De novo mutations involved in post-transcriptional dysregulation contribute to six neuropsychiatric disorders

Fengbiao Mao[1,2¶], Lu Wang[3¶], Xiaolu Zhao[2¶], Luoyuan Xiao[4¶], Xianfeng Li[5], Qi Liu[6], Xin He[7], Rajesh C. Rao[2], Jinchen Li[1], Huajing Teng[1*], Yali Dou[2*] and Zhongsheng Sun[1*]

[1] Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China.

[2] Department of Pathology, University of Michigan, Ann Arbor, MI 48109, USA.

[3] Institute of Life Science, Southeast University, Nanjing 210096, China.

[4] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.

[5] State Key Laboratory of Medical Genetics, Central South University, Changsha, Hunan, 410078, China.

[6] Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA.

[7] Department of Human Genetics, University of Chicago, Chicago, IL, USA.

¶These authors contributed equally to this work

[*] Corresponding authors

E-mail: sunzs@mail.biols.ac.cn (Z.S.S.), yalid@umich.edu (Y.L.D.), tenghj@biols.ac.cn (H.J.T.)

Abstract

While deleterious *de novo* mutations (DNMs) in coding region conferring risk in neuropsychiatric disorders have been revealed by next-generation sequencing, the role of DNMs involved in post-transcriptional regulation in pathogenesis of these disorders remains to be elucidated. By curating 42,871 DNMs from 9,772 core families with six kinds of neuropsychiatric disorders and controls, we identified 2,351 post-transcriptionally impaired DNMs (piDNMs) and prioritized 1,923 candidate genes in these six disorders by employing workflow RBP-Var2. Our results revealed a higher prevalence of piDNM in the probands than that of controls (P = 9.52E-17). Moreover, we identified 214 piDNM-containing genes with enriched co-expression modules and intensive protein-protein interactions (P = 7.75E-07) in at least two of neuropsychiatric disorders. Furthermore, these cross-disorder genes carrying piDNMs could form interaction network centered on RNA binding proteins, suggesting a shared post-transcriptional etiology underlying these disorders. Our findings highlight that piDNMs contribute to the pathogenesis of neuropsychiatric disorders.

Main

In the past decades, there have been increased efforts to better understand the pathogenesis of psychiatric disorders. Next-generation sequencing, which allows genome-wide detection of rare and *de novo* mutations (DNMs), is transforming the pace of genetics of neuropsychiatric disorder by identifying protein-coding mutations that confer risk. Various computational methods have been developed to predict the effects of amino acid substitutions on protein function and to classify corresponding mutations as deleterious or benign. The majority of these approaches rely on evolutionary conservation or protein structural constraints of amino acid substitutions and focus on direct changes in protein coding genes, especially from nonsynonymous mutations which directly affect the gene product[1]. Although previous studies have analyzed the biological and clinical implications of protein truncating mutations identified from next-generation sequencing,[2,3] genetic mutations may not only impact the protein structure or catalytic activity but may also impact transcriptional processses[4,5] via direct or indirect effects on DNA-protein binding[6], histone modifications[7], enhancer cis-regulation[8] and gene-enhancer interactions[9]. Meanwhile, genetic mutations could impair post-transcriptional processes[10,11] such as interactions with RNA binding proteins (RBPs), mRNA splicing, mRNA transport, mRNA stability and mRNA translation.

Hitherto, previous studies of neuropsychiatric disorders have focused on genetic mutation in coding region[12], cis-regulation[13,14], epigenome[15], transcriptome[16,17], and proteome[18], but less is pursued with regard to how these disorders interface with post-transcriptional regulation. Therefore, potential mechanisms of post-transcriptional dysregulation related to pathology and clinical treatment of neuropsychiatric disorders remain largely uninvestigated. Our previous method named RBP-Var[19] and a recent published platform called POSTAR[20]

represents a seminal effort to systematically annotate post-transcriptional regulatory maps and explore the putative roles of RBP-mediated SNVs in human diseases.

To explore association between extremely post-transcriptionally impaired DNMs, namely piDNMs, with psychiatric disorders, we collected whole exome and genome sequencing data from 9,772 core family and curated 42,871 *de novo* mutations from six kinds of neuropsychiatric disorders, including autism spectrum disorder (ASD), epileptic encephalopathy (EE), intellectual disability (ID), schizophrenia (SCZ), brain development defect (DD) and neurodevelopmental defect (NDD) as well as unaffected controls including siblings. By employing our newly updated workflow RBP-Var2, we investigated the potential implication of these *de novo* mutations involved in post-transcriptional regulation in these six neuropsychiatric disorders and found that a subset of *de novo* mutations could be classified as piDNMs. We observed a higher prevalence of piDNMs in the probands of each of all six neuropsychiatric disorders versus controls. In addition, gene ontology (GO) enrichment analyses of cross-disorder genes containing piDNMs in at least two of these disorders revealed several shared crucial pathways involved in the biological processes of neurodevelopment. Meanwhile, weighted gene co-expression network analysis (WGCNA) and protein-protein interactions analysis showed enriched co-expression modules and intensive protein-protein interactions between these cross-disorder genes, respectively, implying that there was convergent machinery of post-transcriptional regulation among six psychiatric disorders. Furthermore, we established an interaction network which is centered on several RBP hubs and encompassed with piDNM-containing genes. In short, DNMs which are deleterious to post-transcriptional regulation extensively contribute to the pathogenesis of neuropsychiatric disorders.

Results

Our ultimate goal is to identify all piDNMs from DNMs in six neuropsychiatric disorders by analyzing a variety of aspects related to post-transcriptional regulation process. To do this, we first developed a comprehensive workflow RBP-Var2 with a statistic algorithm that links the potentially functional roles in post-transcriptional regulation to DNMs. We then used this workflow to identify functional variants from 9,772 core families (7,453 cases and 2,319 controls) with 42,871 *de novo* mutations (29,544 in cases and 13858 in controls, 531 in common) associated with six neuropsychiatric disorders, including 4,209 families and 25,670 DNMs in autism spectrum disorder (ASD), 383 families and 492 DNMs in epileptic encephalopathy (EE), 261 families and 348 DNMs in intellectual disability (ID), 1,077 families and 1,049 DNMs in schizophrenia (SCZ), 1,133 families and 1,531 DNMs in brain development defect (DD), and 390 families and 454 DNMs in neurodevelopmental defect (NDD) (Supplementary Table 1). We further presented an online tool (http://www.rbp-var.biols.ac.cn/) that could rapidly annotate and classify piDNMs, and determine their potential roles in neuropsychiatric disorders.

High prevalence of piDNM in six neuropsychiatric disorders

Firstly, we used our updated workflow RBP-Var2 (see methods) to identify functional piDNMs from 9,772 trios with 42,871 *de novo* mutations (29,041 DNMs in probands and 13,830 DNMs in controls) across six kinds of neuropsychiatric disorders as well as their unaffected controls. We determined DNMs with 1/2 categories predicted by RBP-Var2 as piDNMs and identified 2,351 piDNMs in probands (Supplementary Table 2), of which 64, 18, 2275, 15, 9 were located in 3' UTRs, 5' UTRs, exons, ncRNA exons and splicing sites, respectively. In detail, RBP-Var2 identified 1,410 piDNMs in ASD, 356 piDNMs in DD, 281 piDNMs in SCZ, 103 piDNMs in EE, 129 piDNMs in NDD and 102 piDNMs in ID, whileas only 398 piDNMs were identified in controls groups (Supplementary Table 3) of which 20, 4,

368, 6 were located in 3' UTRs, 5' UTRs, exons and ncRNA exons, respectively. Interestingly, the overall frequency of piDNMs in patients in any one of the six neuropsychiatric disorders were significantly over-represented compared with those in controls (P = 1.07E-16, Figure 1A). With the combined data from 9,772 trios, we also observed that probands groups have much higher odds ratio (OR) of piDNMs compared with controls in all six kinds of neuropsychiatric disorders (Figure 1B; Supplementary Table 4). Dramatically, we found that synonymous piDNMs were significantly enriched in probands in contrast to those in controls (P = 9.53E-14). While in the original DNMs data set before being evaluated by RBP-Var2, the prevalence of those synonymous DNMs was opposite (P = 1.34E-10). To eliminate the effects of loss-of-function (LoF) mutations derived from protein dysfunction, we filtered out those with LoF through all piDNMs and found the non-LoF piDNMs exhibited more significantly higher frequency in probands (P = 3.09E-156) (Supplementary Figure 1), indicating that the significant enrichment of those piDNMs in probands is driven by the contribution of the post-transcriptional effects rather than those LoF DNMs (Figure 1C). In short, piDNMs that are linked to post-transcriptional dysregulation may contribute to the pathogenesis of these six neuropsychiatric disorders.

High sensitivity and specificity of piDNMs identified by RBP-Var2

Our platform RBP-Var2 determined the post-transcriptional dysfunction of piDNMs regardless of considering whether the DNMs give rise to protein truncating. To investigate the accuracy and specificity of DNMs in different regulatory processes, we compared our RBP-Var2 prediction tool with other three popular tools, including SIFT[25], PolyPhen2 (PPH2)[26] and RegulomeDB[21]. RBP-Var2,

RegulomeDB and SIFT/PPH2 were designed to evaluate the impact of mutation on post-transcriptional regulation, transcriptional regulation and protein product, respectively. We found that the frequency of both deleterious nonsynonymous and stop gain/loss DNMs is higher in cases than in controls determined by SIFT (P = 5.37E-62 and P = 4.55E-03, respectively), but only higher frequency of deleterious nonsynonymous DNMs was identified by PPH2 (P = 3.68E-102) (Figure 2A, B). However, RegulomeDB determined significant higher frequency of deleterious nonsynonymous (P = 1.12E-39) and synonymous (P = 3.21E-22), nonframeshift (P < 2.20E-16), splicing (P = 7.62E-04), stop gain/loss (P < 2.20E-16) DNMs rather than frameshift DNMs (P = 4.14E-01) (Figure 2C) in cases compared with controls. In contrast, RBP-Var2 could determine much more deleterious DNMs with the effect of nonsynonymous (P = 6.31E-60) and synonymous (P=8.98E-09), splicing (P < 2.20E-16) and frameshift (P = 1.38E-03) (Figure 2D).

Our results indicated that both RegulomeDB and RBP-Var2 better distinguish deleterious DNMs from benign ones compared to SIFT and PPH2 in cases versus controls. However, the number of deleterious DNMs detected by RegulomeDB is much less than that of RBP-Var2, suggesting that RegulomeDB has a higher false-negative rate than RBP-Var2. Therefore, we performed receiver operating characteristic (ROC) analysis to systemically evaluate the sensitivity and specificity of these four prediction methods. We found that area under curve (AUC) value of SIFT, PPH2, RBP-Var2 and RegulomeDB are 78.27 %, 76.57 %, 82.89 % and 50.77 %, respectively (Supplementary Figure 2). Moreover, the AUC value of SIFT, PPH2, and RBP-Var2 is significantly more sensitive and specific than that of RegulomeDB with P value 1.63E-10, 2.40E-08 and 2.51E-60, respectively. In addition, while the AUC value of RBP-Var2 is significantly higher than that of SIFT and PPH2 with p value 0.049 and 0.019, respectively. Consequently, we compared

the deleterious DNMs detected by these four methods and found that RBP-Var2 detected more piDNMs that overlapped with predictions of SIFT and PPH2 than that of RegulomeDB. Intriguingly, RBP-Var2 could detect an additional 1,238 piDNMs that covered 841 genes and were regarded as benign DNMs by other three methods, accounting for 24.85% of total 4,981 deleterious DNMs detected by all four tools (Supplementary Figure 3A, B). The top-enriched gene ontology of these 841 genes were mitotic cell cycle (P = 2.47E-07), mitotic cell cycle process (P = 3.85E-06), cellular response to stress (P = 5.63E-06), positive regulation of cell cycle (P = 7.50E-06) (Supplementary Figure 3D; Supplementary Table 5), suggesting dysregulation involved in cell cycle and the dysfunction of cellular response may contribute to diverse neural damage, thereby trigger neurodevelopmental disorders[22]. Therefore, the piDNMs detected by RBP-Var2 are distinct and may play significant roles in the post-transcriptional processes of the development of neuropsychiatric disorders.

Shared and distinct features of piDNM across six neuropsychiatric disorders

Our previous study using the NPdenovo database demonstrated that *de novo* mutations predicted as harmful in protein level are shared by four neuropsychiatric disorders[23]. Therefore, we wondered whether there were common piDNMs among six neuropsychiatric disorders. Firstly, we identified 24 genes harboring recurrent piDNMs among 2,351 piDNMs, including nine genes from ASD trios, 10 genes from DD&NDD trios, six genes from ID trios, and one genes from EE trios (Figure 3A; Supplementary Table 6). Only two genes harbored cross-disorder recurrent piDNMs. One was STXBP1 which occurred in both DD and EE, and another was CTNNB1 which occurred in both ASD and DD. However, we identified 312 genes carrying at

least two piDNMs in all disorders among each of 1,923 candidate genes (accounting for 16.22%), including 252 genes involved in ASD, 38 genes involved in EE, 40 genes involved in ID, 68 involved in SCZ trios and 132 genes involved in DD&NDD (Supplementary Table 7). Among these 312 genes, we identified 35 high risk genes with *P* value < 1E-4 derived from TADA program (Transmission And De novo Association)[24] (Figure 3B).

Subsequently, by comparing the genes harboring piDNMs across six disorders, we found 214 genes significantly shared by at least two kinds of disorders rather than random overlaps (permutation test, $P < 1.00E-5$ based on random resampling, Figure 3C, D). Furthermore, the genes harboring piDNMs in controls also have significantly fewer overlaps with the 214 shared genes in the six disorders than random overlaps ($P = 0.007$, Supplementary Figure 4A). In addition, the number of shared genes for any pairwise comparison is significantly higher than random overlaps ($P < 0.05$) except for the comparison of EE versus SCZ ($P = 0.17864$) (Supplementary Figure 5). On the contrary, the genes harboring piDNMs in controls have significantly fewer overlaps with disorders than random overlaps: ASD ($P < 1.00E-5$, Supplementary Figure 4B), ID ($P < 1.00E-5$, Supplementary Figure 4C), EE ($P = 1.20E-4$, Supplementary Figure 4D), SCZ ($P < 1.00E-5$, Supplementary Figure 4E) and DD&NDD ($P < 1.00E-5$, Supplementary Figure 4F). Our observation suggests that there are common genes harboring piDNMs in these six neuropsychiatric disorders.

The shared symptoms in the six neuropsychiatric disorders suggest there is common underlying molecular mechanism that may implicate important regulators of pathogenesis. Thus, we performed ClueGO analysis for these shared genes and found they were mainly enriched in biological processes in epigenetic modification and neuronal and synaptic functions (Supplementary Table 8). These epigenetic

regulating genes are composed of ARID1B, CTCF, CTNNB1, EP300, MECP2, DNMT3A, KMT2D, MYO1E, CNOT1, TNRC18, JARID2, PHF19, TNRC6A, DOT1L, KMT2B (Supplementary Table 9). Moreover, most of these epigenetic modification genes, especially for genes with $P_{\mathrm{TADA}} < 0.05$, have been previously linked with neuropsychiatric disorders[25-28]. Interestingly, shared genes in ClueGO enrichment analyses are intensive linkages among these significant pathways as some of piDNM-containing genes could play roles in more than one of these pathways (Supplementary Figure 7).

Next, to investigate the biological pathways involved in each disorder-specific genes containing piDNM, we carried out GO enrichment analysis with terms in biological process (Supplementary Table 10-12). The most significantly enriched category of ASD-specific genes was "organelle organization" (P = 1.00E-15) and the second top enriched category in the ASD-specific genes was "cell cycle" (P = 2.30E-12). We also found that "chromosome organization" (P = 3.30E-08), "regulation of chromosome organization" (P = 1.80E-05) and "regulation of chromatin organization" (P = 6.60E-05) are the three most significantly dysregulated biological pathways of genes unique to DD and NDD. With respect to genes specific to SCZ, it is actually no surprise that the top three of the enrichment results are related to autophagy, because autophagy is an essential natural destructive process and have been suggested to play a key role in the pathophysiology of schizophrenia[29,30]. Due to the insufficient number of genes, GO enrichment analysis for genes unique to ID or EE was not performed. As each disorder-specific genes containing piDNM could be overrepresented into different biological pathway, our observation suggests that piDNMs may also contribute the distinct phenotype of each of the six psychiatric disorders although the explicit mechanism need to be further explored.

Convergent co-expression modules across six neuropsychiatric disorders

As co-expression of genes has been used to explore the common and distinct molecular mechanism in neuropsychiatric disorders[31], we performed weighted gene co-expression network analysis (WGCNA)[32] with 214 cross-disorder genes containing piDNMs by using up-to-date BrainSpan developmental transcriptome (n=524): gene expression in 16 human brain structures across 31 developmental stages[33]. The results of WGCNA deciphered two gene modules with distinct spatiotemporal expression patterns (Figure 4A, B; Supplementary Figure 8). The turquoise module (n=146 genes) is characterized by high expression during early fetal development (8-37 postconceptional weeks) in most of brain structures (Figure 4C). The blue module (n=58 genes) is delineated to low expression in early fetal development (postconceptional week 8) and early childhood (3 year) in most of brain structures (Figure 4D). The turquoise module was enriched for epigenetic regulation of gene expression (q = 3.99E-08), establishment or maintenance of cell polarity (q = 8.24E-04), Notch binding (q = 1.33E-3), Thyroid hormone signaling pathway (q = 1.75E-3), beta-catenin-TCF complex assembly (q = 5.81E-3), chromatin remodeling (q = 1.30E-2), Notch signaling pathway (q = 2.51E-2), motor activity (q = 3.21E-2) and other enriched GO terms (Supplementary Figure 9; Supplementary Table 13). The blue module was enriched visual learning (q = 4.92E-07), regulation of synaptic plasticity (q = 1.91E-06), actin filament-based movement (q = 6.52E-06), adult locomotory behavior (q = 3.86E-04), neuromuscular process (q = 8.61E-04), Long-term depression (q = 1.02E-3), positive regulation of dendrite development (q = 1.35E-3), negative regulation of autophagy (q = 3.74E-3), negative regulation of G-protein coupled receptor protein signaling pathway (q = 5.07E-3)

and other neurodevelopment related pathways (Supplementary Figure 10; Supplementary Table 14).

In addition, we constructed the co-expression network by using gene pairs with correlation weight of at least 0.3 (Supplementary Table 15). Out of the 214 cross-disorder genes containing piDNMs, 206 were clustered in the co-expression networks (Supplementary Table 16). Interestingly, we found that majority of the co-expression hubs were histone modifiers, such as KAT6B, KDM5B, SETDB1, TRRAP, EP300, and transcriptional regulators, including CTCF, SMARCA4, ARID1B, ADNP, CHAMP1 (Supplementary Figure 11). Consequently, the link among these six highly heterogeneous neuropsychiatric disorders may be represented by biological processes controlled by these hub genes in co-expression networks. Moreover, by employing our newly developed EpiDenovo database[34], we found that most of the genes were highly expressed in the early embryonic development (Supplementary Figure 12), indicating that these genes are crucial in cell differentiation.

Common network of protein-protein interactions across six neuropsychiatric disorders

The co-expression results indicated that these 214 piDNM-related cross-disorder proteins may have intensive protein-protein interactions (PPI). To identify common biological processes that potentially contribute to disease pathogenesis, we investigated protein-protein interactions within these 214 cross-disorder genes containing piDNMs by database BioGRID. Our results revealed that 128 out of 214 (59.81%) cross-disorder genes represent an interconnected network on the level of available direct/genetic protein-protein interactions (Figure 5A; Supplementary Table 17). Furthermore, we determined several crucial hubs of protein-protein

interactions, including CUL3, CTNNB1, HECW2, HNRNPU, EP300, SMARCA4, ARRB2, PTPRF, MYH9 and NOTCH1 (Figure 5A), which may control common biological processes among these six neuropsychiatric disorders. Indeed, these 128 cross-disorder proteins are enriched in nervous system phenotypes, including abnormal synaptic transmission, abnormal nervous system development, abnormal neuron morphology and abnormal brain morphology, and behavior/neurological phenotype such as abnormal motor coordination/balance (two sided Fisher's Exact Test, $q < 0.05$, Supplementary Figure 13A). Similarly, these 128 genes in interaction network are enriched in nervous system phenotype including abnormal nervous system development and abnormal brain morphology (two sided Fisher's Exact Test, $q < 0.05$, Supplementary Figure 13B). Dramatically, HNRNPU belongs to the subfamily of ubiquitously expressed heterogeneous nuclear ribonucleoproteins (hnRNPs) which are RNA binding proteins and they form complexes with heterogeneous nuclear RNA (hnRNA). These proteins are associated with pre-mRNAs in the nucleus and appear to influence pre-mRNA processing and other aspects of mRNA metabolism and transport.

By investigating the expression of these co-interacting genes in the human cortex of 12 ASD patients (accession number GSE64018) and 13 normal donators (accession number GSE76852) from Gene Expression Omnibus (GEO), we identified 97 significantly differential expression genes (accounting for 75.78% of the 128 PPI genes we mentioned above) between ASD patients and normal controls (Student's t-test, $q < 0.05$, Figure 5B). And 86 of these PPI genes were highly expressed in ASD patients while only 11 genes were down-regulated in ASD patients when compared with normal controls (Supplementary Table 18), implying that most of these PPI genes were abnormally expressed in ASD patients.

Shared networks between piDNM and RBPs

According to previous studies, dysregulation or mutations of RBPs can cause a range of developmental and neurological diseases[35,36]. Meanwhile, mutations in RNA targets of RBPs, which could disturb the interaction between RBPs and their mRNA targets, would affect mRNA metabolism and protein homeostasis in neurons by impair RNA transport and translation in neuropathological disorders[37-39]. Indeed, mutations that alter sequence context of RBP binding sites on target RNA commonly affect RBP binding and regulation[40]. For instance, regulation depending on FMRP-related activity during fetal brain development might be particularly vulnerable to genetic perturbations, with severe mutations resulting in disruption of developmental canalization[41]. Hence, we constructed a regulatory network between piDNMs and RBPs based on predicted binding sites of RBPs to investigate the genetic perturbations of mRNA-RBP interactions in six disorders (Figure 6). We identified several crucial common RBP hubs that contribute to the pathogenesis of the six neuropsychiatric disorders, including EIF4A3, FMR1, PTBP1, AGO1/2, ELAVL1, IGF2BP1/3, WDR33 and FXR2. Genes with piDNMs in different disorders could be regulated by the same RBP hub while one candidate genes may be regulated by different RBP hubs (Figure 6). In addition, all of these RBP hubs were highly expressed in early fetal development stages (8-37 postconceptional weeks) based on BrainSpan developmental transcriptome (Supplementary Figure 14), suggesting that these RBP hubs may play important roles in the early stages of brain development.

