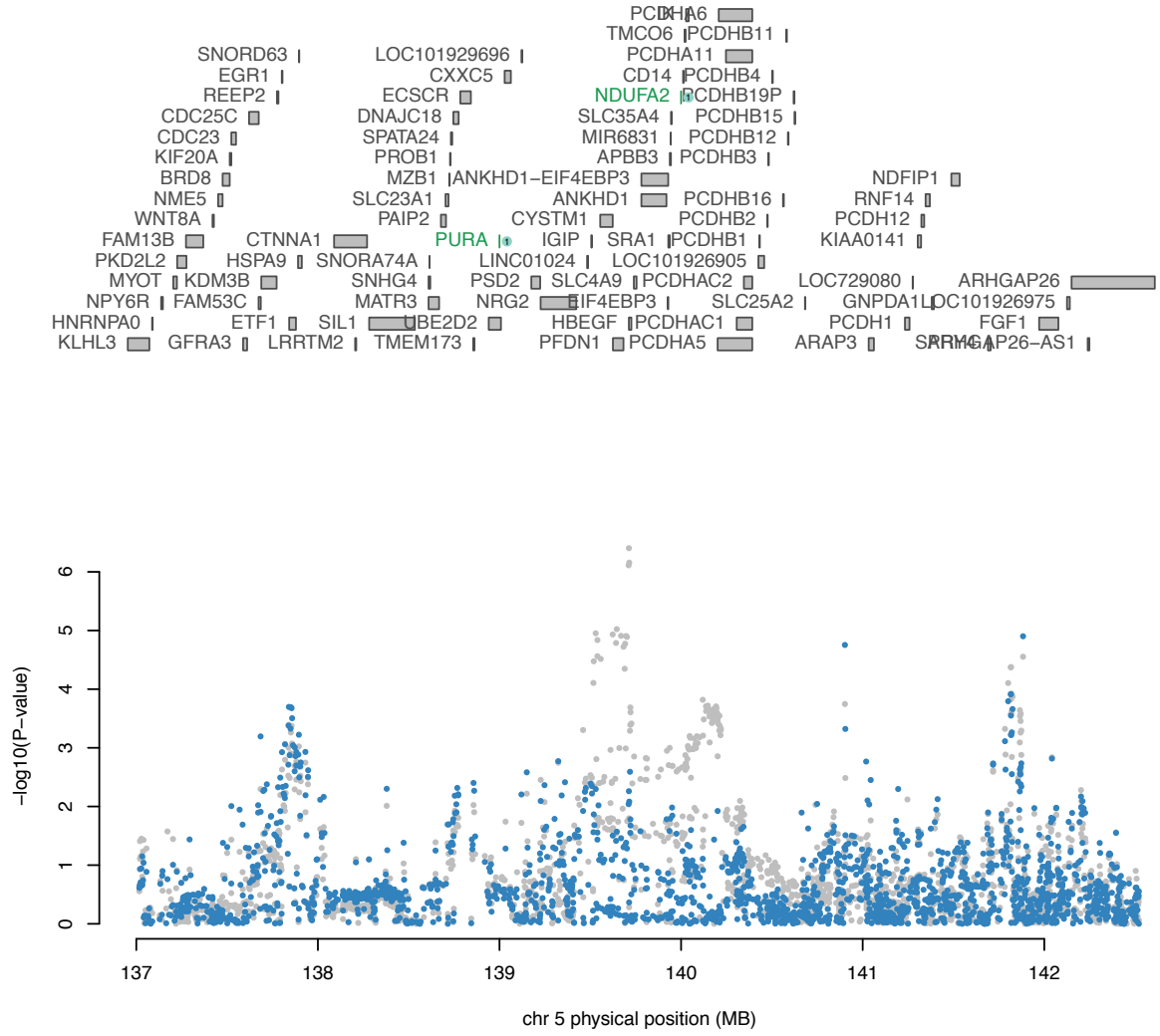

Figure S11. Conditional analysis of IGAP AD GWAS results for splicing effects for *NDUFA2.* The AD GWAS effect is mostly explained by splicing effect at *NDUFA2* (chr5:140025303:140026841:clu_34750). Shown here are AD GWAS P-value (in blue) and conditioned P-value (in grey). The GWAS P-value at *NDUFA2* is suggestive in the original IGAP study ($p<10^{-5}$).