The available data resource

To make our findings easily accessible to the research community, we have developed RBP-Var2 platform (http://www.rbp-var.biols.ac.cn/) for storage and retrieval of piDNMs, candidate genes, for exploring the genetic etiology of neuropsychiatric disorders in post-transcriptional regulation. The expression and

epigenetic profile of genes related to regulatory *de novo* mutations and early embryonic development have been deposited in our previously published database EpiDenovo (http://www.epidenovo.biols.ac.cn/)[34].

Discussion

In this study, we systematically analyzed the damaging effect of *de novo* mutations at the level of post-transcriptional regulation in six neuropsychiatric disorders. We discovered 2,351 piDNMs in the total of 1,923 genes that displayed deleterious effect on both post-transcriptional regulation and protein function, of which 1,034 (43.43%) piDNMs were also determined as extremely deleterious on the structure and function of protein predicted by NPdenovo, thereby resulting in identifying the rest of 1,317 piDNMs to have impact on post-transcriptional regulation only. Among the total 1,923 genes containing piDNMs, 862 (44.83%) genes were predicted to simultaneously alter the structure and function of correspondingly coded proteins predicted by NPdenovo. Our observations indicate that 55.17% of piDNMs are not expected to affect the structure and function of protein but are predicted to affect post-transcriptional regulation. Moreover, probands have higher odds ratio of piDNMs than controls in all kinds of neuropsychiatric disorders. Therefore, piDNMs are one of major contributors to etiology for neuropsychiatry disorders, which is distinct from the mutation effect on protein functions.

We applied RBP-Var2 algorithm to annotate and interpret *de novo* variants in subjects with six neuropsychiatric disorders and control subjects. We observed a strong enrichment of pathogenic or likely pathogenic variants in affected subjects. In comparison with accuracy of RBP-Var2 (82.89%), other prediction algorithms such as SIFT or PPH2 have moderate accuracy (76~78%) while RegulomeDB has

much lower accuracy (50.77 %) to differentiate affected from the control subjects. This observation suggests that the potential pathogenicity of genetic variants is mainly due to post-transcriptional dysregulation and malfunction of protein, but not disruption of transcription process. Therefore, deciphering multiple biological layers of deleteriousness, particularly post-transcriptional regulation and structure/function of protein, may improve the accuracy to predict disease related genetic variations.

To date, it has been a challenge to estimate the deleteriousness of synonymous and UTRs mutations though such mutations have been widely acknowledged to alter protein expression, conformation and function[42]. Nevertheless, our RBP-Var2 tool identified 518 synonymous DNMs and 82 UTR's DNMs, which were extremely deleterious in post-transcriptional regulation. Moreover, synonymous damaging DNMs were significantly prominent in probands compared with that in controls, supported by predictions of both RBP-Var2 and RegulomeDB. Meanwhile, *de novo* insertions and deletions (InDels), especially frameshift patterns are taken for granted to be deleterious. Indeed, *de novo* frameshift InDels are more frequent in neuropsychiatric disorders compared to non-frameshift InDels[43], which were demonstrated by predictions of RBP-Var2 but not SIFT or PPH2 (Figure 2).

Most interestingly, we discovered that some epigenetic pathways are enriched among these piDNM-containing genes, such as those that regulate gene expression and histone modification. This finding is consistent with a previous report in which more than 68% of ASD cases shared a common acetylome aberrations at >5,000 cis-regulatory regions in prefrontal and temporal cortex[44]. Such common "epimutations" may be induced by either perturbations of epigenetic regulations (including post-transcriptional regulations) due to genetic mutations of substrates or the disruptions of epigenetic modifications due to the genetic mutation of epigenetic genes. Actually, our observations suggest that alterations of "epimutations" were associated with the

dysregulation of post-transcription. This hypnosis is constant with the observation that several recurrent piDNM-containing genes are non-epigenetic genes, including CHD8, CHD2, SYNGAP1, ADNP, POGZ, ANK2 and DYRK1A. Moreover, we also discovered several recurrent epigenetic genes including KMT2A, KMT2B, KMT2C, KAT6B, KDM3B, JARID2, DNMT3A and MECP2, that contain piDNMs which may play important roles in the genome-wide aberrations of epigenetic landscapes through disruption of the post-transcription regulation. Furthermore, WGCNA analysis revealed that major hubs of the co-expression network for these 214 piDNM-containing genes were histone modifiers by using BrainSpan developmental transcriptome. These data indicate that piDNM-containing genes are co-expressed with genes frequently involved in epigenetic regulation of common cellular and biological process in neuropsychiatric disorders.

Importantly, these 214 piDNM-containing genes harbor intensive protein-protein interactions in physics and shared regulatory networks between piDNMs and RBPs in six neuropsychiatric disorders. We identified several RBP hubs of regulatory networks between piDNM-containing genes and RBP proteins, including EIF4A3, FMRP, PTBP1, AGO1/2, ELAVL1, IGF2BP1/3, WDR33 and FXR2. Taking FMRP for example, it is a well-known pathogenic gene of Fragile X syndrome which co-occurs with autism in many cases and its targets are highly enriched for *de novo* mutations in ASD[41]. The mutations of any RBP hubs may result in multiple disorders and mutations of RBP-targeting genes may disrupt interactions with multiple RBPs. Fortunately, we have constructed the regulatory networks among multiple RBPs and their target RNAs with *de novo* mutations that could significantly disturb the regulatory networks in six neuropsychiatric disorders.

Alterations in expression or mutations in either RBPs or their binding sites in target transcripts have been reported to be the cause of several human diseases such

as muscular atrophies, neurological disorders and cancer[45]. However, it is still a challenge to decipher the effect of genetic mutation on post-transcriptional regulation. In this study, our method sheds light on evaluation of post-transcriptional impact of genetic mutations especially for synonymous mutations. In addition, as small molecules can be rapidly designed to selectively target RNAs and affect RNA-RBP interactions, leading to anti-cancer strategies[46]. Therefore, the discovery of disease-causing mutations in RNAs is yielding a wealth of new potential therapeutic targets, providing new RNA-based tools for developing therapeutics.

Materials and methods

Data collection.

For this study, 9,772 trios were recruited from previous WES/WGS studies, comprising 7,453 parent-probands trios associated with six neuropsychiatric disorders and 2,319 control trios. After deduplication, a total of 29,041 DNMs in probands and 13,830 DNMs in controls were identified for subsequent analysis.

RBP-Var2 algorithm

To better interpret and comprise the catalog of *de novo* mutation, we developed a new heuristic scoring system based on the functional confidence of variants and mathematical algorithms. The scoring system represents with increasing confidence that a variant lies in a functional location and probably results in a functional consequence (i.e. alteration of RBP binding and a gene regulatory effect). For example, we consider variants that are known eQTLs (expression quantitative trait locus) as significant and label them as category 1. Within category 1, subcategories indicate additional annotations ranging from the most informational variants (1a, variant may change the motif for RBP binding) to the least informational variants (1e, variant only has a motif for RBP binding). In mathematical algorithms, we employed LS-GKM[47,48] to predict the impact of DNMs on the binding of specific RBPs by calculating the delta SVM scores. Moreover, for single-base mutations, we employed the RNAsnp[49] with default parameters to estimate the mutation effects on local RNA secondary structure in terms of empirical P-values based on global folding in correlation measured by using RNAfold[50]. For insertions and deletions, based on changes of minimal free energy, we evaluated the effects on RNA secondary structure in terms of empirical P-values calculated from cumulative probabilities of the Poisson distribution. Only the functional DNM produces >1

change in gkm-SVM scores for the effect of RBP binding or *P*-value < 0.1 for the effect of RNA secondary structure change were determined to be a piDNM. Only DNMs occurred in exonic or UTR regions were included in our analysis process.

Identification of deleterious piDNMs and comparison with variants predicted by other methods.

To determine the likelihood of causing a deleterious mutation in post-transcriptional regulation for all SNVs and InDels, our newly updated program RBP-Var2 was utilized to assign an exclusive rank for each mutation and only those mutations categorized into rank 1 or 2 were considered as piDNMs. In comparison with those mutations involved in the disruption of gene function or transcriptional regulation, several programs such as SIFT, PolyPhen2 and RegulomeDB were used to analyze the same dataset of DNMs as the input for RBP-Var2. We only kept the mutations qualified as "damaging" from the result of SIFT and "possibly damaging" or "probably damaging" from PolyPhen2. In the case of RegulomeDB, mutations labeled as category 1 and 2 were retained. Next, we classified the type of mutation (frameshift, nonframeshift, nonsynonymous, synonymous, splicing and stop) and located regions (UTR3, UTR5, exonic, ncRNA exonic and splicing) in order to determine distribution of piDNMs, genetic variants and other regulatory variants. The number of variants in cases versus controls was illustrated by bar chart (***: $P < 0.001$, **: $0.001 < P < 0.01$, *: $0.01 < P < 0.05$, binomial test).

TADA analysis of DNMs in six disorders

The TADA program (Transmission And De novo Association), which predicts risk genes accurately on the basis of allele frequencies, gene-specific penetrance, and mutation rate, was used to calculate the P-value for the likelihood of each gene contributing to the all six disorders with default parameters.

ROC curves and specificity/sensitivity estimation

We screened a positive (non-neutral) test set of likely casual mutations in Mendelian disease from the ClinVar database (v20170130). From a total of 237,308 mutations in ClinVar database, we picked up 145 exonic mutations presenting in our curated DNMs in probands. Our negative (neutral) set of likely non-casual variants was built from DNMs of unaffected siblings in six neuropsychiatric disorders. In order to avoid rare deleterious DNMs, we selected only DNMs in controls with a minor allele frequency of at least 0.01 in 1000 genome (1000g2014oct), resulting in a set of 921 exonic variants. Then, we employed R package pROC to analyze and compare ROC curves.

Overlaps of piDNMs between disorders.

In order to elucidate the overlap of genes among any two of the six disorders as well as each disorder and the control, we shuffled the intersections of genes and repeated this procedure 100,000 times. During each permutation, we randomly selected the same number of genes as the actual situation from the entire genome for each disorder and the control, then *p* values were calculated as the proportion of permutations during which the observed number of genes was above/below the number of genes in simulated situations.

Functional enrichment analysis

A gene harboring piDNMs would be selected into our candidate gene set to conduct functional enrichment analysis if it occurred in at least two of the six disorders. GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway enrichments analysis was implemented by Cytoscape (version 3.4.0) plugin ClueGO (version 2.3.0) and *P* values calculated by hypergeometric test was corrected to be q values by Benjamini–Hochberg procedure for reducing the false discovery rate resulted from multiple hypothesis testing.

Co-expression and spatiotemporal specificity

Normalized gene-expression of 16 human brain regions were determined by RNA sequencing and obtained from BrainSpan (http://www.brainspan.org). We extracted expression for 206 out of 214 extreme damaging cross-disorder genes and employed R-package WGCNA (weighted correlation network analysis) with a power of five to cluster the spatiotemporal-expression patterns and prenatal laminar-expression profiles. The expression level for each gene and development stage (only stages with expression data for all 16 structures were selected, n = 14) was presented across all brain regions.

Protein-protein interaction and phenotype enrichment

Cross-disorder genes containing piDNMs were subjected to esyN analysis for generating a network of protein interactions and the network of protein interaction was created by using physical and genetic interactions of H. sapiens curated in BioGRID. Cytoscape (version 3.4.0) was used to analyze and visualize protein-protein interaction networks. Overrepresentation of mouse-mutant phenotypes was evaluated using the web tool MamPhea for the genes in the PPI network and for all cross-disorder genes containing piDNMs. Rest of genome was used as background and q values were calculated from *P* values by Benjamini-Hochberg correction.

Gene-RBP interaction network

Cytoscape (version3.4.0) was utilized for visualization of the associations between genes harboring piDNMs in the six neuropsychiatric disorders and the corresponding regulatory RBPs.

URLs

RBP-Var2, http://www.rbp-var.biols.ac.cn/ ; NPdenovo, http://www.wzgenomics.cn/NPdenovo/index.php; EpiDenovo: http://www.epidenovo.biols.ac.cn/; BioGRID, https://thebiogrid.org/; MamPhea , http://evol.nhri.org.tw/phenome/index.jsp?platform=mmus; BrainSpan, http://www.brainspan.org; ClinVar, https://www.ncbi.nlm.nih.gov/clinvar/; 1000Genomes, http://www.internationalgenome.org/; WGCNA, https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/; esyN, http://www.esyn.org/; Cytoscape, http://www.cytoscape.org/; TADA, http://wpicr.wpic.pitt.edu/WPICCompGen/TADA/TADA_homepage.htm; ClueGO, http://apps.cytoscape.org/apps/cluego; pROC, http://web.expasy.org/pROC/; R, https://www.r-project.org/; Perl, https://www.perl.org/.

Contributions

F.M. and L.W. participated in the design and execution of analyses, produced the figures, participated in the interpretation of results and edited the manuscript. F.M. developed computational code employed in the analyses. L.W. developed the statistical framework and drew the figures. X.Z. participated in the interpretation of results, the oversight of analyses. L.X., X. L. and Q.L. developed and improved the online platform of RBP-Var2. X.H., H.J.T. and R.C.R. provided professional guidance in the writing and refining of the manuscript. J.L. and H.J.T. collected the DNMs from literature and database. Z.S.S. and Y.L.D. conceived the study, participated in the design of analyses, oversaw the study and the interpretation of results, and drafted and edited the manuscript. All authors reviewed and edited the final manuscript.

Acknowledgments

The project was funded by National Key R&D Program of China (No. 2016YFC0900400). We thank Kun Zhang for his help in TADA analysis and thank Leisheng Shi for his help in data analysis. It is greatly appreciated for the professional suggestions from Professor Weixiang Guo in the Institute of Genetics and Developmental Biology, Chinese Academy of Sciences.

Competing interests

The authors declare no competing financial interests.

Figure legends

Figure 1. The abundance of DNMs in different RBP-Var2 categories. (A) The quantities of DNMs in patients with neuropsychiatric disorders are significantly larger than that in normal controls, shown in several categories classified by RBP-Var2. (B) The bar plot corresponds to the odds ratios indicating the enrichment of piDNMs in patients from each of the five neuropsychiatric disorders. (C) The relative amount of LoF and non-LoF piDNMs in five neuropsychiatric disorders.

Figure 2. Performance comparison of the ability to distinguish severe DNMs between RBP-Var2 and three other tools. (A) Different kinds of DNMs affecting protein function predicted by SIFT. The Y-axis corresponds to the proportion of each kind of mutations within the total number of damaging DNMs predicted by SIFT. (B) Different kinds of DNMs that affect protein function predicted by PolyPhen2. The Y-axis corresponds to the proportion of each kind of mutations within the total number of damaging DNMs predicted by PolyPhen2. (C) The DNMs predicted as functional elements involved in transcriptional regulation by RegulomeDB are categorized into different functional types. The Y-axis corresponds to the proportion of each kind of mutations within the total number of damaging DNMs predicted by

RegulomeDB. (D) The DNMs classified as either level 1 or 2 (piDNMs) are categorized into different functional types. The Y-axis corresponds to the proportion of each kind of mutations within the total number of damaging piDNMs. The P values were measured by two-sided binomial test. DNMs predicted in both cases and controls are excluded in the comparison and the DNMs labeled as "unknown" are not demonstrated in the bar plot.

Figure 3. Candidate genes selected by RBP-Var2 involved in six neuropsychiatric disorders. (A) Scatter plot of 24 genes harboring recurrent piDNMs among 2,351 piDNMs. The Y-axis corresponds to the -log10(P value) calculated by TADA. The X-axis stands for the TADA output of -log10(mutation rate). (B) Scatter plot of 312 recurrent genes among 1,923 candidate genes. The Y-axis corresponds to the -log10(P value) calculated by TADA. The X-axis stands for the TADA output of -log10(mutation rate). (C) Venn diagram representing the distribution of candidate genes shared among five neuropsychiatric disorders. (D) Permutation test for the validity of the overlap between the candidate genes involved in the five neuropsychiatric disorders. We shuffled the genes of each disorder and calculated the shared genes between the five disorders, repeating this procedure for 100,000 times to get the null distribution. The vertical dash line stands for the observed value corresponding to a P value of the permutation test.

Figure 4. Weighted co-expression analysis of 214 shared genes. (A) Heat map visualization of the co-expression network of 214 shared genes. The more saturated color corresponds to the highly expressed genes. (B) Hierarchical clustering dendrogram of the two color-coded gene modules displayed in (A). (C, D) The two spatiotemporal expression patterns (Turquoise module and Blue module) for network genes based on RNA-seq data from BrainSpan, and corresponding to 17 developmental stages across 16 subregions. (E, F) Characterization of neocortical

expression profiles (Turquoise module and Blue module) for 214 piDNM-related genes. Four subregions of the developing neocortex, delineating nine layers per subregion, were analyzed. A1C, primary auditory cortex; AMY, amygdaloid complex; CPo, outer cortical plate; CPi, inner cortical plate; DFC, dorsolateral prefrontal cortex; HIP, hippocampus; IPC, posteroinferior parietal cortex; ITC, inferolateral temporal cortex; M1C, primary motor cortex; MD, mediodorsal nucleus of thalamus; MFC, anterior cingulate cortex; MZ, marginal zone; OFC, orbital frontal cortex; STC, posterior superior temporal cortex; SG, subpial granular zone; SP, subplate zone; IZ, subplate zone; STR, striatum; SZo, outer subventricular zone; SZi, inner subventricular zone; S1C, primary somatosensory cortex; V1C, primary visual cortex; VFC, ventrolateral prefrontal cortex; VZ, ventricular zone.

Figure 5. Protein-protein interaction network analysis. (A) The network of interactions between pairs of proteins of 128 out 214 shared genes. (B) Heat map showing significantly differential expression of 97 out of 128 genes involved in the protein-protein interaction network.

Figure 6. Interaction network of RBPs and genes with piDNMs. Different roles of the nodes are reflected by distinguishable geometric shapes and colors. The magenta vertical arrow stands for the RNA binding proteins. Disks with different colors represent the genes with piDNMs involved in different kinds of disorders.

Supplementary Figure 1. Excess of piDNMs in probands. The distinction of synonymous mutations and mutations in UTRs were analyzed in DNMs and piDNMs between probands and healthy controls. Overall moderate piDNMs and piDNMs (LoF mutations not included) were also displayed. P values were calculated assuming a binomial distribution.

Supplementary Figure 2. ROC curve showing the performance of the predictions of SIFT, PPH2, RBP-Var2 and RegulomeDB.

Supplementary Figure 3. Overlap of DNMs identified by different tools. (A) Venn diagram depicting the overlap between the DNMs predicted by SIFT, PPH2, RBP-Var2 and RegulomeDB. (B) Venn diagram depicting the overlap between the genes predicted by SIFT, PPH2, RBP-Var2 and RegulomeDB. (C) The pie chart shows the distribution of all non-LoF piDNMs. The non-LoF piDNMs detected by RBP-Var2 alone account for 52% in this distribution (light red), while the non-LoF piDNMs identified by SIFT and Polyphen2 both take up 27.6% of all (dark moderate blue). (D) Pathway enrichment analysis of the 841 genes unique to the prediction of RBP-Var2.

Supplementary Figure 4. Test of the significance of the genes shared between each pair of the five disorders. (A-G) Permutation test for the validity of the gene overlap between each pair of the five disorders. We shuffled the genes of each disorder and calculated the shared genes between each pair, repeating this procedure for 100,000 times to get the null distribution. The vertical dash line stands for the observed value corresponding to a P value of the permutation test.

Supplementary Figure 5. Test of the significance of the 214 candidate genes involved in the five neuropsychiatric disorders. (A-E) Permutation test for the validity of the gene overlap between each disorder and the normal control. (F) Permutation test for the validity of the gene overlap between the 214 shared genes and the normal control.

Supplementary Figure 6. Pie chart of the pathway enrichment analysis for the 214 shared genes.

Supplementary Figure 7. Interaction network of the gene enrichment analysis for the 214 shared genes.

Supplementary Figure 8. Relationship between Co-expression modules. (A) MDS plot of genes in turquoise module and blue module. (B) Relationship between module eigengenes. (C) Clustering tree based of the module eigengenes. (D) heatmap of adjacency Eigengene.

Supplementary Figure 9. The two developmental expression patterns (Turquoise module and Blue module) for network genes based on RNA-seq data from EpiDenovo.

Supplementary Figure 10. The co-expression network of 206 out of the 214 shared genes. Different size of the node is representative of the number of connections between the gene and others.

Supplementary Figure 11. Pie chart showing the enrichment analysis of the genes clustered in the turquoise module from the weighted co-expression analysis.

Supplementary Figure 12. Pie chart showing the enrichment analysis of the genes clustered in the blue module from the weighted co-expression analysis.

Supplementary Figure 13. Mammalian phenotype enrichment analysis of selected genes. (A) Mammalian phenotype enrichment of 214 cross-disorder piDNMs genes. (B) Mammalian phenotype enrichment of 128 genes in interaction network.