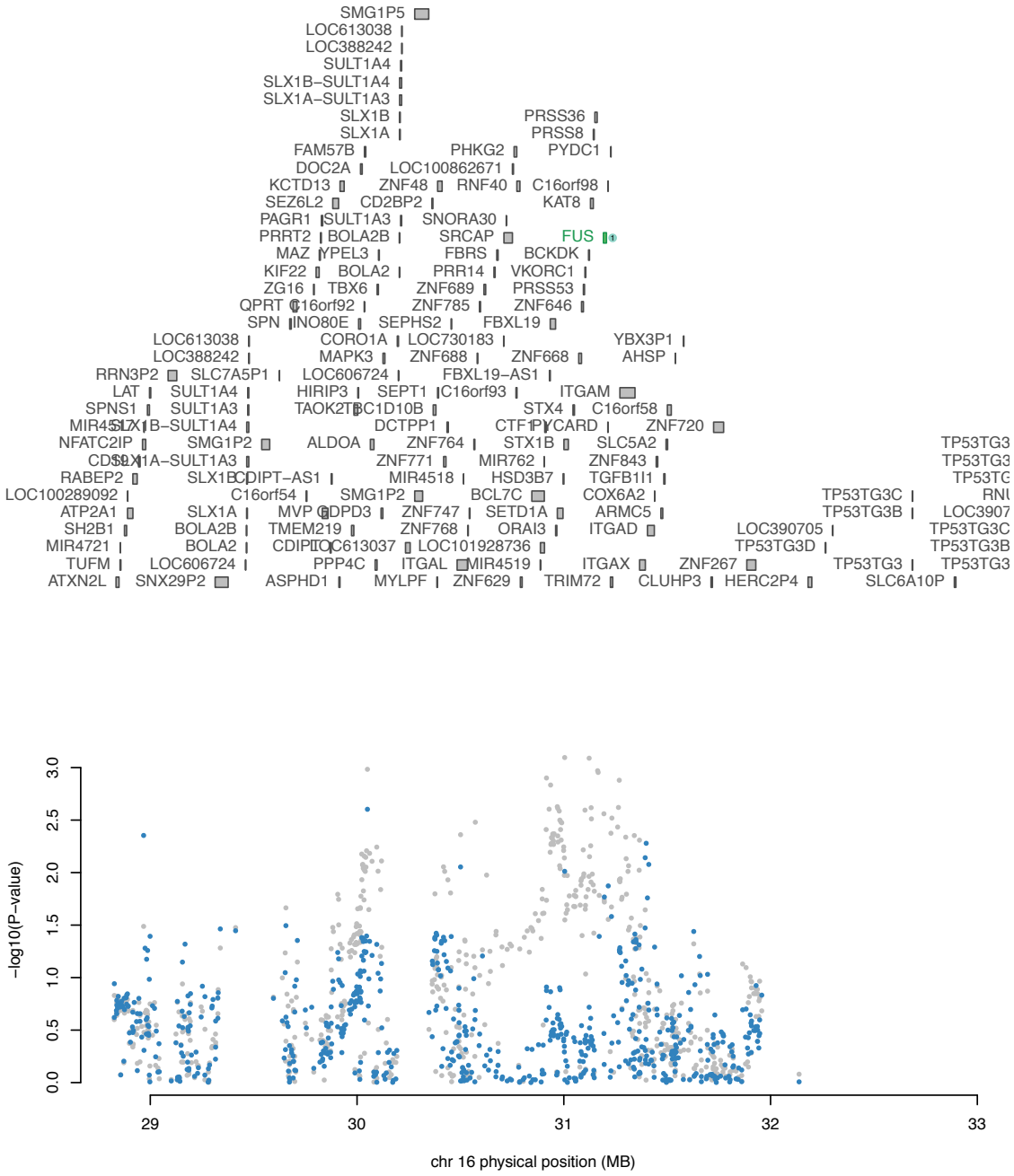

Figure S12. Conditional analysis of IGAP AD GWAS results for splicing effects for *FUS.* The AD GWAS effect is mostly explained by splicing effect at *FUS* (chr16:31194180:31195179:clu_15168). Shown here are AD GWAS P-value (in blue) and conditioned P-value (in grey). The GWAS P-value at *FUS* is suggestive in the original IGAP study ($p<10^{-4}$).