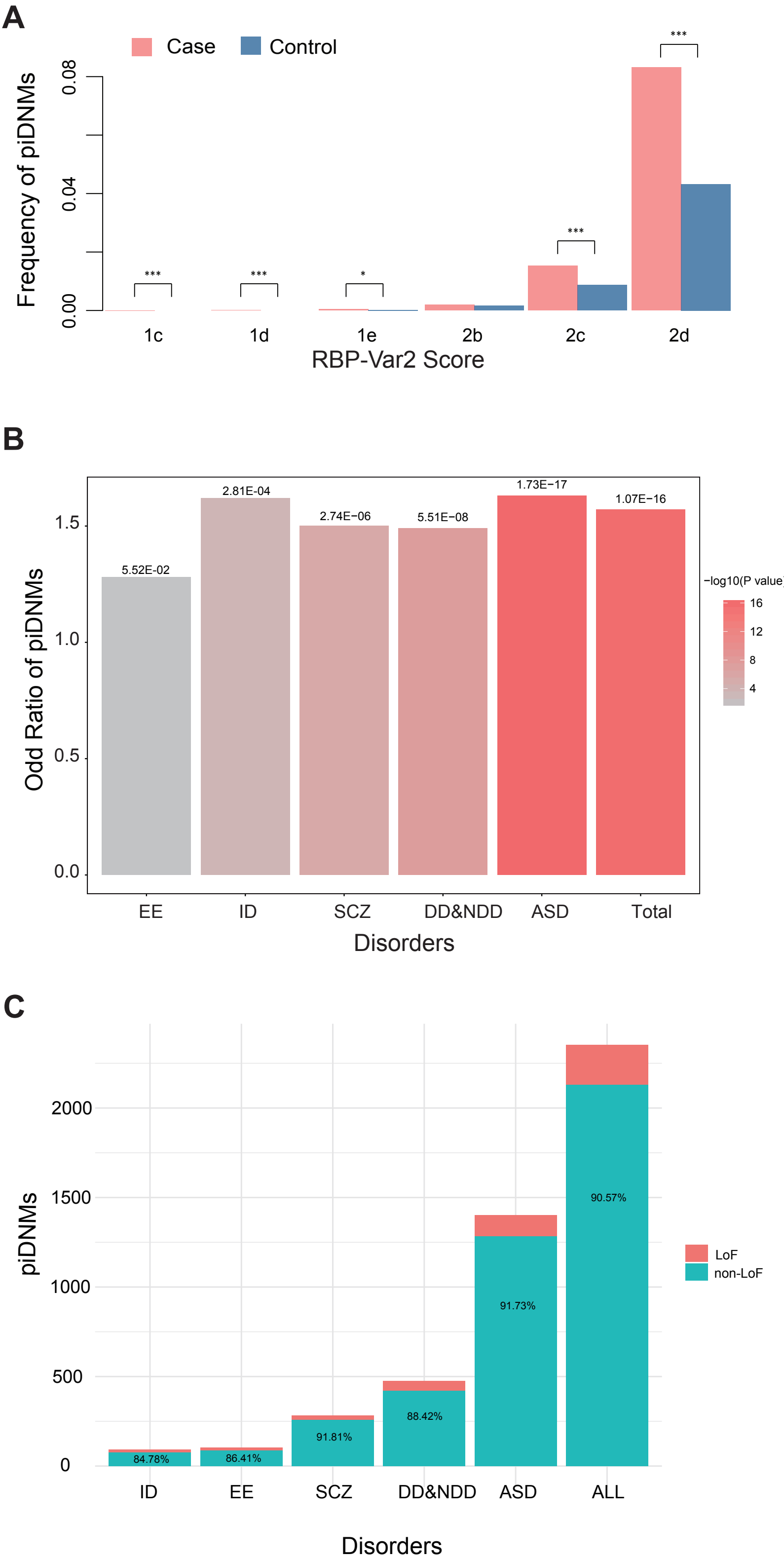
Supplementary Figure 14. Heat map of the expression of the crucial RBP hub genes during the early fetal development stages.

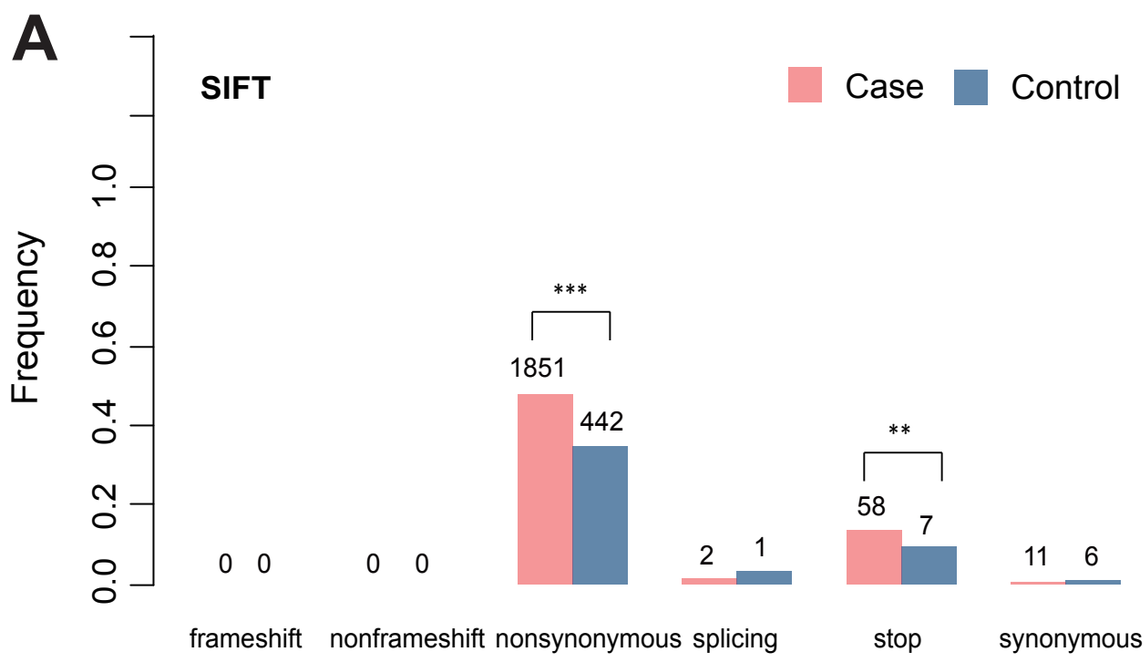
Reference

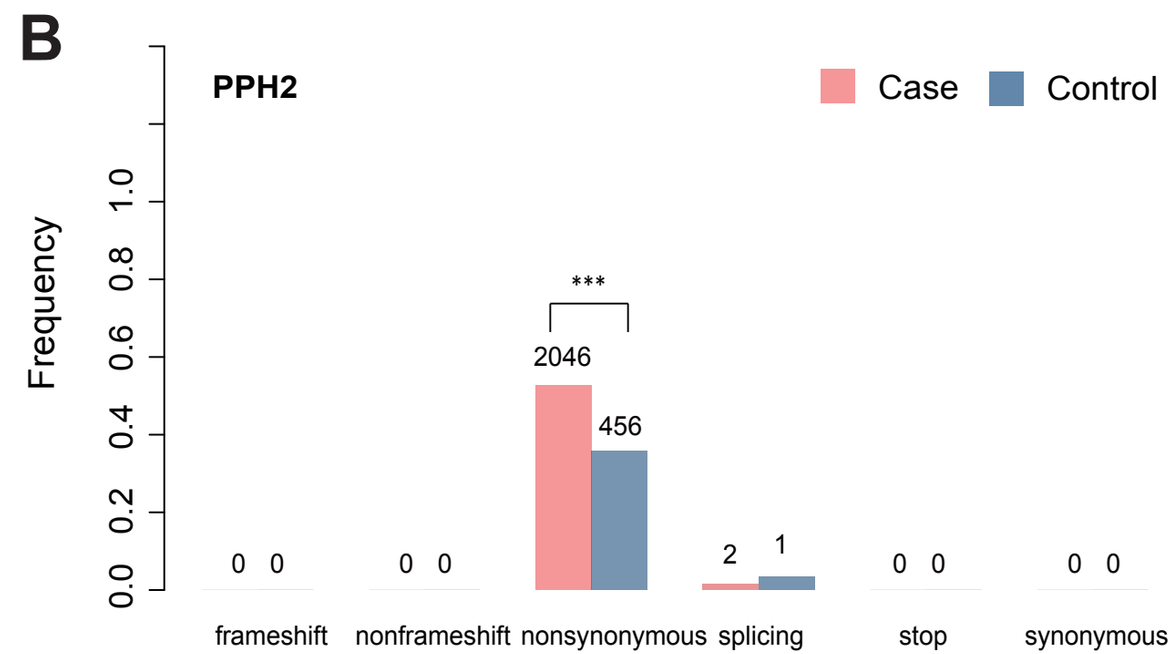
1. Gnad, F., Baucom, A., Mukhyala, K., Manning, G. & Zhang, Z.M. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *Bmc Genomics* **14**(2013).
2. Sullivan, P.F., Daly, M.J. & O'Donovan, M. DISEASE MECHANISMS Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Reviews Genetics* **13**, 537-551 (2012).
3. Heinzen, E.L., Neale, B.M., Traynelis, S.F., Allen, A.S. & Goldstein, D.B. The Genetics of Neuropsychiatric Diseases: Looking In and Beyond the Exome. *Annual Review of Neuroscience, Vol 38* **38**, 47-68 (2015).
4. Chen, X.F., Zhang, Y.W., Xu, H.X. & Bu, G.J. Transcriptional regulation and its misregulation in Alzheimer's disease. *Molecular Brain* **6**(2013).
5. Lee, T.I. & Young, R.A. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **152**, 1237-1251 (2013).
6. Mathelier, A. *et al.* Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. *Genome Biology* **16**(2015).
7. Portela, A. & Esteller, M. Epigenetic modifications and human disease. *Nature Biotechnology* **28**, 1057-1068 (2010).
8. Sakabe, N.J., Savic, D. & Nobrega, M.A. Transcriptional enhancers in development and disease. *Genome Biology* **13**(2012).
9. Lupianez, D.G. *et al.* Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell* **161**, 1012-1025 (2015).
10. Linder, B., Fischer, U. & Gehring, N.H. mRNA metabolism and neuronal disease. *Febs Letters* **589**, 1598-1606 (2015).
11. Halbeisen, R.E., Galgano, A., Scherrer, T. & Gerber, A.P. Post-transcriptional gene regulation: From genome-wide studies to principles. *Cellular and Molecular Life Sciences* **65**, 798-813 (2008).
12. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-21 (2014).
13. Gamazon, E.R. *et al.* Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Molecular Psychiatry* **18**, 340-346 (2013).
14. Bonder, M.J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat Genet* (2016).
15. Loke, Y.J., Hannan, A.J. & Craig, J.M. The Role of Epigenetic Change in Autism Spectrum Disorders. *Front Neurol* **6**, 107 (2015).
16. Gupta, S. *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat Commun* **5**, 5748 (2014).
17. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380-4 (2011).
18. Corbett, B.A. *et al.* A proteomic study of serum from children with autism showing differential expression of apolipoproteins and complement proteins. *Mol Psychiatry* **12**, 292-306 (2007).
19. Mao, F.B. *et al.* RBP-Var: a database of functional variants involved in regulation mediated by RNA-binding proteins. *Nucleic Acids Research* **44**, D154-D163 (2016).
20. Hu, B., Yang, Y.-C.T., Huang, Y., Zhu, Y. & Lu, Z.J. POSTAR: a platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Research*, gkw888 (2016).

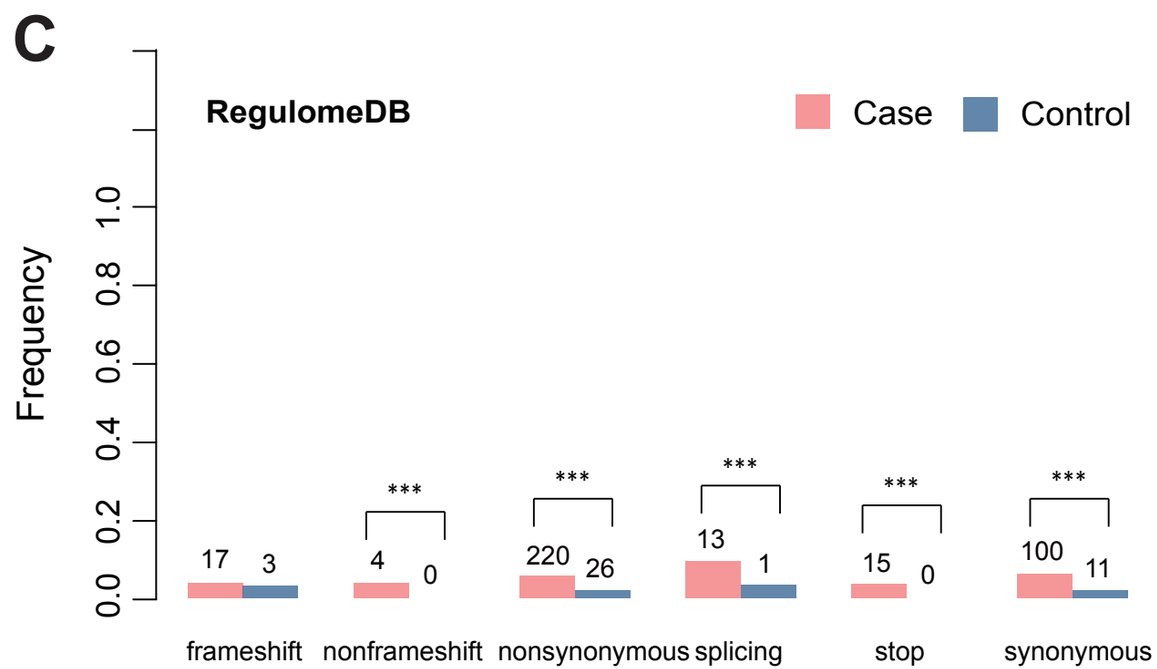
21. Boyle, A.P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research* **22**, 1790-1797 (2012).
22. Wang, W. *et al.* Neural cell cycle dysregulation and central nervous system diseases. *Prog Neurobiol* **89**, 1-17 (2009).
23. Li, J.C. *et al.* Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. *Molecular Psychiatry* **21**, 290-297 (2016).
24. He, X. *et al.* Integrated Model of De Novo and Inherited Genetic Variants Yields Greater Power to Identify Risk Genes. *Plos Genetics* **9**(2013).
25. Hsieh, J. & Eisch, A.J. Epigenetics, hippocampal neurogenesis, and neuropsychiatric disorders: Unraveling the genome to understand the mind. *Neurobiology of Disease* **39**, 73-84 (2010).
26. Brookes, E. & Shi, Y. Diverse Epigenetic Mechanisms of Human Disease. *Annual Review of Genetics, Vol 48* **48**, 237-268 (2014).
27. Jakovcevski, M. & Akbarian, S. Epigenetic mechanisms in neurological disease. *Nature Medicine* **18**, 1194-1204 (2012).
28. Peter, C.J. & Akbarian, S. Balancing histone methylation activities in psychiatric disorders. *Trends in Molecular Medicine* **17**, 372-379 (2011).
29. Merenlender-Wagner, A. *et al.* Autophagy has a key role in the pathophysiology of schizophrenia. *Mol Psychiatry* **20**, 126-32 (2015).
30. Merenlender-Wagner, A. *et al.* New horizons in schizophrenia treatment: autophagy protection is coupled with behavioral improvements in a mouse model of schizophrenia. *Autophagy* **10**, 2324-32 (2014).
31. Lotan, A. *et al.* Neuroinformatic analyses of common and distinct genetic components associated with major neuropsychiatric disorders. *Frontiers in Neuroscience* **8**(2014).
32. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *Bmc Bioinformatics* **9**(2008).
33. Kang, H.J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483-9 (2011).
34. Mao, F. *et al.* EpiDenovo: a platform for linking regulatory de novo mutations to developmental epigenetics and diseases. *Nucleic Acids Res* (2017).
35. Castello, A., Fischer, B., Hentze, M.W. & Preiss, T. RNA-binding proteins in Mendelian disease. *Trends in Genetics* **29**, 318-327 (2013).
36. Klein, M.E., Monday, H. & Jordan, B.A. Proteostasis and RNA Binding Proteins in Synaptic Plasticity and in the Pathogenesis of Neuropsychiatric Disorders. *Neural Plasticity* (2016).
37. Lukong, K.E., Chang, K.W., Khandjian, E.W. & Richard, S. RNA-binding proteins in human genetic disease. *Trends in Genetics* **24**, 416-425 (2008).
38. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nature Reviews Genetics* **15**, 829-845 (2014).
39. Neelamraju, Y., Hashemikhabir, S. & Janga, S.C. The human RBPome: From genes and proteins to human disease. *Journal of Proteomics* **127**, 61-70 (2015).
40. Taliaferro, J.M. *et al.* RNA Sequence Context Effects Measured In Vitro Predict In Vivo Protein Binding and Regulation. *Molecular Cell* **64**, 294-306 (2016).
41. Parikshak, N.N., Gandal, M.J. & Geschwind, D.H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nature Reviews Genetics* **16**, 441-458 (2015).
42. Sauna, Z.E. & Kimchi-Sarfaty, C. Understanding the contribution of synonymous mutations to human disease. *Nature Reviews Genetics* **12**, 683-691 (2011).
43. Kosmicki, J.A. *et al.* Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat Genet* (2017).

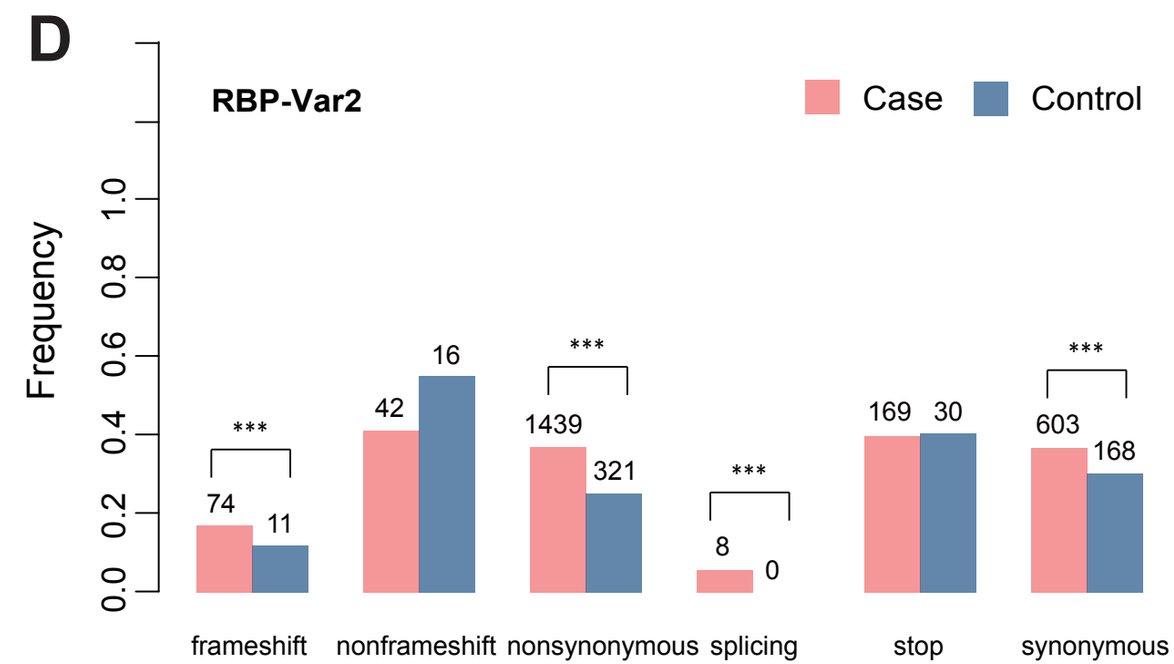
44. Sun, W. *et al.* Histone Acetylome-wide Association Study of Autism Spectrum Disorder. *Cell* **167**, 1385-1397 e11 (2016).
45. Kechavarzi, B. & Janga, S.C. Dissecting the expression landscape of RNA-binding proteins in human cancers. *Genome Biology* **15**(2014).
46. Costales, M.G. *et al.* Small Molecule Inhibition of microRNA-210 Reprograms an Oncogenic Hypoxic Circuit. *J Am Chem Soc* **139**, 3446-3455 (2017).
47. Lee, D. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **32**, 2196-2198 (2016).
48. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nature Genetics* **47**, 955-+ (2015).
49. Sabarinathan, R. *et al.* RNAsnp: Efficient Detection of Local RNA Secondary Structure Changes Induced by SNPs (vol 34, pg 546, 2013). *Human Mutation* **34**, 925-925 (2013).
50. Gruber, A.R., Lorenz, R., Bernhart, S.H., Neuboock, R. & Hofacker, I.L. The Vienna RNA Websuite. *Nucleic Acids Research* **36**, W70-W74 (2008).

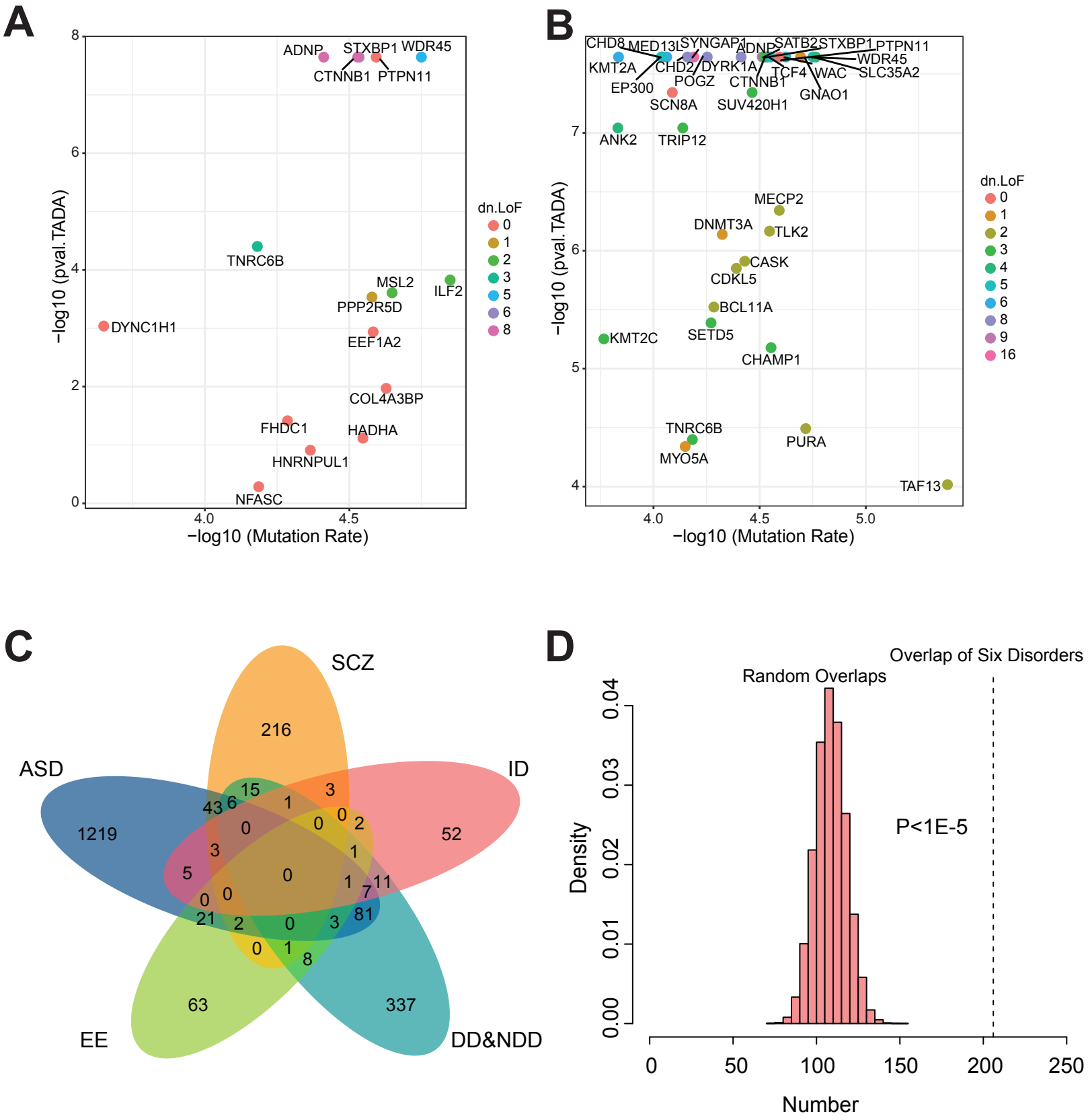

A
Case
Control
Frequency of piDNMs
0.00
0.04
0.08
***
***
*
***
***
1c
1d
1e
2b
2c
2d
RBP-Var2 Score
B
Odd Ratio of piDNMs
0.0
0.5
1.0
1.5
5.52E-02
2.81E-04
2.74E−06
5.51E−08
1.73E−17
1.07E−16
−log10(P value)
16
12
8
4
EE
ID
SCZ
DD&NDD
ASD
Total
Disorders
C
piDNMs
0
500
1000
1500
2000
84.78%
86.41%
91.81%
88.42%
91.73%
90.57%
LoF
non-LoF
ID
EE
SCZ
DD&NDD
ASD
ALL
Disorders

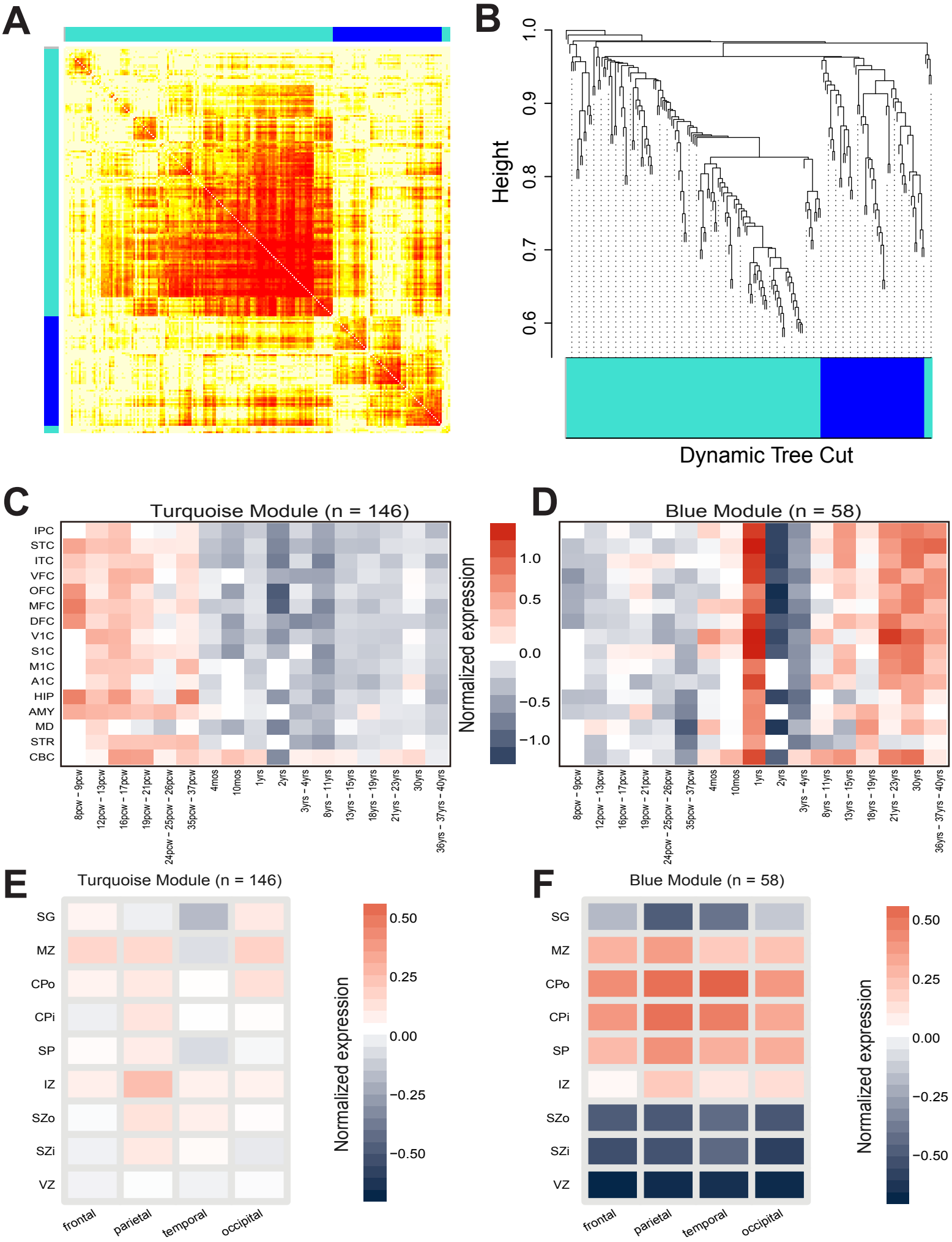

A
SIFT
Case
Control
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
***
**
1851
442
58
7
0 0
0 0
2 1
11 6
frameshift
nonframeshift
nonsynonymous
splicing
stop
synonymous


B
PPH2
Case
Control
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
***
2046
456
0 0
0 0
2 1
0 0
0 0
frameshift
nonframeshift
nonsynonymous
splicing
stop
synonymous

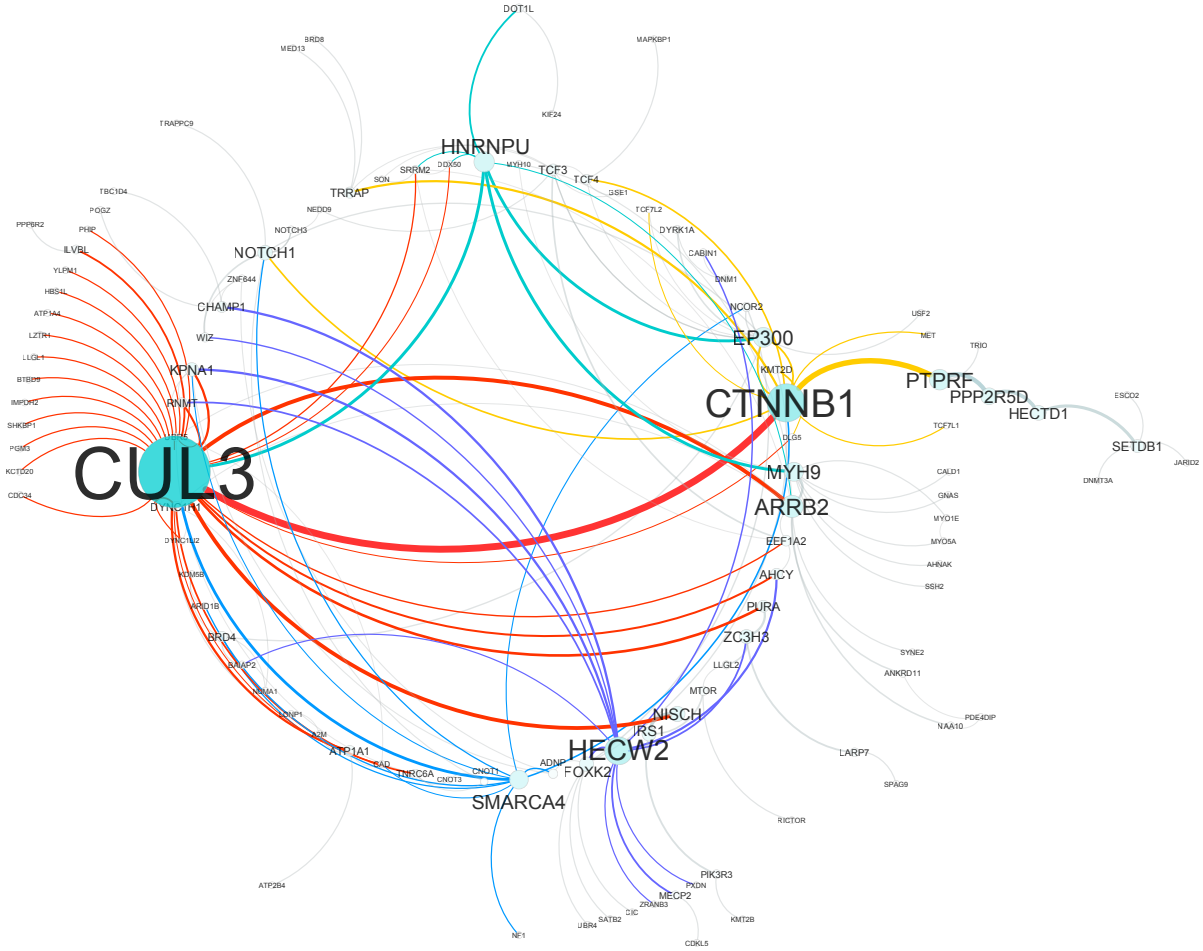

C
RegulomeDB
Case
Control
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
17 3
4 0
220 26
13 1
15 0
100 11
***
frameshift
nonframeshift
nonsynonymous
splicing
stop
synonymous


D
RBP-Var2
Case
Control
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
74 11
42 16
1439 321
8 0
169 30
603 168
***
frameshift
nonframeshift
nonsynonymous
splicing
stop
synonymous

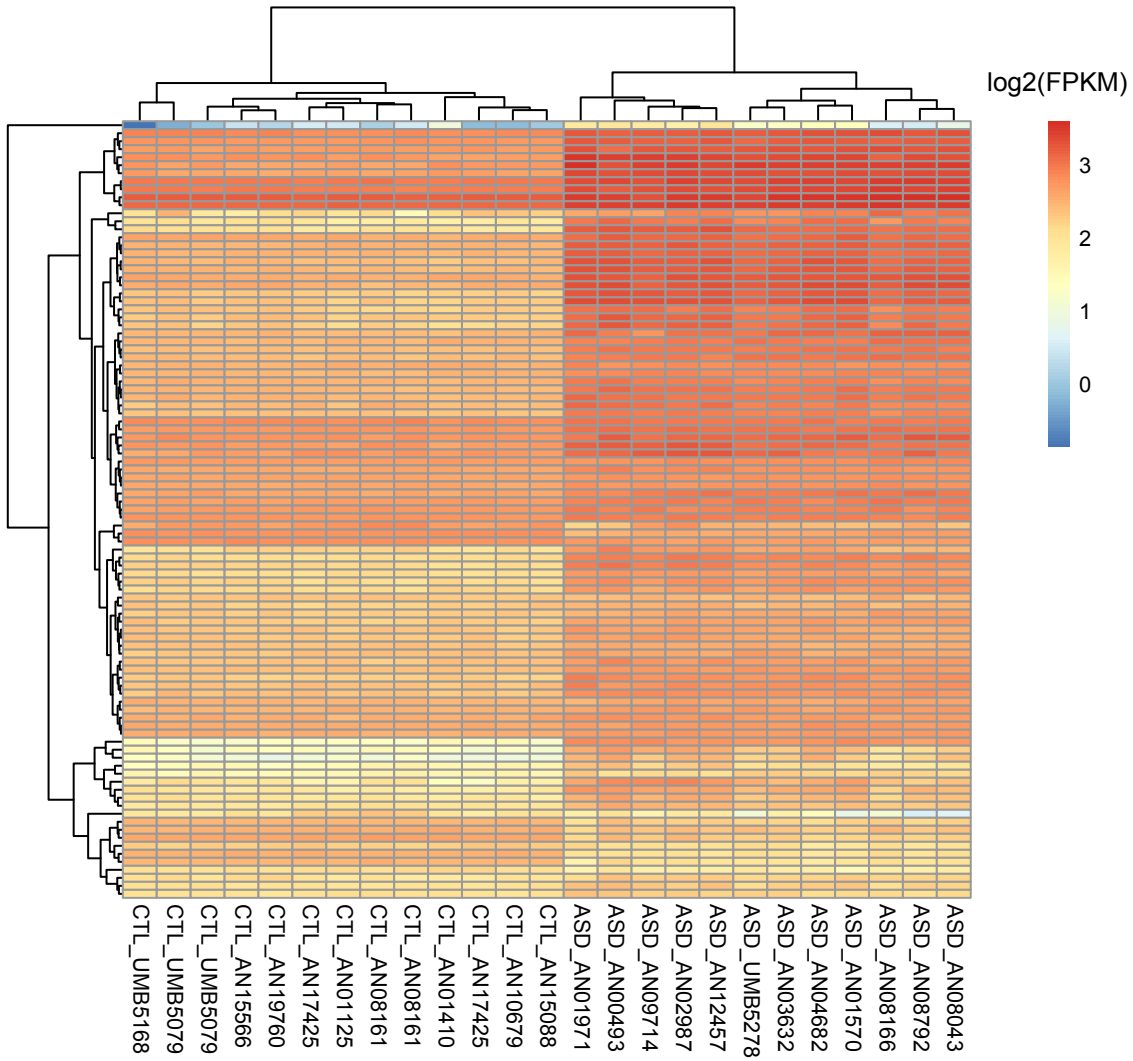

A
ADNP
STXBP1
WDR45
CTNNB1
PTPN11
TNRC6B
MSL2
ILF2
PPP2R5D
DYNC1H1
EEF1A2
COL4A3BP
FHDC1
HADHA
HNRNPUL1
NFASC
−log10 (pval.TADA)
−log10 (Mutation Rate)
dn.LoF
0
1
2
3
5
6
8
B
CHD8
MED13L
SYNGAP1
ADNP
SATB2
STXBP1
PTPN11
WDR45
KMT2A
CHD2
DYRK1A
TCF4
WAC
SLC35A2
EP300
POGZ
CTNNB1
GNAO1
SCN8A
SUV420H1
ANK2
TRIP12
MECP2
DNMT3A
TLK2
CASK
CDKL5
BCL11A
KMT2C
SETD5
CHAMP1
TNRC6B
PURA
MYO5A
TAF13
−log10 (pval.TADA)
−log10 (Mutation Rate)
dn.LoF
0
1
2
3
4
5
6
8
9
16
C
SCZ
216
ASD
1219
ID
52
EE
63
DD&NDD
337
D
Overlap of Six Disorders
Random Overlaps
P<1E-5
Density
Number

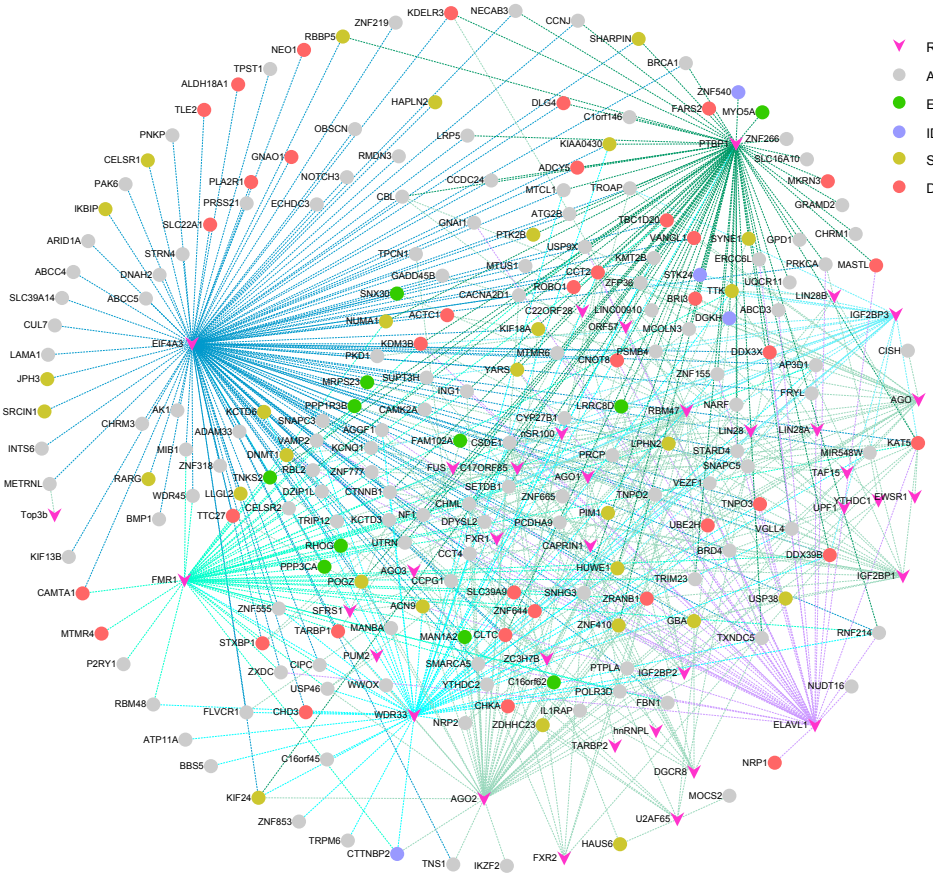

A
B
Height
1.0
0.9
0.8
0.7
0.6
Dynamic Tree Cut
C
Turquoise Module (n = 146)
IPC
STC
ITC
VFC
OFC
MFC
DFC
V1C
S1C
M1C
A1C
HIP
AMY
MD
STR
CBC
8pcw – 9pcw
12pcw – 13pcw
16pcw – 17pcw
19pcw – 21pcw
24pcw – 25pcw – 26pcw
35pcw – 37pcw
4mos
10mos
1yrs
2yrs
3yrs – 4yrs
8yrs – 11yrs
13yrs – 15yrs
18yrs – 19yrs
21yrs – 23yrs
30yrs
36yrs – 37yrs – 40yrs
Normalized expression
1.0
0.5
0.0
−0.5
−1.0
D
Blue Module (n = 58)
E
Turquoise Module (n = 146)
SG
MZ
CPo
CPi
SP
IZ
SZo
SZi
VZ
frontal
parietal
temporal
occipital
0.50
0.25
0.00
−0.25
−0.50
F
Blue Module (n = 58)