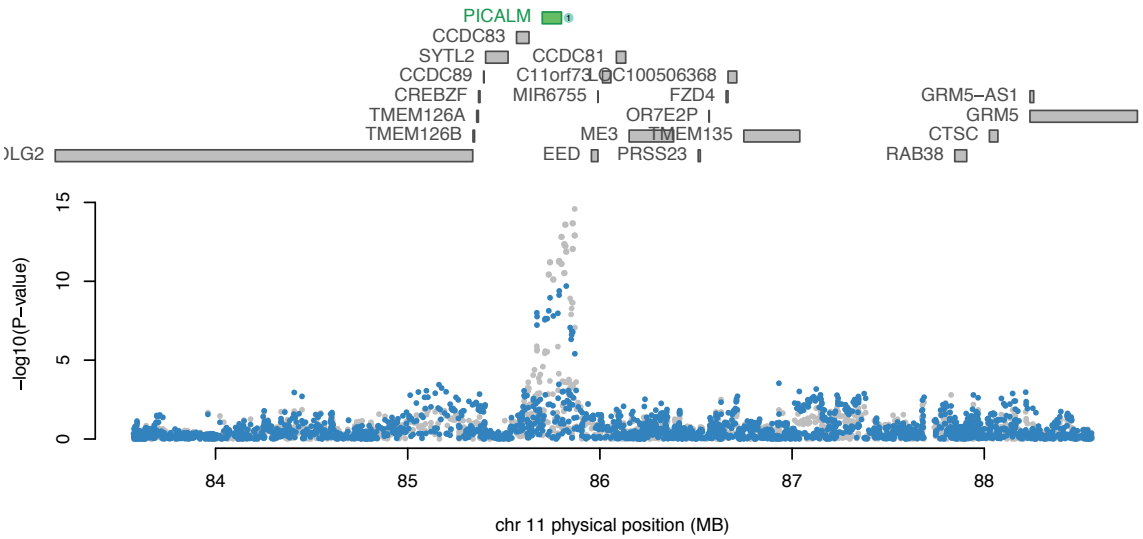

Figure S13. Conditional analysis of IGAP AD GWAS results for splicing effects for *PICALM.* The AD GWAS effect is mostly explained by splicing effect at *PICALM* (chr11:85737409:85742511:clu_7404). There is a secondary effect in this locus. Shown here are AD GWAS P-value (in blue) and conditioned P-value (in grey).