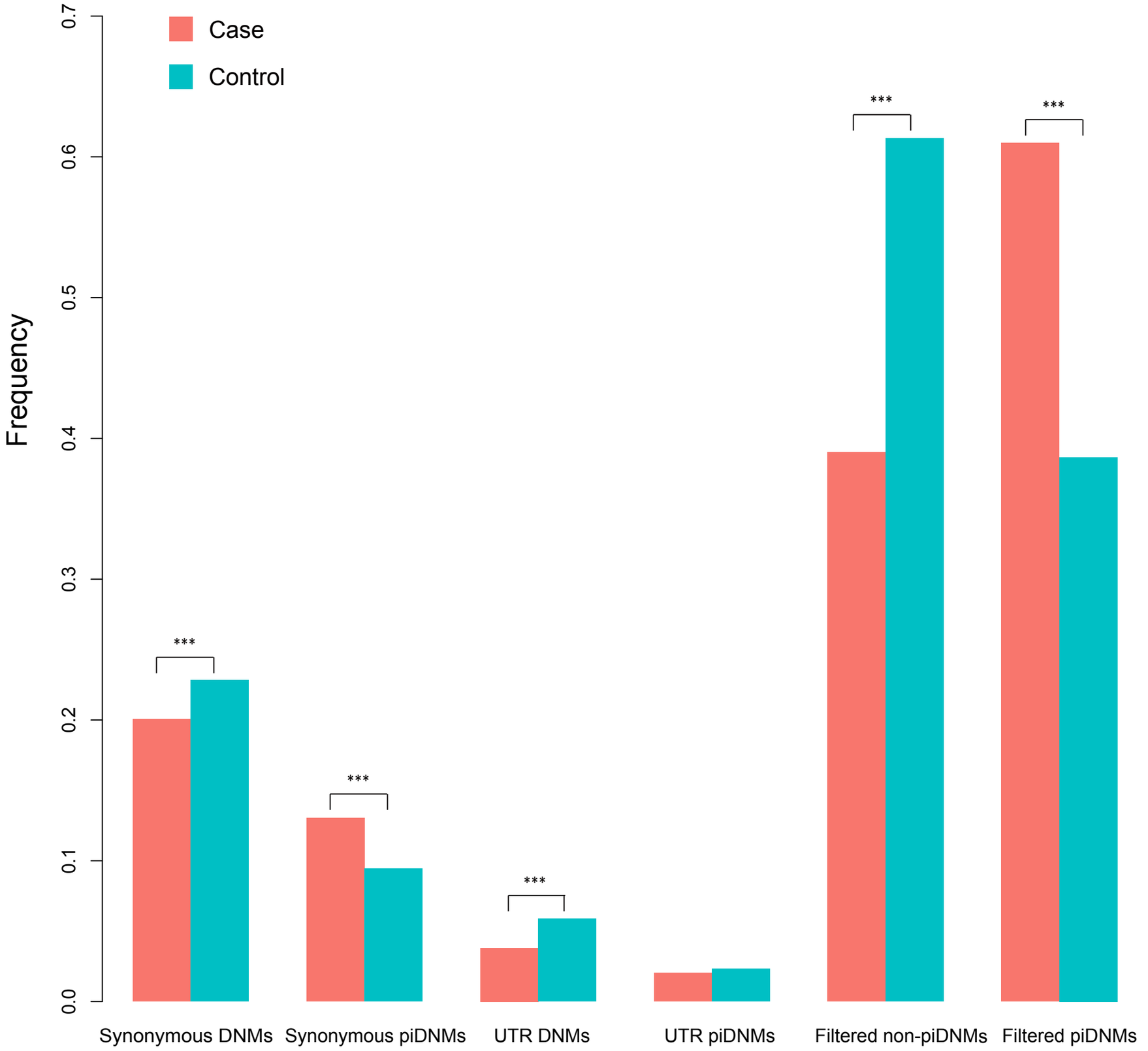
A

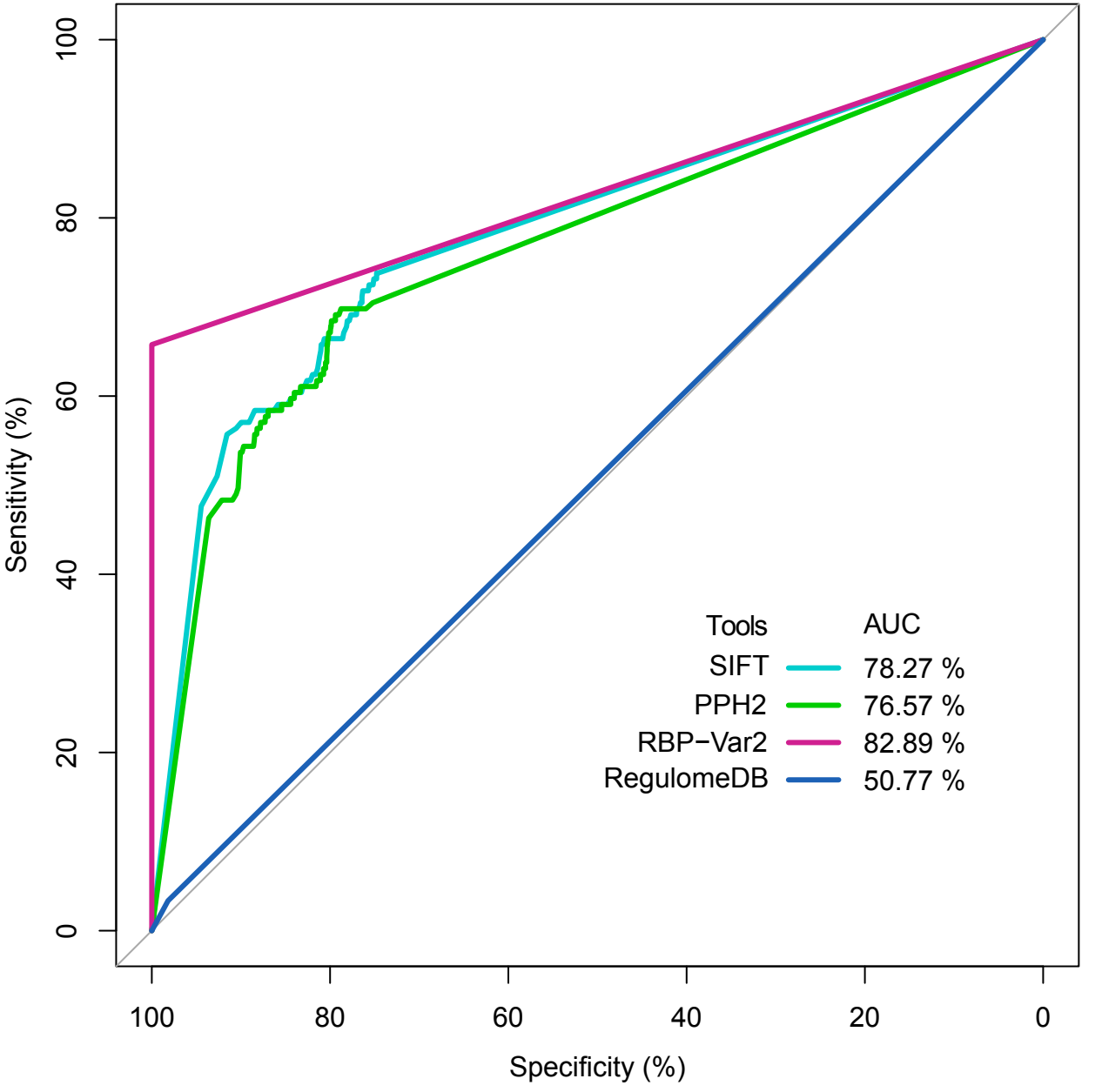

DOT1L
MAPKBP1
KIF24
HNRNPU
TCF3
TCF4
TRRAP
NOTCH1
CHAMP1
WIZ
KPNA4
RNMT
CUL3
DYNC1H1
EP300
CTNNB1
PTPRF
PPP2R5D
HECTD1
SETDB1
MYH9
ARRB2
EEF1A2
AHCY
PURA
ZC3H3
BRD4
MTOR
NISCH
IRS1
HECW2
FOXK2
SMARCA4
ATP1A1
LARP7
RICTOR
PIK3R3
MECP2
KMT2B
CDKL5

B

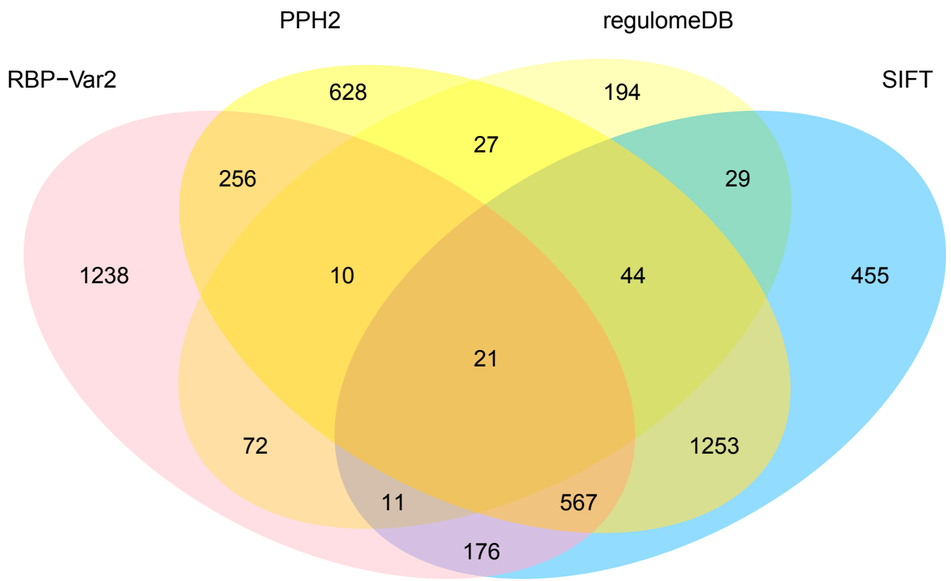

log2(FPKM)
3
2
1
0
CTL_UMB5168
CTL_UMB5079
CTL_UMB5079
CTL_AN15566
CTL_AN19760
CTL_AN17425
CTL_AN01125
CTL_AN08161
CTL_AN08161
CTL_AN01410
CTL_AN17425
CTL_AN10679
CTL_AN15088
ASD_AN01971
ASD_AN00493
ASD_AN09714
ASD_AN02987
ASD_AN12457
ASD_UMB5278
ASD_AN03632
ASD_AN04682
ASD_AN01570
ASD_AN08166
ASD_AN08792
ASD_AN08043


RBP
ASD
EE
ID
SCZ
DD & NDD
KDELR3
NECAB3
CCNJ
ZNF219
RBBP5
SHARPIN
NEO1
BRCA1
TPST1
ALDH18A1
ZNF540
HAPLN2
DLG4
FARS2
MYD5A
TLE2
C1orf146
OBSCN
LRP5
PTBP1
ZNF266
PNKP
KIAA0430
SLC16A10
CELSR1
GNAO1
RMDN3
ADCY5
MKRN3
PAK6
PLA2R1
NOTCH3
CCDC24
TROAP
GRAMD2
IKBIP
PRSS21
ECHDC3
CBL
MTCL1
ATG2B
SLC22A1
GNAI1
TBC1D20
CHRM1
ARID1A
PTK2B
VANGL1
SYNE1
GPD1
STRN4
USP9X
TPCN1
KMT2B
ERCC6L
PRKCA
MASTL
ABCC4
DNAH2
MTUS1
CCT2
STK24
GADD45B
ZFP36
UQCR11
SLC39A14
ABCC5
SNX30
CACNA2D1
ROBO1
TTK
LIN28B
BRI3
ABCD3
CUL7
NUMA1
ACTC1
C22ORF28
LINC00910
DGKH
IGF2BP3
KIF18A
ORF57
MCOLN3
EIF4A3
KDM3B
MTMR6
PSMB4
DDX3X
CISH
LAMA1
PKD1
CNOT8
AP3D1
JPH3
YARS
ZNF155
MRPS23
SUPT3H
ING1
FRYL
SRCIN1
PPP1R3B
CAMK2A
LRRC8D
RBM47
NARF
AGO
AK1
KCTD6
SNAPC3
CYP27B1
CHRM3
ADAM33
AGGF1
rSR100
LIN28
LIN28A
INTS6
VAMP2
FAM102A
CSDE1
LPHN2
KAT5
MIB1
DNMT1
KCNQ1
PRCP
STARD4
MIR548W
ZNF318
RBL2
ZNF777
FUS
C17ORF85
SNAPC5
METRNL
RARG
TNKS2
AGO1
TAF15
WDR45
LLGL2
DZIP1L
CTNNB1
SETDB1
VEZF1
TNPO2
EWSR1
Top3b
CELSR2
CHML
ZNF665
TNPO3
UPF1
YTHDC1
BMP1
TTC27
TRIP12
KCTD3
NF1
DPYSL2
PCDHA9
PIM1
UBE2H
VGLL4
KIF13B
RHOG
UTRN
FXR1
CAPRIN1
BRD4
DDX39B
PPP3CA
CCT4
HUWE1
IGF2BP1
FMR1
AGO3
TRIM23
CAMTA1
POGZ
CCPG1
SLC39A9
SNHG3
ZRANB1
USP38
ZNF555
SFRS1
ACN9
ZNF644
GBA
MTMR4
TARBP1
MANBA
MAN1A2
CLTC
ZNF410
TXNDC5
RNF214
STXBP1
PUM2
ZC3H7B
P2RY1
CIPC
SMARCA5
PTPLA
IGF2BP2
ZXDC
WWOX
YTHDC2
C16orf62
NUDT16
USP46
POLR3D
RBM48
CHKA
FBN1
FLVCR1
CHD3
WDR33
IL1RAP
ELAVL1
NRP2
ZDHHC23
hnRNPL
ATP11A
TARBP2
C16orf45
NRP1
BBS5
DGCR8
KIF24
AGO2
MOCS2
ZNF853
U2AF65
TRPM6
HAUS6
CTTNBP2
FXR2
TNS1
IKZF2

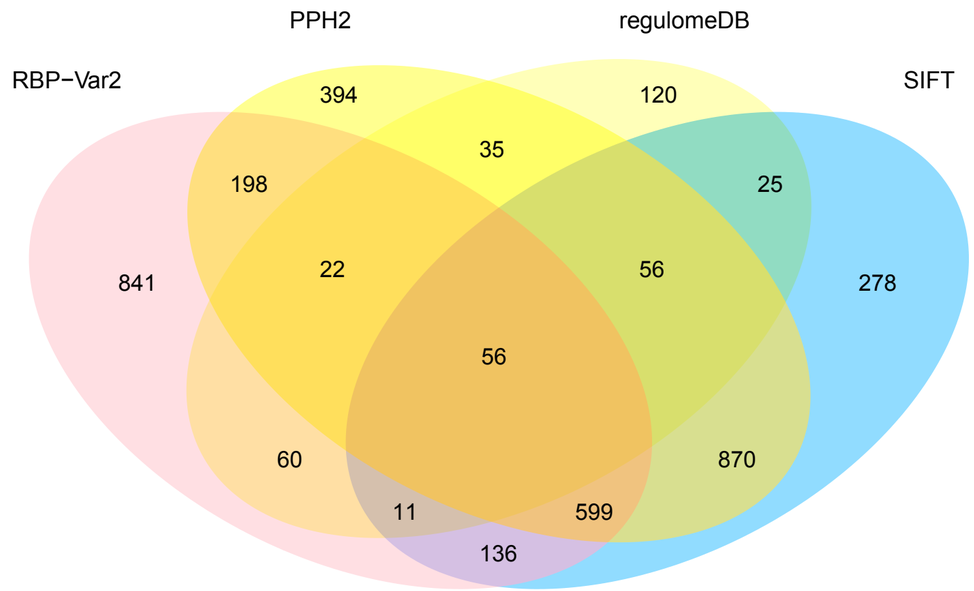

Case
Control
Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
***
***
***
***
***
Synonymous DNMs
Synonymous piDNMs
UTR DNMs
UTR piDNMs
Filtered non-piDNMs
Filtered piDNMs


Sensitivity (%)
Specificity (%)
100
80
60
40
20
0
Tools
AUC
SIFT
78.27 %
PPH2
76.57 %
RBP−Var2
82.89 %
RegulomeDB
50.77 %

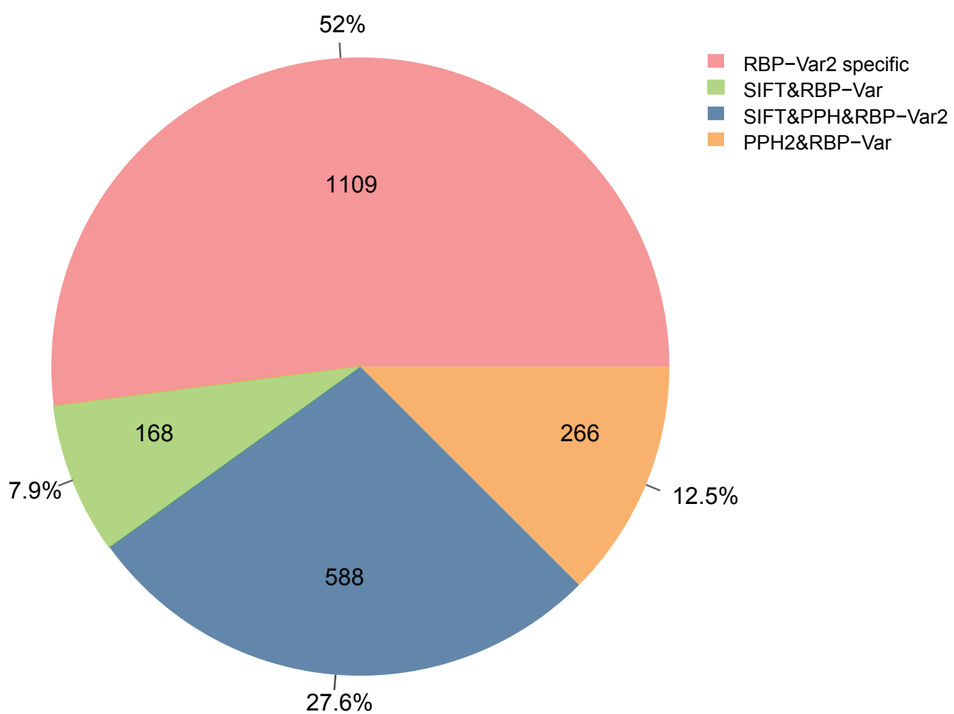
A


PPH2
regulomeDB
RBP−Var2
SIFT
628
194
27
256
29
1238
10
44
455
21
72
1253
11
567
176

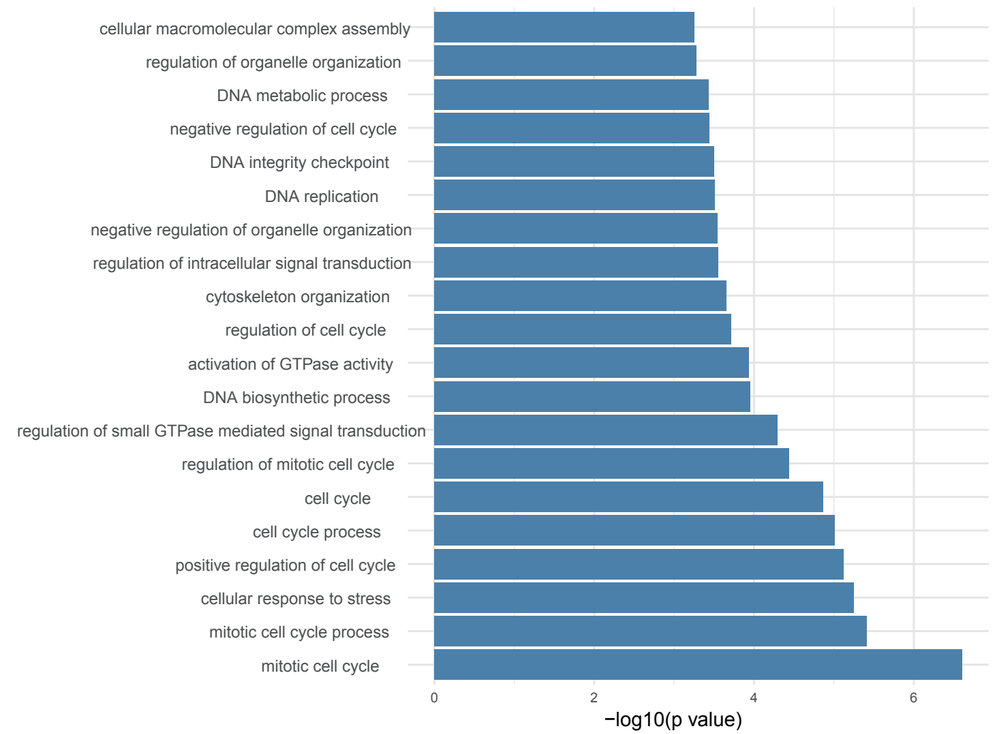
B

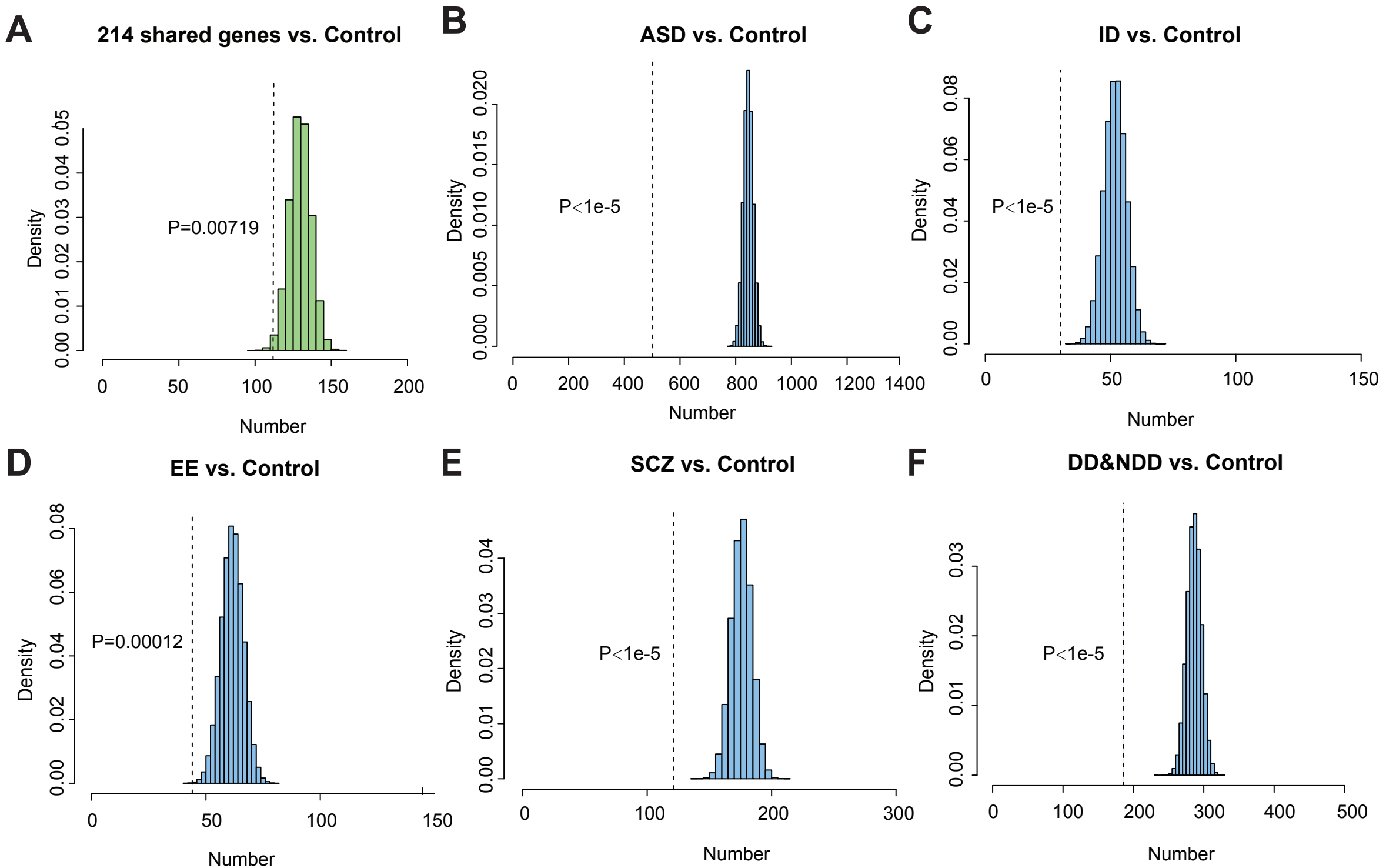

PPH2
regulomeDB
RBP−Var2
SIFT
394
120
35
198
25
841
22
56
278
56
60
870
11
599
136

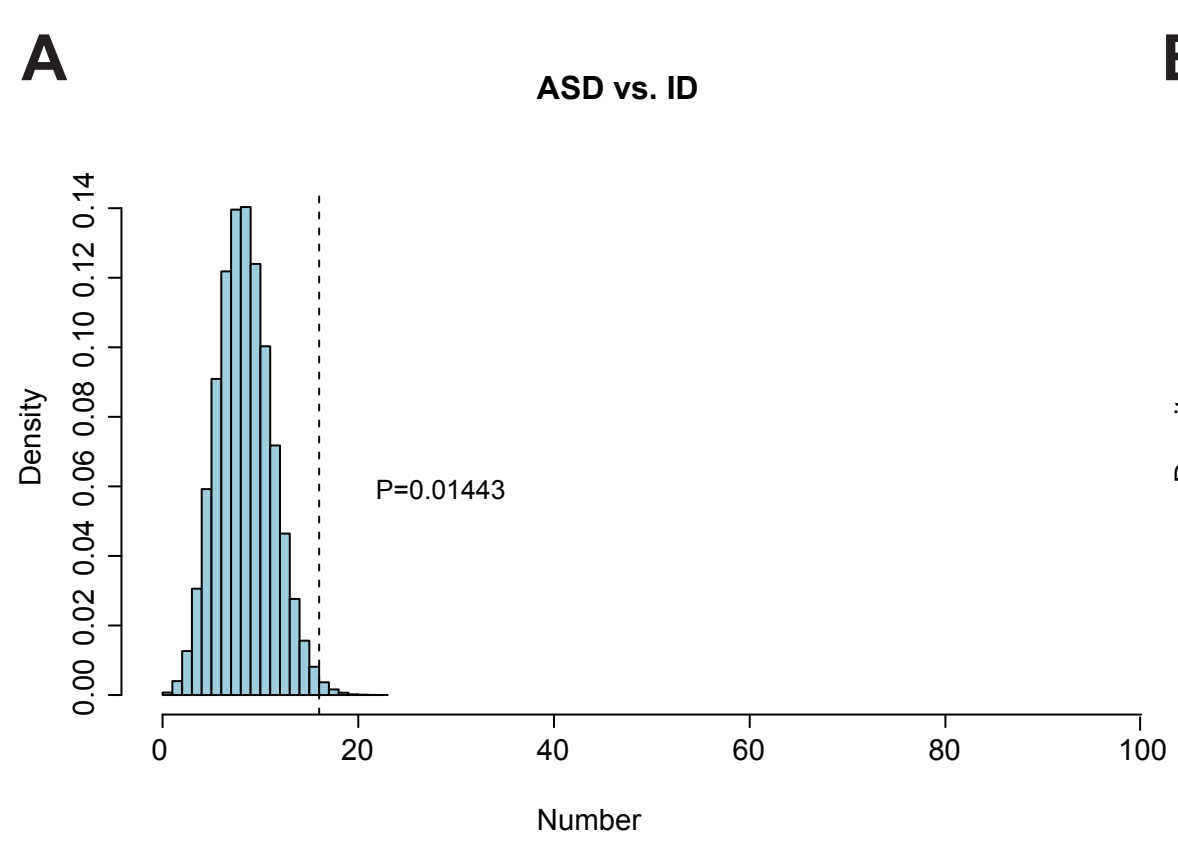
C

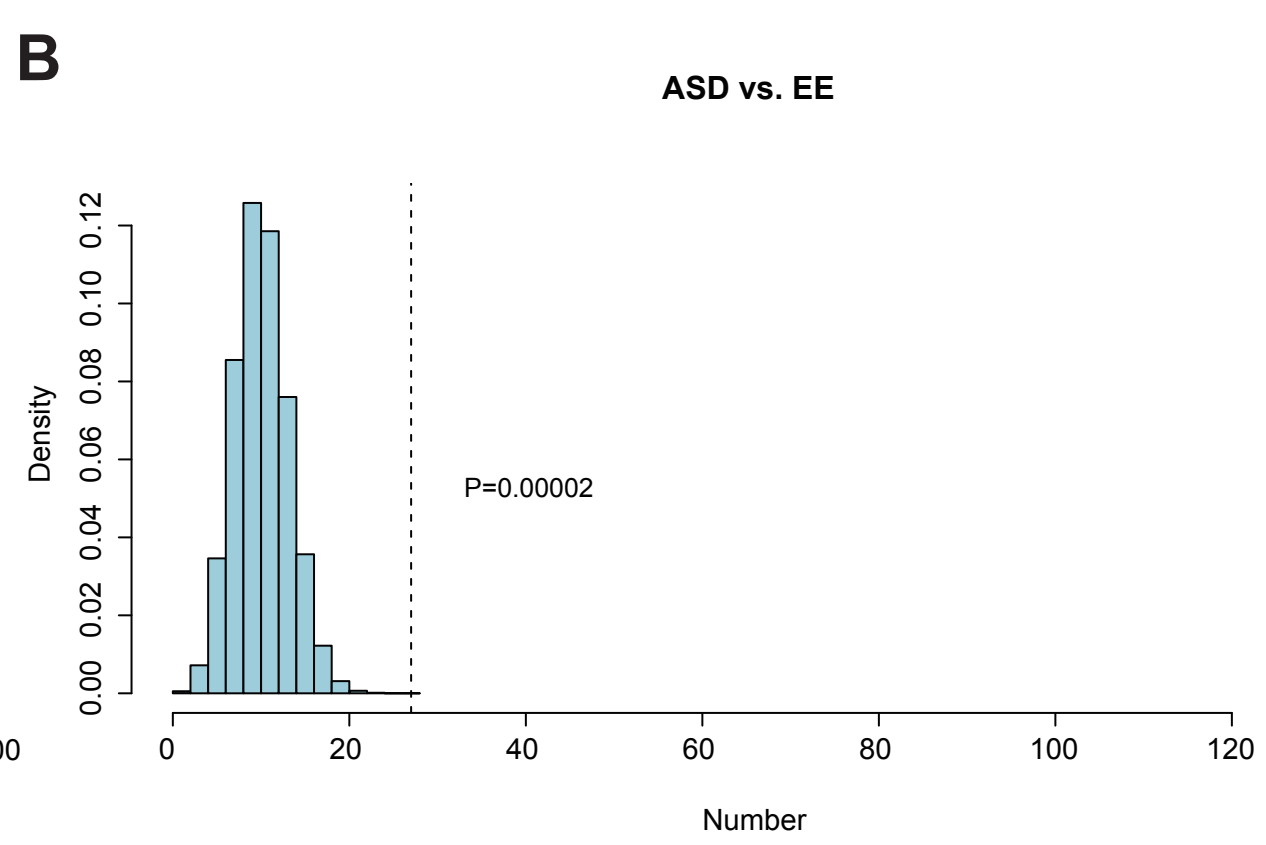

52%
RBP−Var2 specific
SIFT&RBP−Var
SIFT&PPH&RBP−Var2
PPH2&RBP−Var
1109
168
266
7.9%
12.5%
588
27.6%

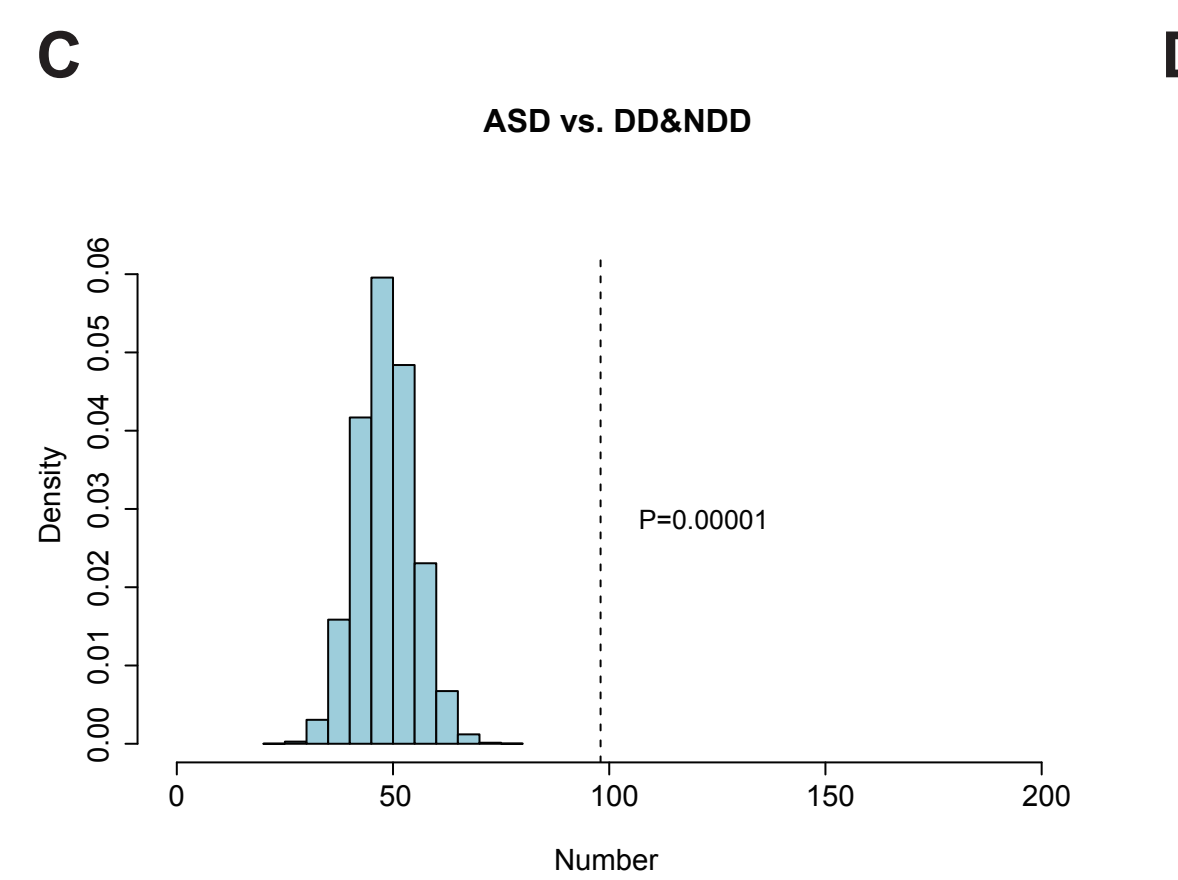
D

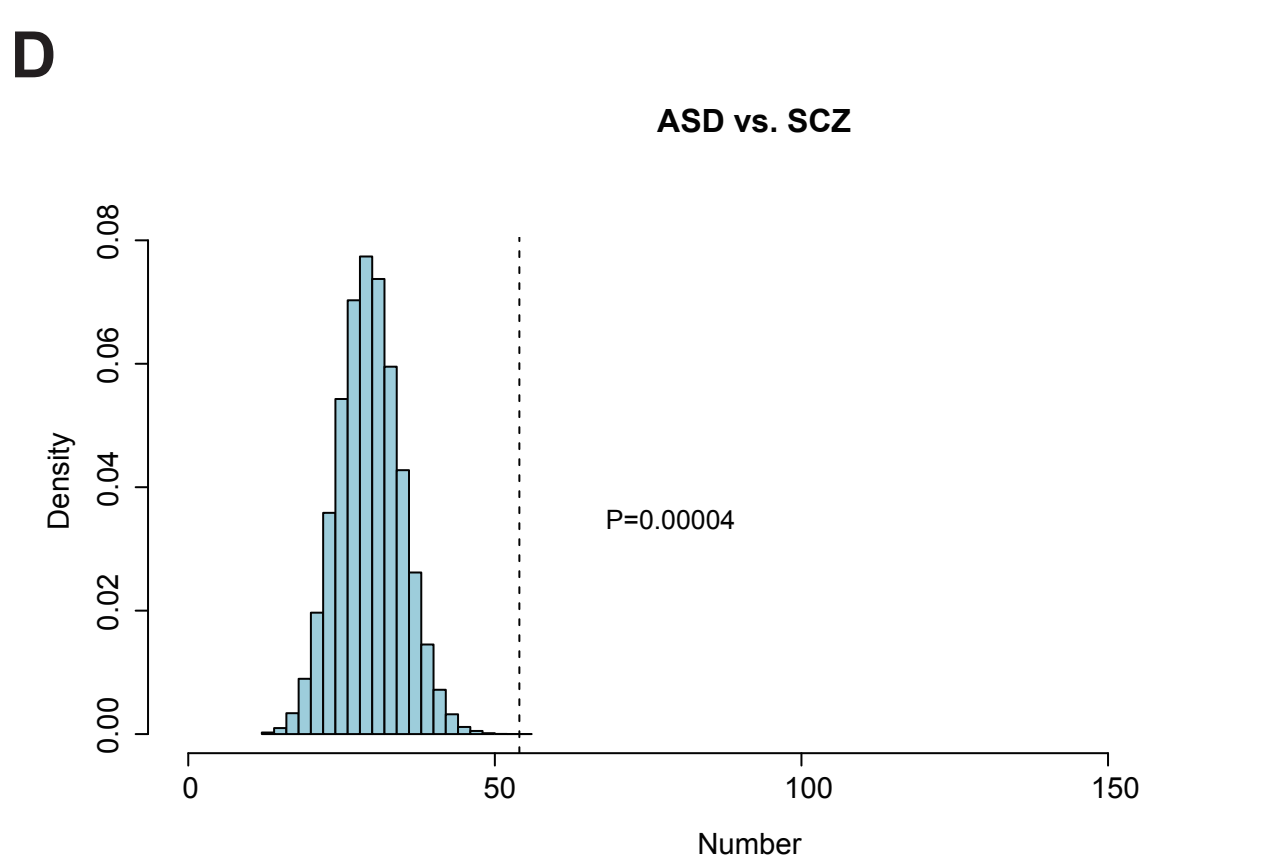

cellular macromolecular complex assembly
regulation of organelle organization
DNA metabolic process
negative regulation of cell cycle
DNA integrity checkpoint
DNA replication
negative regulation of organelle organization
regulation of intracellular signal transduction
cytoskeleton organization
regulation of cell cycle
activation of GTPase activity
DNA biosynthetic process
regulation of small GTPase mediated signal transduction
regulation of mitotic cell cycle
cell cycle
cell cycle process
positive regulation of cell cycle
cellular response to stress
mitotic cell cycle process
mitotic cell cycle
0
2
4
6
−log10(p value)

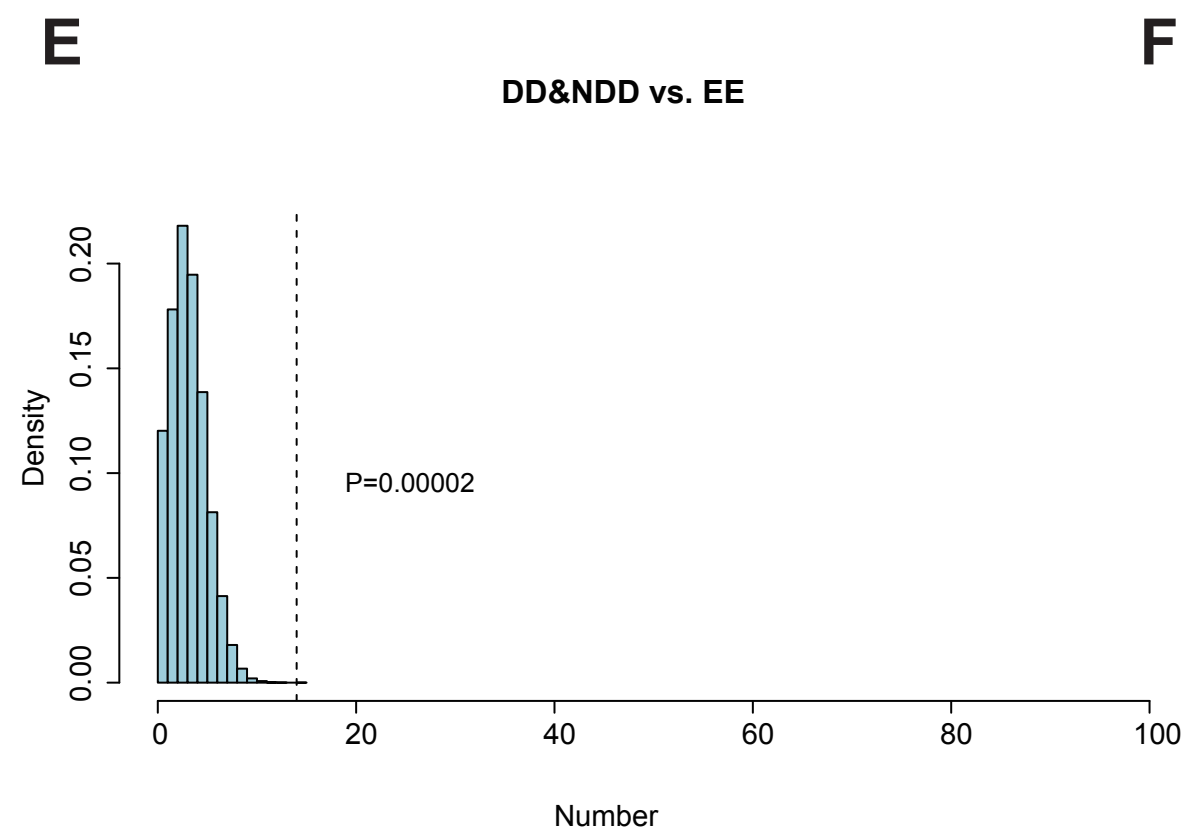

A
214 shared genes vs. Control
P=0.00719
Density
Number
B
ASD vs. Control
P<1e-5
Density
Number
C
ID vs. Control
P<1e-5
Density
Number
D
EE vs. Control
P=0.00012
Density
Number
E
SCZ vs. Control
P<1e-5
Density
Number
F
DD&NDD vs. Control
P<1e-5
Density
Number