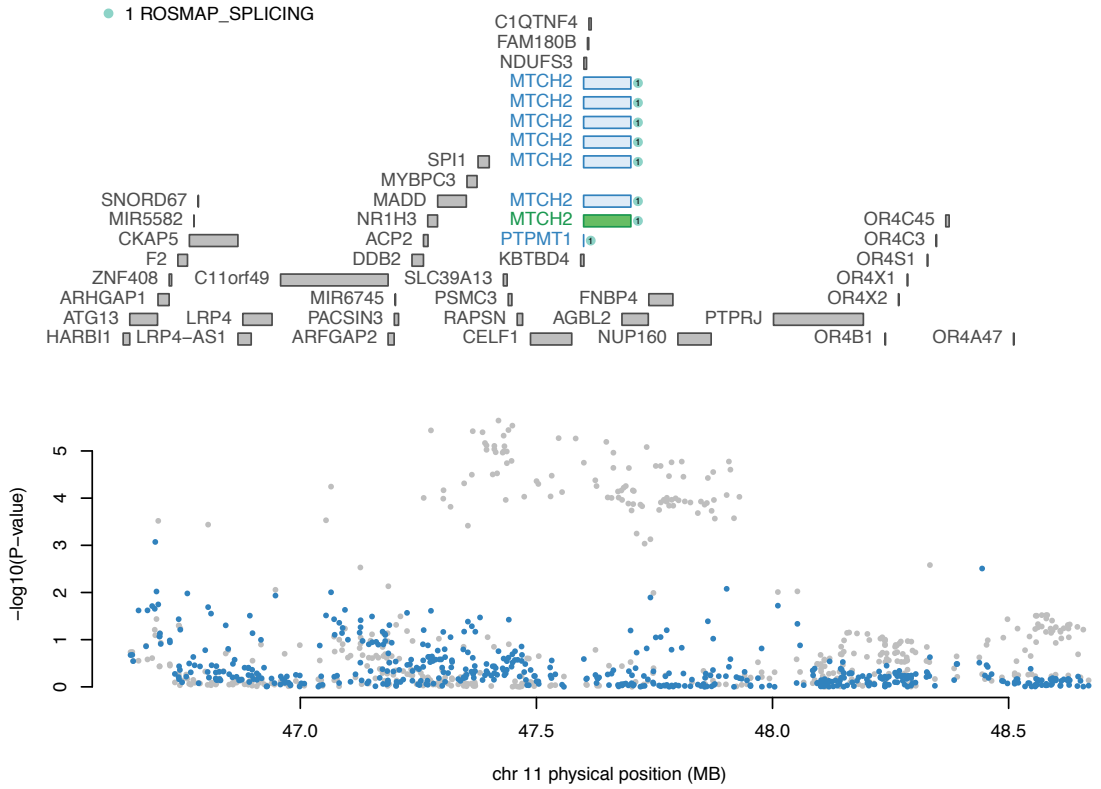

Figure S14. Conditional analysis of IGAP AD GWAS results for splicing effects for *MTCH2.* The AD GWAS effect is mostly explained by splicing effect at *MTCH2* (11:47627806:47637515:clu_6184). Shown here are AD GWAS P-value (in blue) and conditioned P-value (in grey).

References

1. Lim, A.S. *et al.* Diurnal and seasonal molecular rhythms in human neocortex and their relation to Alzheimer's disease. *Nat Commun* **8**, 14931 (2017).
2. Ng, B. *et al.* Brain xQTL map: integrating the genetic architecture of the human brain transcriptome and epigenome *bioArxiv* (2017).
3. Levin, J.Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7**, 709-15 (2010).
4. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
5. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
6. Brown, J., Pirrung, M. & McCue, L.A. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics* (2017).
7. De Jager, P.L. *et al.* Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nat Neurosci* **17**, 1156-63 (2014).
8. Teschendorff, A.E. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**, 189-96 (2013).
9. Mecham, B.H., Nelson, P.S. & Storey, J.D. Supervised normalization of microarrays. *Bioinformatics* **26**, 1308-15 (2010).
10. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-95 (2010).
11. Kharchenko, P.V., Tolstorukov, M.Y. & Park, P.J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* **26**, 1351-9 (2008).
12. Mirra, S.S. *et al.* The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer's disease. *Neurology* **41**, 479-86 (1991).
13. Braak, H., Alafuzoff, I., Arzberger, T., Kretzschmar, H. & Del Tredici, K. Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. *Acta Neuropathol* **112**, 389-404 (2006).
14. Haroutunian, V. *et al.* Regional distribution of neuritic plaques in the nondemented elderly and subjects with very mild Alzheimer disease. *Arch Neurol* **55**, 1185-91 (1998).
15. Morris, J.C. *et al.* The consortium to establish a registry for Alzheimer's disease (CERAD). Part IV. Rates of cognitive change in the longitudinal assessment of probable Alzheimer's disease. *Neurology* **43**, 2457-65 (1993).
16. Liao, Y., Smyth, G.K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-30 (2014).
17. Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci* **19**, 1442-1453 (2016).

18. Powchik, P. *et al.* Postmortem studies in schizophrenia. *Schizophr Bull* **24**, 325-41 (1998).
19. Purohit, D.P. *et al.* Alzheimer disease and related neurodegenerative diseases in elderly patients with schizophrenia: a postmortem neuropathologic study of 100 cases. *Arch Gen Psychiatry* **55**, 205-11 (1998).
20. Kimoto, S., Bazmi, H.H. & Lewis, D.A. Lower expression of glutamic acid decarboxylase 67 in the prefrontal cortex in schizophrenia: contribution of altered regulation by Zif268. *Am J Psychiatry* **171**, 969-78 (2014).
21. Glantz, L.A., Austin, M.C. & Lewis, D.A. Normal cellular levels of synaptophysin mRNA expression in the prefrontal cortex of subjects with schizophrenia. *Biol Psychiatry* **48**, 389-97 (2000).
22. Volk, D.W., Austin, M.C., Pierri, J.N., Sampson, A.R. & Lewis, D.A. Decreased glutamic acid decarboxylase67 messenger RNA expression in a subset of prefrontal cortical gamma-aminobutyric acid neurons in subjects with schizophrenia. *Arch Gen Psychiatry* **57**, 237-45 (2000).
23. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-81 (2011).
24. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
25. Hoffman, G.E. & Schadt, E.E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483 (2016).