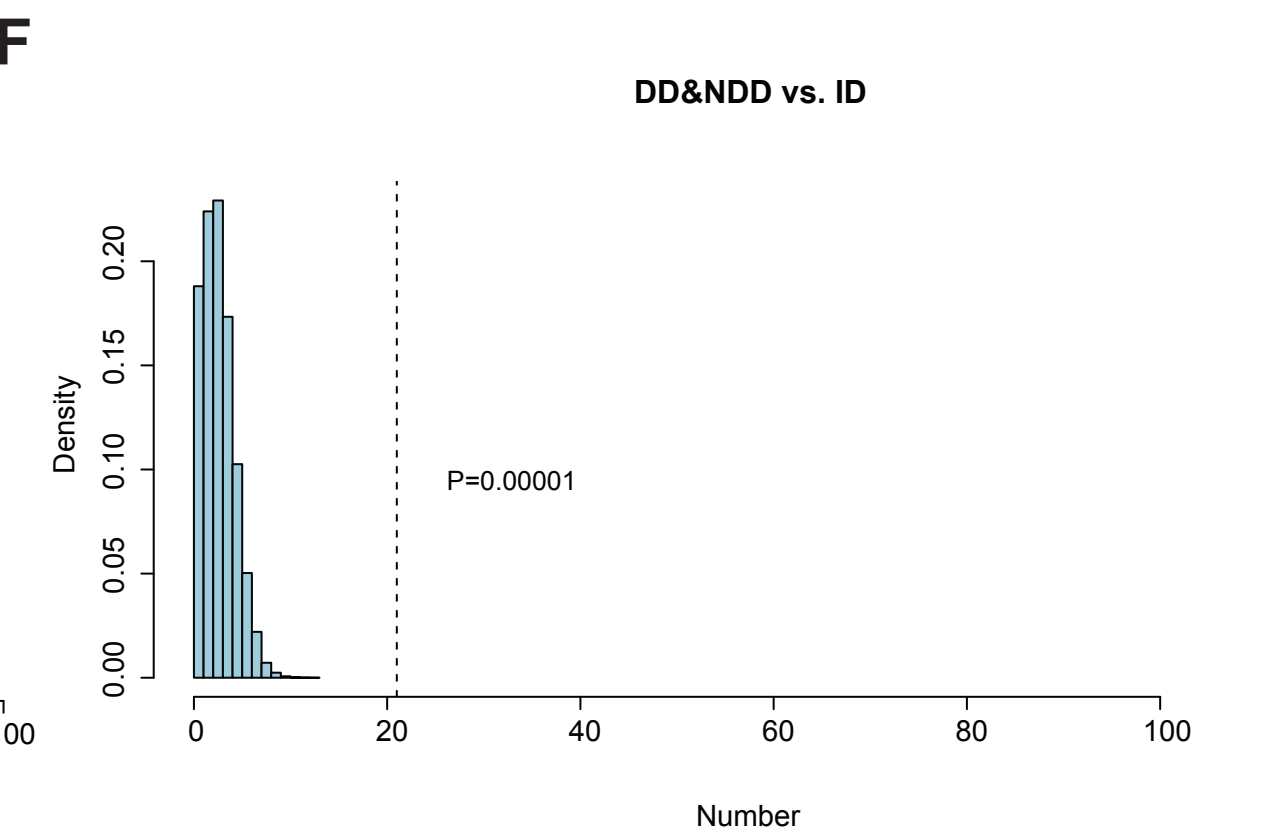

A
ASD vs. ID
P=0.01443
Density
Number

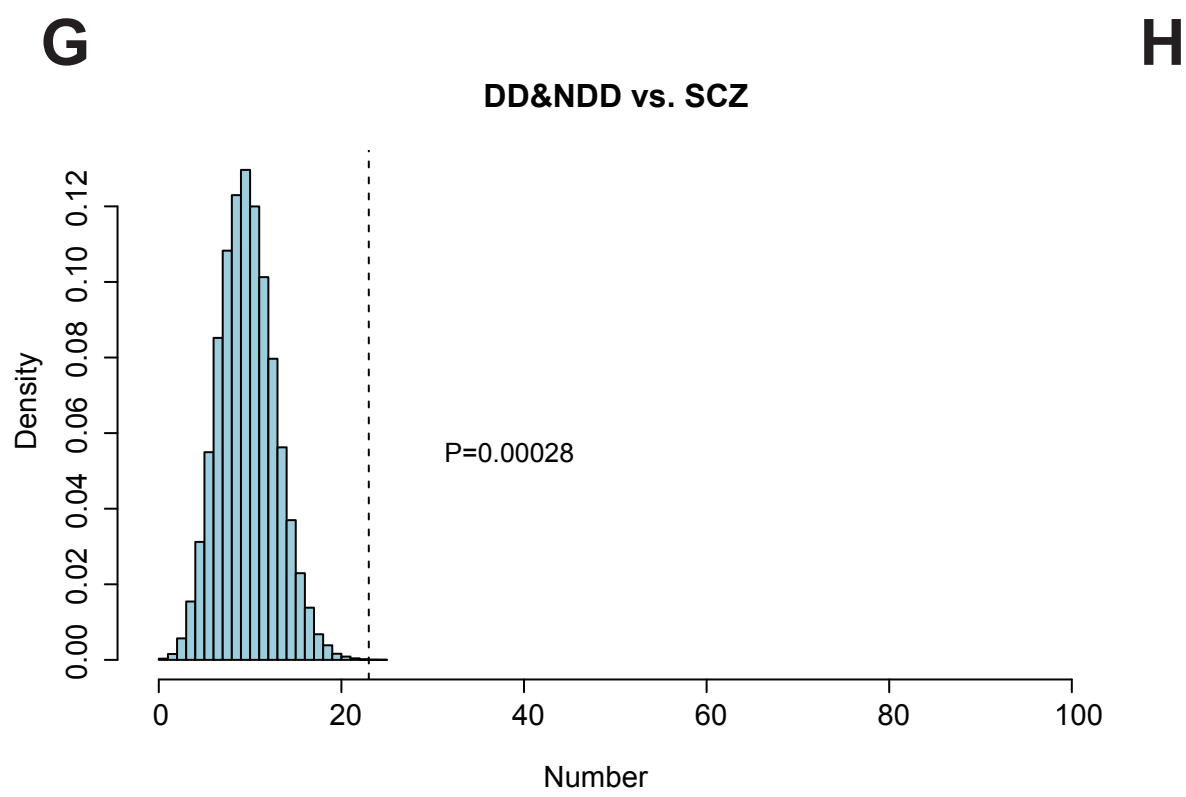

B
ASD vs. EE
P=0.00002
Density
Number

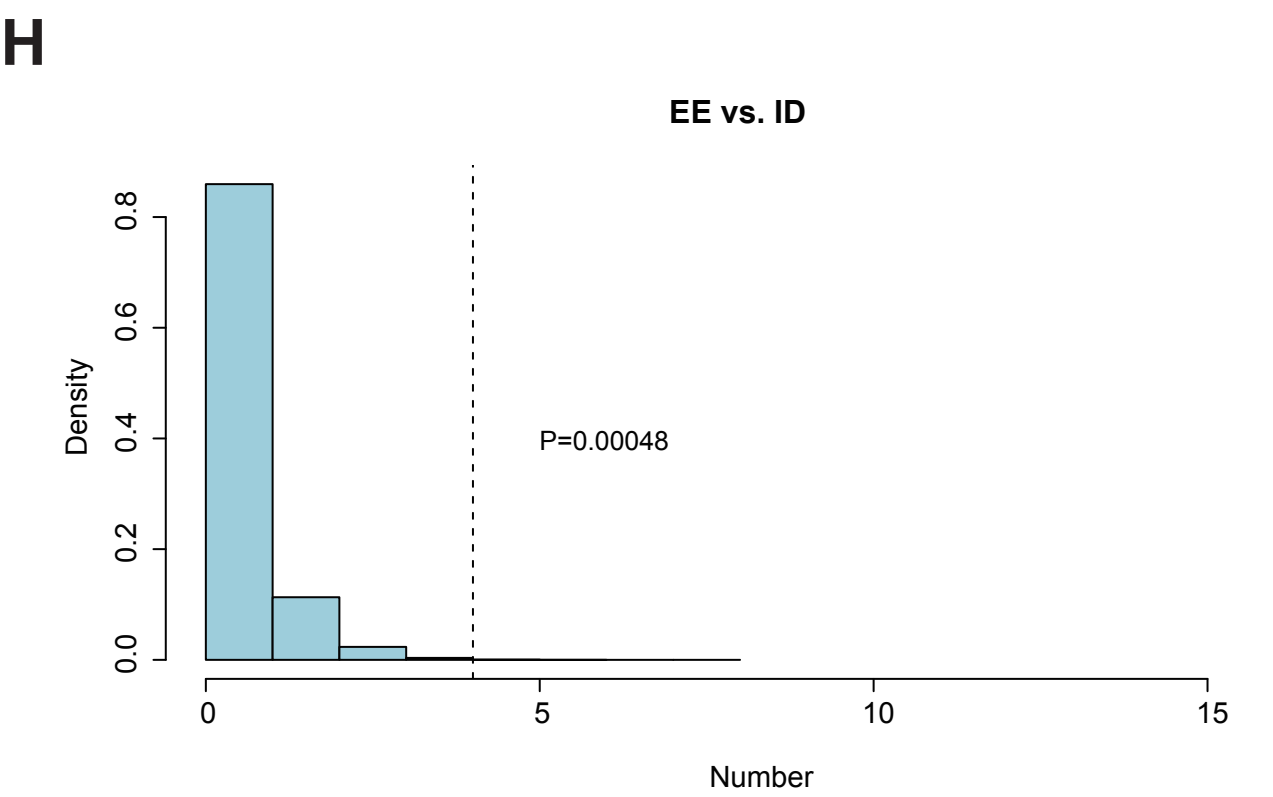

C
ASD vs. DD&NDD
P=0.00001
Density
Number

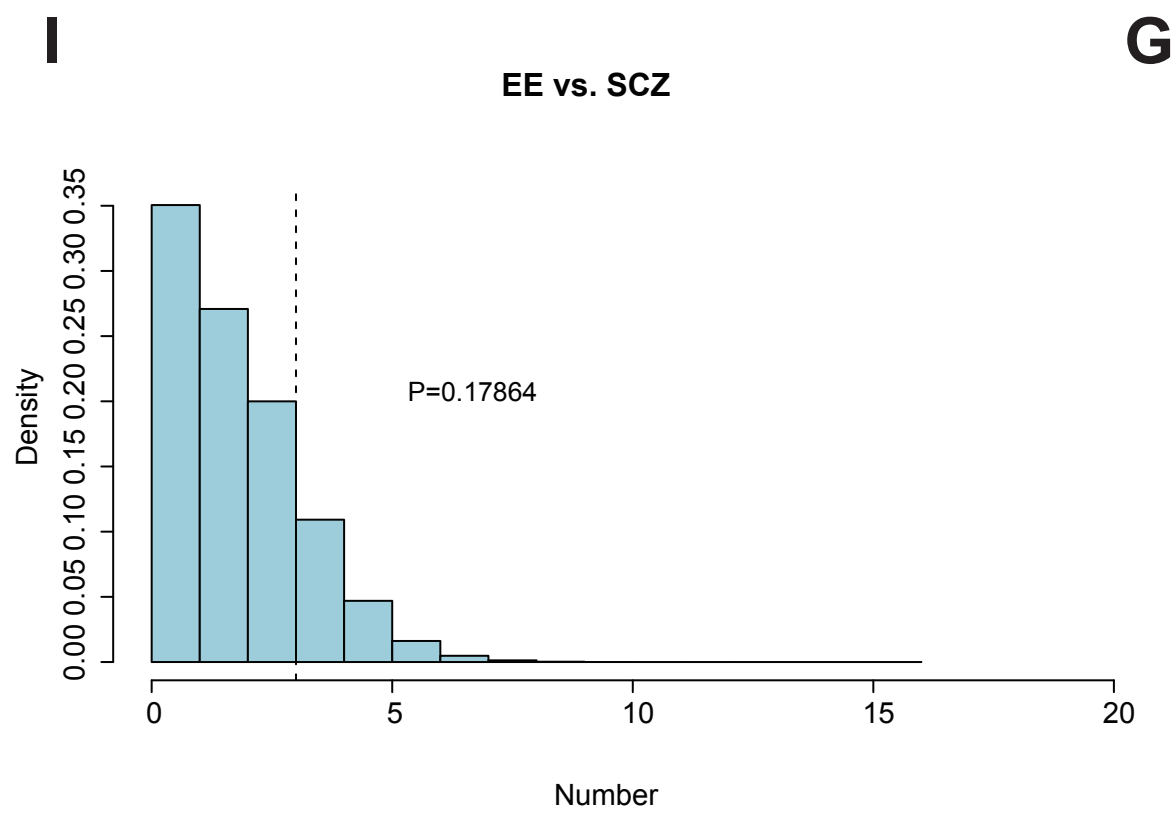

D
ASD vs. SCZ
P=0.00004
Density
Number

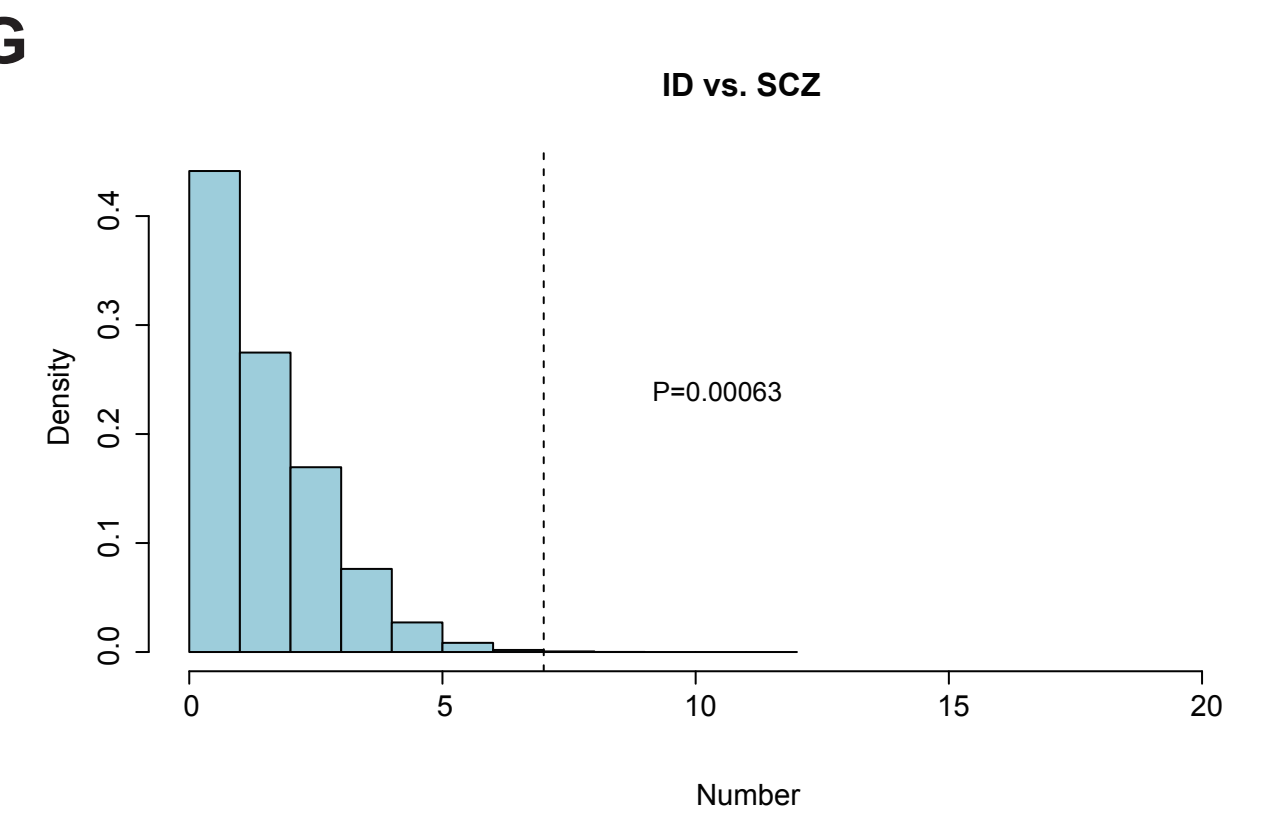

E
DD&NDD vs. EE
P=0.00002
Density
Number

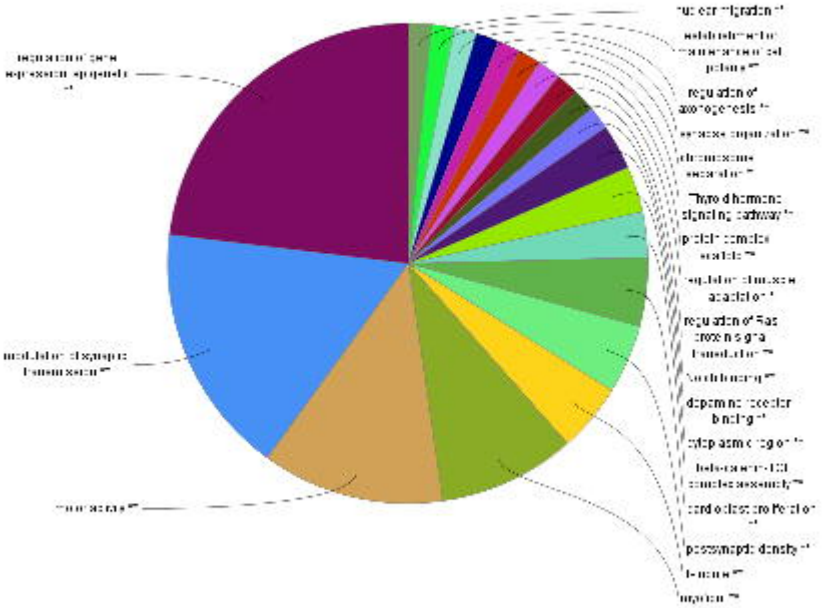

F
DD&NDD vs. ID
P=0.00001
Density
Number

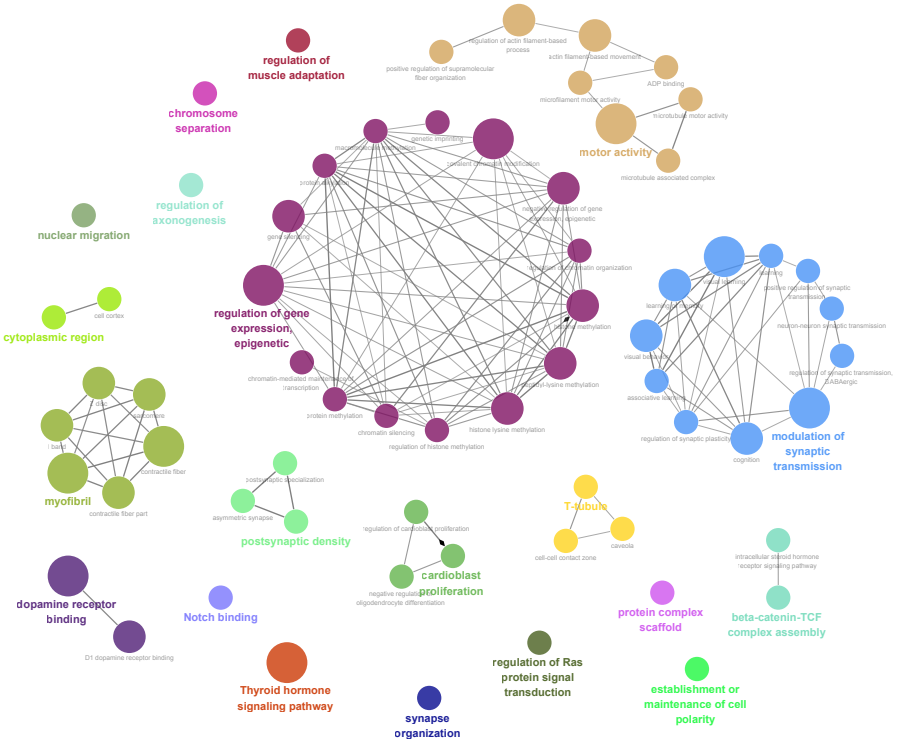

G
DD&NDD vs. SCZ
P=0.00028
Density
Number

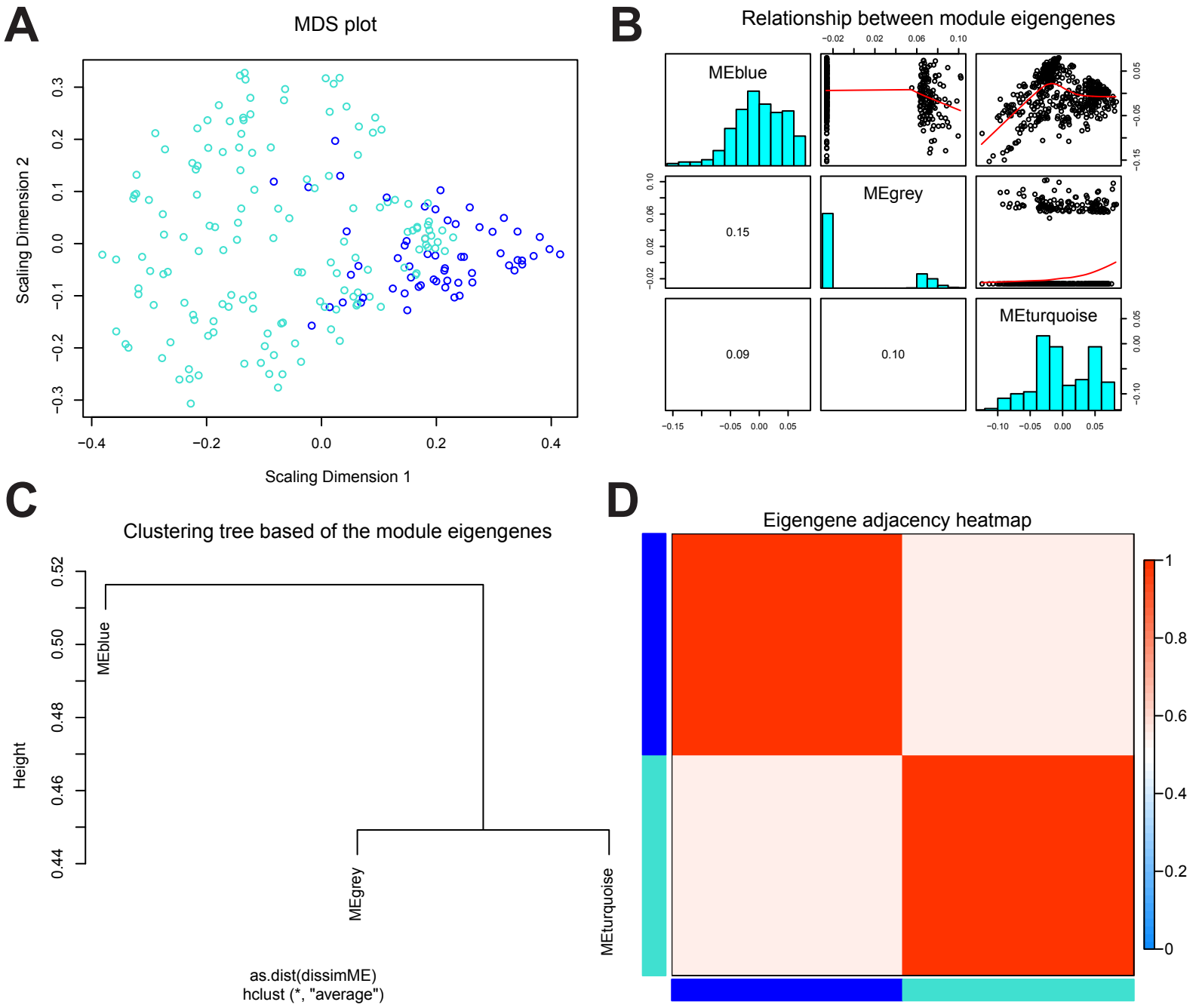

H
EE vs. ID
P=0.00048
Density
Number

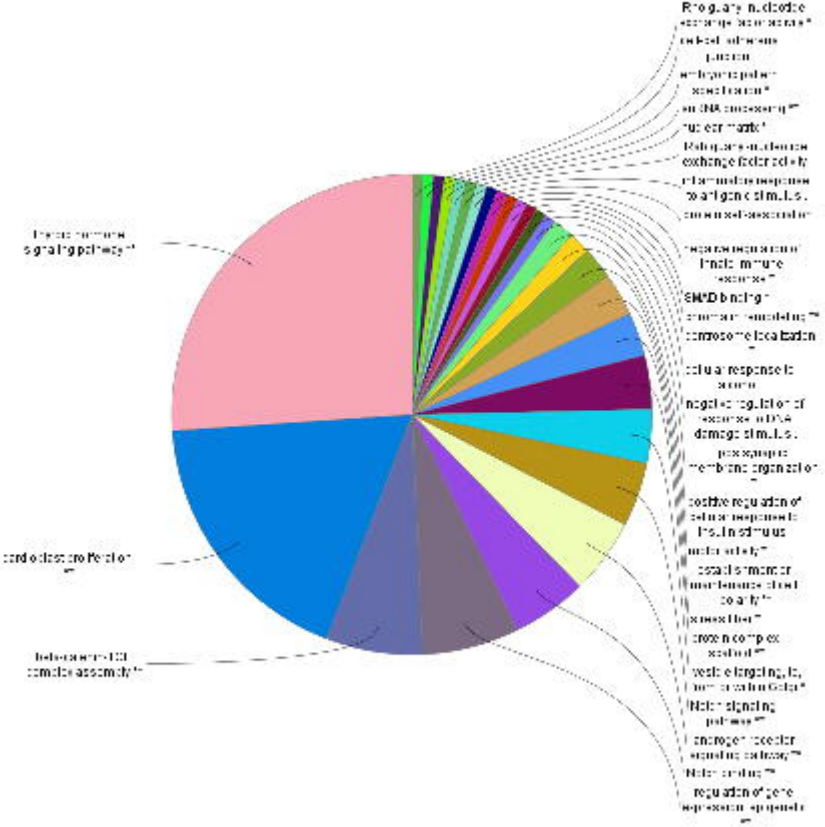

I
EE vs. SCZ
P=0.17864
Density
Number

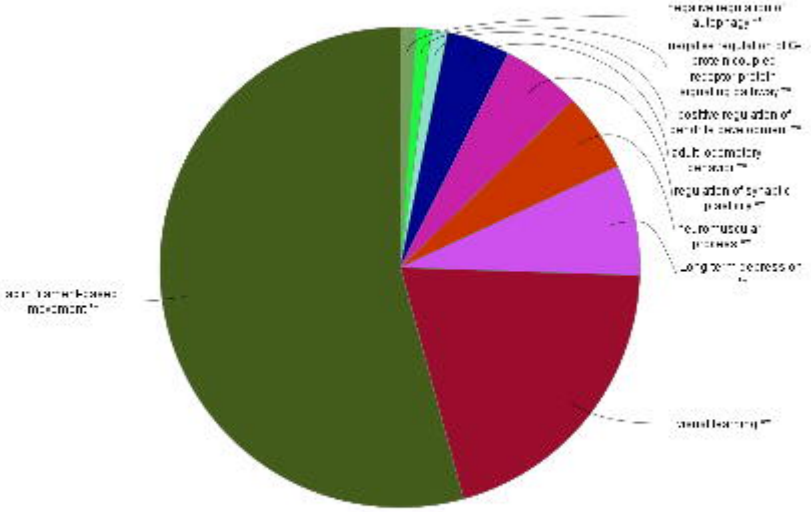

G
ID vs. SCZ
P=0.00063
Density
Number

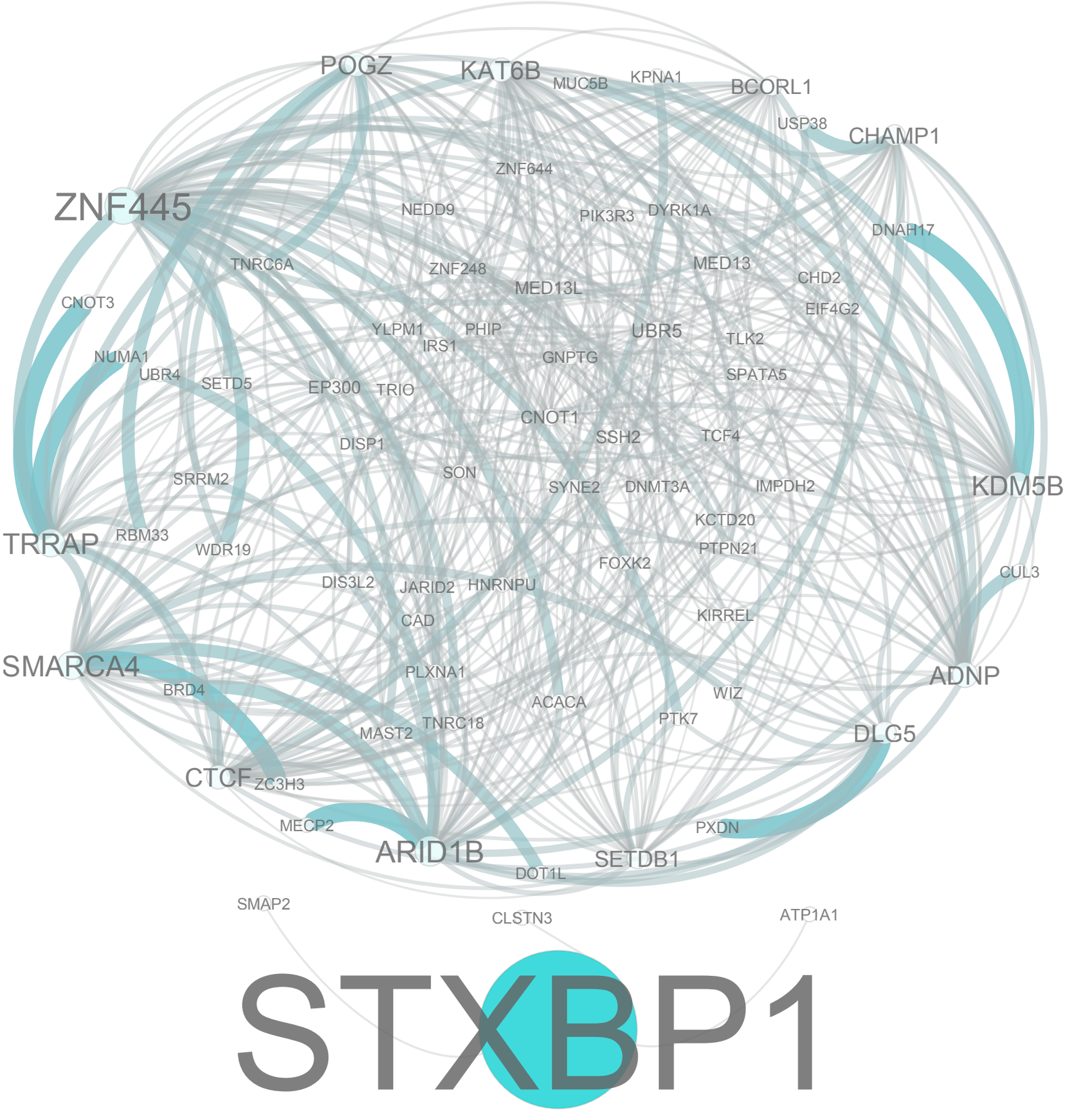

regulation of axonogenesis
Thyroid hormone signaling pathway
regulation of Ras protein signal transduction
dopamine receptor binding
cytoplasmic region
cardioblast proliferation
postsynaptic density

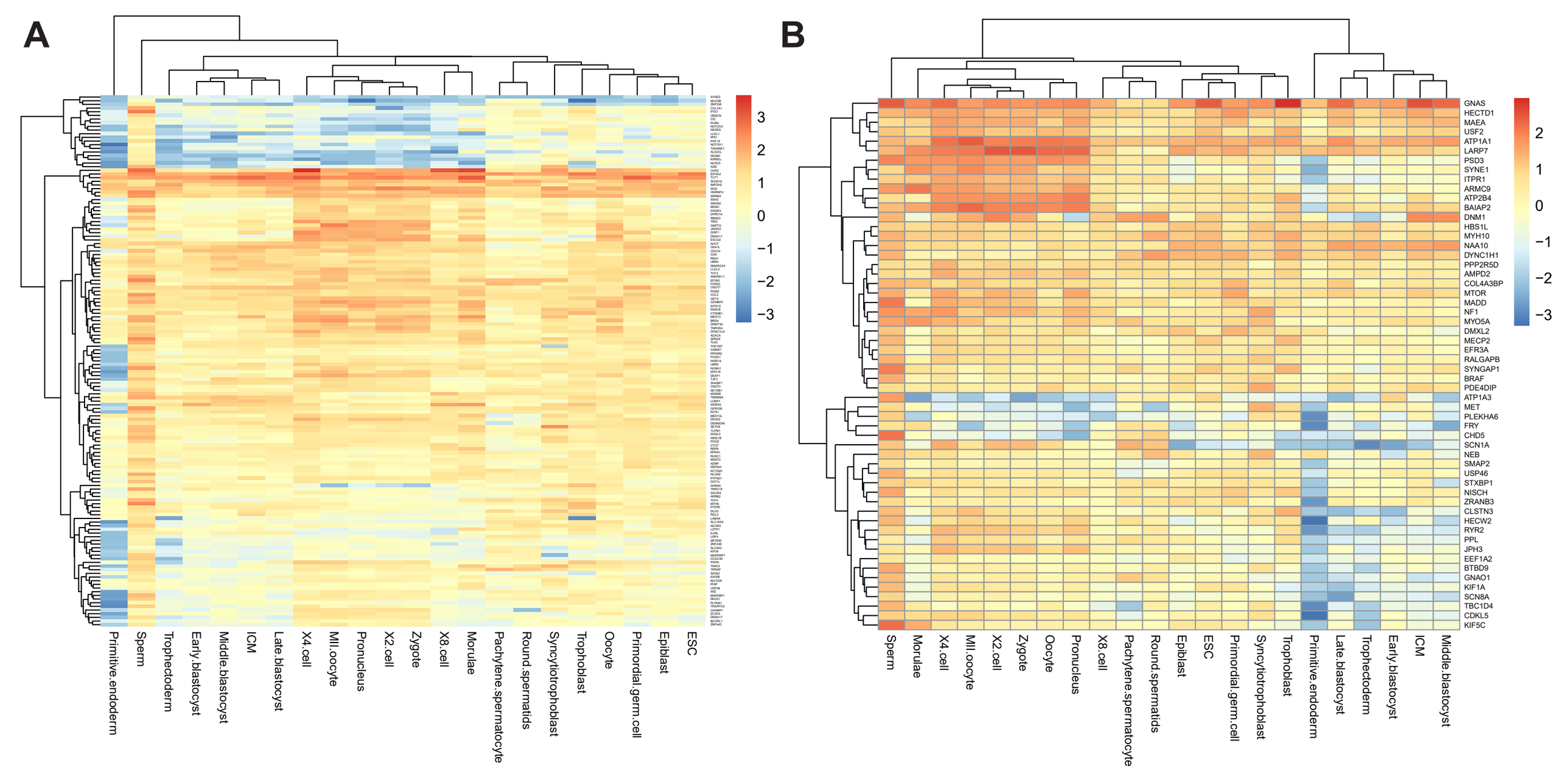

regulation of muscle adaptation
chromosome separation
regulation of axonogenesis
nuclear migration
cytoplasmic region
cell cortex
regulation of gene expression, epigenetic
chromatin-mediated maintenance of transcription
protein methylation
chromatin silencing
regulation of histone methylation
histone lysine methylation
protein-lysine methylation
genetic imprinting
macromolecular methylation
covalent chromatin modification
protein methylation
gene silencing
regulation of gene expression, epigenetic
regulation of actin filament-based process
positive regulation of supramolecular fiber organization
actin filament-based movement
ADP binding
microfilament motor activity
microtubule motor activity
motor activity
microtubule associated complex
visual learning
learning
positive regulation of synaptic transmission
learning or memory
neuron-neuron synaptic transmission
visual behavior
regulation of synaptic transmission, GABAergic
associative learning
regulation of synaptic plasticity
cognition
modulation of synaptic transmission
i disc
a/i contacts
i band
contractile fiber
myofibril
contractile fiber part
postsynaptic specialization
asymmetric synapse
postsynaptic density
regulation of cardioblast proliferation
cardioblast proliferation
negative regulation of oligodendrocyte differentiation
T-tubule
caveola
cell-cell contact zone
intracellular steroid hormone receptor signaling pathway
beta-catenin-TCF complex assembly
dopamine receptor binding
D1 dopamine receptor binding
Notch binding
protein complex scaffold
regulation of Ras protein signal transduction
establishment or maintenance of cell polarity
Thyroid hormone signaling pathway
synapse organization

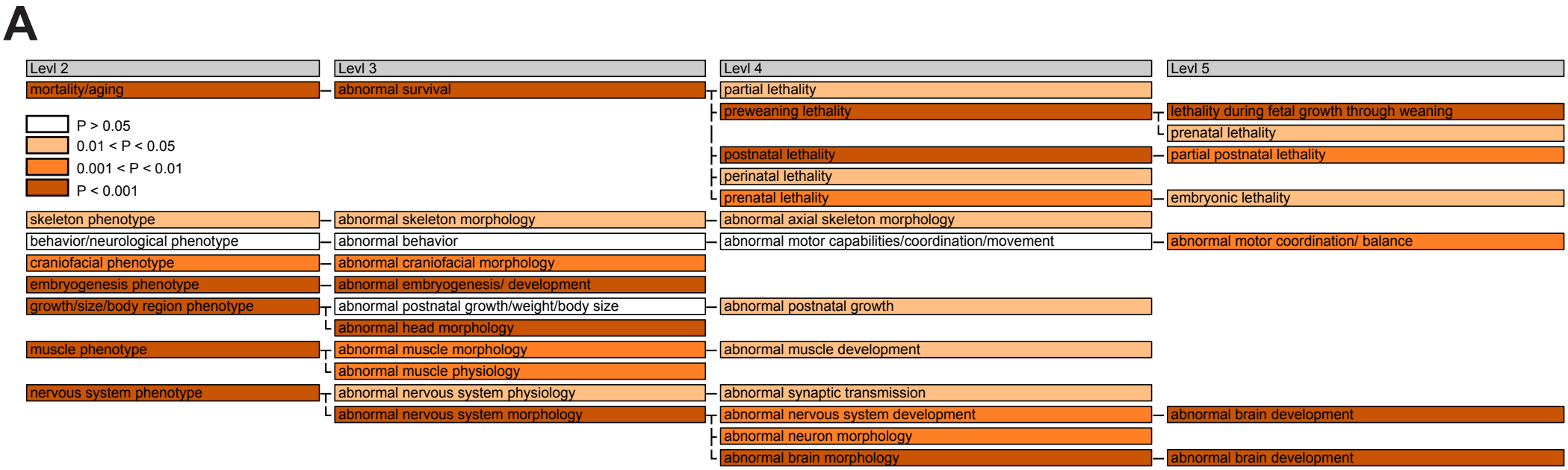

A
MDS plot
Scaling Dimension 2
Scaling Dimension 1
B
Relationship between module eigengenes
MEblue
MEgrey
MEturquoise
0.15
0.09
0.10
C
Clustering tree based of the module eigengenes
Height
MEblue
MEgrey
MEturquoise
as.dist(dissimME)
hclust (*, "average")
D
Eigengene adjacency heatmap

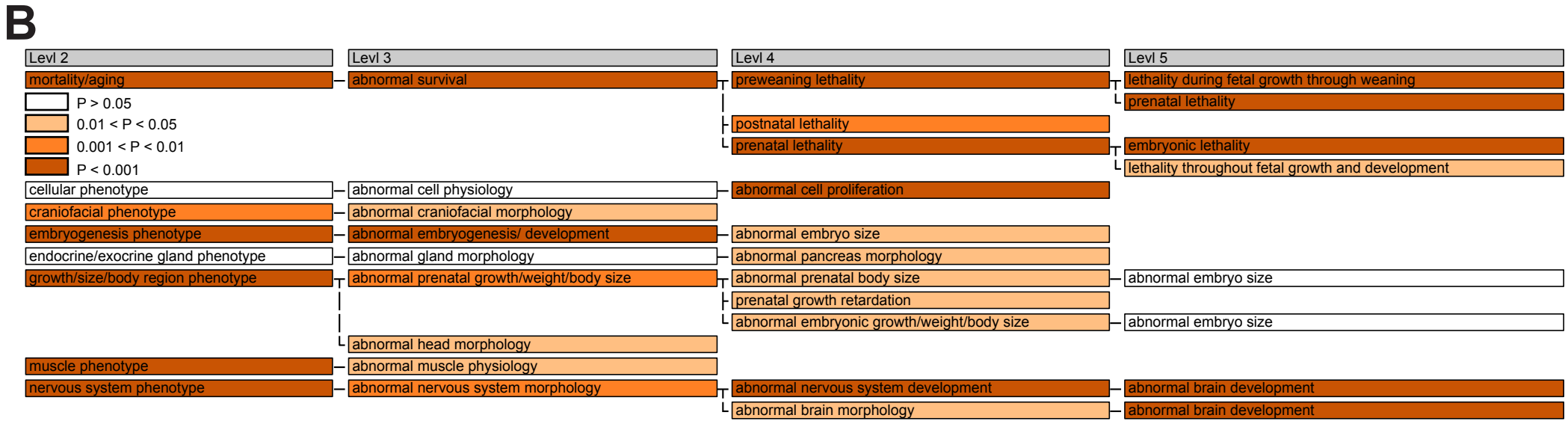

Rho guanyl nucleotide
exchange factor activity
SMAD binding
cellular response to
negative regulation of
damage stimulus
membrane organization
positive regulation of
insulin stimulus
protein complex
Notch signaling
pathway
androgen receptor
signaling pathway
cardioblast proliferation
complex assembly

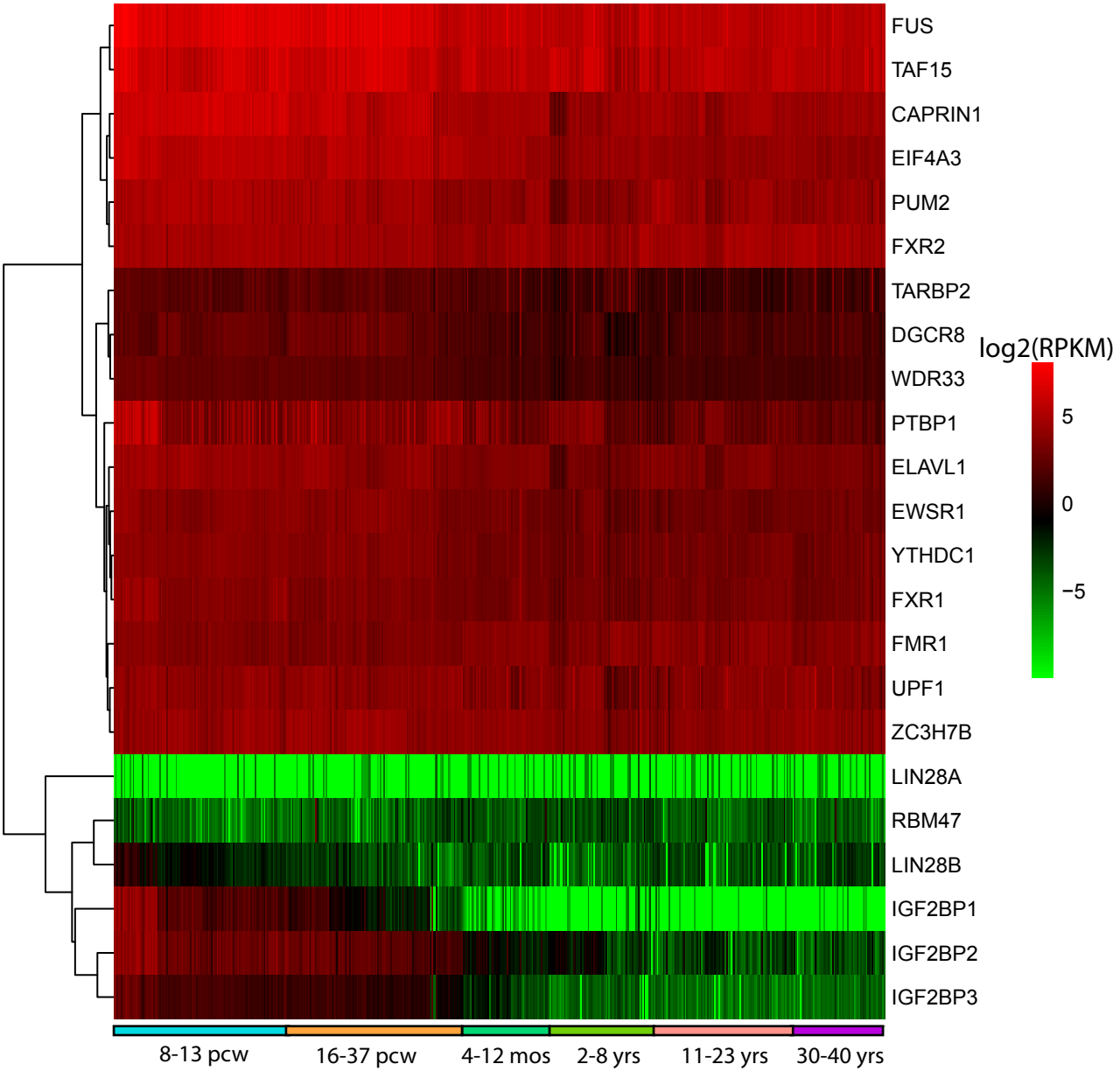



POGZ
KAT6B
MUC5B
KPNA1
BCORL1
USP38
CHAMP1
ZNF644
ZNF445
NEDD9
PIK3R3
DYRK1A
DNAH17
TNRC6A
ZNF248
MED13
CHD2
MED13L
CNOT3
EIF4G2
YLPM1
PHIP
UBR5
TLK2
IRS1
NUMA1
GNPTG
UBR4
SETD5
EP300
TRIO
SPATA5
CNOT1
SSH2
TCF4
DISP1
SON
SRRM2
SYNE2
DNMT3A
IMPDH2
KDM5B
KCTD20
TRRAP
RBM33
WDR19
PTPN21
FOXK2
CUL3
DIS3L2
JARID2
HNRNPU
KIRREL
CAD
SMARCA4
PLXNA1
ADNP
BRD4
WIZ
ACACA
PTK7
TNRC18
MAST2
DLG5
CTCF
ZC3H3
MECP2
PXDN
ARID1B
SETDB1
DOT1L
SMAP2
CLSTN3
ATP1A1
STXBP1


A
3
2
1
0
−1
−2
−3
Primitive.endoderm
Sperm
Trophectoderm
Early.blastocyst
Middle.blastocyst
ICM
Late.blastocyst
X4.cell
MII.oocyte
Pronucleus
X2.cell
Zygote
X8.cell
Morulae
Pachytene.spermatocyte
Round.spermatids
Syncytiotrophoblast
Trophoblast
Oocyte
Primordial.germ.cell
Epiblast
ESC
B
2
1
0
−1
−2
−3
GNAS
HECTD1
MAEA
USF2
ATP1A1
LARP7
PSD3
SYNE1
ITPR1
ARMC9
ATP2B4
BAIAP2
DNM1
HBS1L
MYH10
NAA10
DYNC1H1
PPP2R5D
AMPD2
COL4A3BP
MTOR
MADD
NF1
MYO5A
DMXL2
MECP2
EFR3A
RALGAPB
SYNGAP1
BRAF
PDE4DIP
ATP1A3
MET
PLEKHA6
FRY
CHD5
SCN1A
NEB
SMAP2
USP46
STXBP1
NISCH
ZRANB3
CLSTN3
HECW2
RYR2
PPL
JPH3
EEF1A2
BTBD9
GNAO1
KIF1A
SCN8A
TBC1D4
CDKL5
KIF5C
Sperm
Morulae
X4.cell
MII.oocyte
X2.cell
Zygote
Oocyte
Pronucleus
X8.cell
Pachytene.spermatocyte
Round.spermatids
Epiblast
ESC
Primordial.germ.cell
Syncytiotrophoblast
Trophoblast
Primitive.endoderm
Late.blastocyst
Trophectoderm
Early.blastocyst
ICM
Middle.blastocyst


A
Levl 2
Levl 3
Levl 4
Levl 5
mortality/aging
abnormal survival
partial lethality
preweaning lethality
lethality during fetal growth through weaning
prenatal lethality
postnatal lethality
partial postnatal lethality
perinatal lethality
prenatal lethality
embryonic lethality
P > 0.05
0.01 < P < 0.05
0.001 < P < 0.01
P < 0.001
skeleton phenotype
abnormal skeleton morphology
abnormal axial skeleton morphology
behavior/neurological phenotype
abnormal behavior
abnormal motor capabilities/coordination/movement
abnormal motor coordination/ balance
craniofacial phenotype
abnormal craniofacial morphology
embryogenesis phenotype
abnormal embryogenesis/ development
growth/size/body region phenotype
abnormal postnatal growth/weight/body size
abnormal postnatal growth
abnormal head morphology
muscle phenotype
abnormal muscle morphology
abnormal muscle development
abnormal muscle physiology
nervous system phenotype
abnormal nervous system physiology
abnormal synaptic transmission
abnormal nervous system morphology
abnormal nervous system development
abnormal brain development
abnormal neuron morphology
abnormal brain morphology
abnormal brain development


B
Levl 2
Levl 3
Levl 4
Levl 5
mortality/aging
abnormal survival
preweaning lethality
lethality during fetal growth through weaning
prenatal lethality
P > 0.05
0.01 < P < 0.05
0.001 < P < 0.01
P < 0.001
postnatal lethality
prenatal lethality
embryonic lethality
lethality throughout fetal growth and development
cellular phenotype
abnormal cell physiology
abnormal cell proliferation
craniofacial phenotype
abnormal craniofacial morphology
embryogenesis phenotype
abnormal embryogenesis/ development
abnormal embryo size
endocrine/exocrine gland phenotype
abnormal gland morphology
abnormal pancreas morphology
growth/size/body region phenotype
abnormal prenatal growth/weight/body size
abnormal prenatal body size
abnormal embryo size
prenatal growth retardation
abnormal embryonic growth/weight/body size
abnormal embryo size
abnormal head morphology
muscle phenotype
abnormal muscle physiology
nervous system phenotype
abnormal nervous system morphology
abnormal nervous system development
abnormal brain development
abnormal brain morphology
abnormal brain development


FUS
TAF15
CAPRIN1
EIF4A3
PUM2
FXR2
TARBP2
DGCR8
WDR33
PTBP1
ELAVL1
EWSR1
YTHDC1
FXR1
FMR1
UPF1
ZC3H7B
LIN28A
RBM47
LIN28B
IGF2BP1
IGF2BP2
IGF2BP3
log2(RPKM)
5
0
−5
8-13 pcw
16-37 pcw
4-12 mos
2-8 yrs
11-23 yrs
30-40 yrs