Integrative genomics identifies a convergent molecular subtype that links epigenomic with transcriptomic differences in autism

Gokul Ramaswami[1], Hyejung Won[1,8], Michael J. Gandal[1,2], Jillian Haney[1,2], Jerry C. Wang[1], Chloe C.Y. Wong[3], Wenjie Sun[4], Shyam Prabhakar[4], Jonathan Mill[5], Daniel H. Geschwind[1,6,7,*]

Affilations
1. Program in Neurogenetics, Department of Neurology, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA.

2. Department of Psychiatry, Semel Institute, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA

3. Institute of Psychiatry, Psychology and Neuroscience, King's College London, De Crespigny Park, London, UK

4. Computational and Systems Biology, Genome Institute of Singapore, Singapore

5. University of Exeter Medical School, University of Exeter, Exeter, UK

6. Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA

7. Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA

8. Present address: Department of Genetics and UNC Neuroscience Center, University of North Carolina, Chapel Hill, NC, USA

*. Correspondence should be addressed to D.H.G: dhg@mednet.ucla.edu

Abstract

Autism spectrum disorder (ASD) is a phenotypically and genetically heterogeneous neurodevelopmental disorder. Despite this heterogeneity, previous studies have shown patterns of molecular convergence in post-mortem brain tissue from autistic subjects. Here, we integrate genome-wide measures of mRNA expression, miRNA expression, DNA methylation, and histone acetylation from ASD and control brains to identify a convergent molecular subtype of ASD with shared dysregulation across both the epigenome and transcriptome. Focusing on this convergent subtype, we substantially expand the repertoire of differentially expressed genes in ASD and identify a component of upregulated immune processes that are associated with hypomethylation. We utilize eQTL and chromosome conformation datasets to link differentially acetylated regions with their cognate genes and identify an enrichment of ASD genetic risk variants in hyperacetylated noncoding regulatory regions linked to neuronal genes. These findings help elucidate how diverse genetic risk factors converge onto specific molecular processes in ASD.

ASD is a prevalent neurodevelopmental disorder characterized by impaired social interactions with repetitive and restrictive behaviors[1]. Although ASD is highly heritable, its genetic etiology is complex, with approximately 1,000 risk genes implicated[2]. Assessment of ASD risk is challenging due to its genetic architecture which encompasses alleles of varying frequencies (common, rare, very rare) and inheritance patterns (Mendelian autosomal and X-linked, additive, de novo)[3-5] that likely interact together within individuals and families[6,7].

Surprisingly, despite this genetic complexity, molecular studies have identified consistent patterns of changes in post-mortem brain tissue from ASD subjects[8-12]. At the transcriptomic level, ASD brains exhibit downregulation of genes involved in neuronal activity with a concomitant upregulation of genes involved in microglial and astrocyte-mediated inflammation[9,13]. Additionally, there is a shared pattern of microRNA (miRNA) dysregulation directly targeting downregulated neuronal genes as well as upregulated astrocyte genes[12]. At the epigenomic level, ASD brains exhibit DNA methylation differences in genomic regions related to immunity and neuronal regulation[11,14]. Additionally, there are differences in histone acetylation (H3K27ac) associated with genes involved in synaptic transmission and morphogenesis[10]. To date, these molecular datasets have not been comprehensively integrated and analyzed together, which could provide a better understanding of how epigenetic changes directly regulate expression of their cognate genes and how these processes are related. Additionally, despite evidence for shared patterns of molecular dysregulation, only approximately two-thirds of ASD brain samples exhibit this major shared molecular pattern, indicating the potential for distinct molecular subtypes. Such heterogeneity among ASD cases

would also be expected to reduce power to identify disease-related signals, providing another rationale for identification of subtypes.

Systems level integration of multi-omic datasets has been a successful strategy to identify molecular subtypes and elucidate causal mechanisms in cancer[15,16]. However, it has not yet been applied to neurodevelopmental disorders, including ASD. In this study, we utilize Similarity Network Fusion (SNF), a powerful integrative method that has identified molecular subtypes when integrating transcriptomic with epigenomic datasets in cancer[17], to integrate mRNA expression, miRNA expression, DNA methylation, and histone acetylation datasets from ASD brain (**Figure 1a**). This unbiased data driven analysis identifies two distinct molecular subtypes of ASD, one, which represents the majority of cases, showing a cohesive molecular pattern, and the other without consistent changes in molecular measures. By analyzing ASD brains according to subtype, which significantly reduces heterogeneity, we are able to identify three-fold more differentially expressed mRNA genes as compared to previous analyses. We identify differentially expressed miRNAs, differentially methylated promoters and gene bodies, as well as differentially acetylated genomic regions and assess the extent to which these regulatory mechanisms influence gene expression in ASD. Finally, we find an enrichment of ASD genetic risk in regulatory regions linked to neuronal genes that are hyperacetylated in ASD brains, suggesting a casual role for these elements.

Results

Integration of transcriptomic and epigenomic datasets from ASD and control brains

We integrated previously published datasets on mRNA expression[9], miRNA expression[12], DNA methylation[11], and histone acetylation[10] from a cohort of 48 ASD and 45 control brains (**Supplementary Table 1 and Figure 1a**). We only analyzed the samples originating from the frontal and temporal cortex, because previous studies found ASD dysregulated features were predominantly localized to the cerebral cortex and substantially attenuated in the cerebellum. For mRNA and miRNA expression, we used normalized gene quantifications and differential expression summary statistics from the previous studies[9,12] (**Methods**).

For DNA methylation, we used the normalized probe quantifications from the previous study[11] and collapsed probe level measurements onto 21880 gene promoters and 24458 gene bodies to facilitate comparisons between methylation and expression (**Methods**). An initial differential methylation analysis identified 2578 and 1262 differentially methylated promoters and gene bodies, respectively, at an FDR < 10% (**Methods, Supplementary Figure 1a-b, and Supplementary Table 4**). The genes with differential promoter methylation were largely distinct from genes with differential gene body methylation (**Supplementary Figure 1e-f**). However, the loadings for each sample along the first principal component of differential promoter and gene body methylation were almost identical (**Supplementary Figure 1d**) and not correlated with any potential confounders (**Supplementary Figure 1c**), suggesting a coherent regulatory mechanism.

For histone acetylation, we reprocessed the dataset[10] entirely (**Methods**) by mirroring established ChIP-seq processing pipelines from the ENCODE[18] and Roadmap Epigenomics projects[19]. We quantified 50773 consensus H3K27ac peaks across the genome, which largely overlapped

H3K27ac peaks identified in the previous study (**Supplementary Figure 2a-c**), identifying 274 differentially acetylated regions at an FDR < 20% (**Methods, Supplementary Figure 2d, and Supplementary Table 5**). Although this was thirty-fold fewer differentially acetylated regions than previously identified (**Supplementary Figure 2e**), we show that this depletion is an artifact of subtype heterogeneity in the ASD samples (see below *Subtype-specific histone acetylation differences in ASD*).

Identification of two ASD molecular subtypes

In the previous molecular studies[9-12], we noticed that approximately two-thirds of ASD brain samples clustered together based on the differential signal for each dataset. To formally assess ASD molecular heterogeneity across the four different datasets, we used SNF[17] to integrate differential mRNA expression, miRNA expression, DNA methylation, and histone acetylation for 30 ASD and 17 control samples that were present in all 4 molecular datasets (**Methods**). SNF creates an integrative sample-sample similarity network by quantifying sample-sample relationships within each individual dataset and then integrating these sample-sample relationships across all of the datasets[17]. The clustering of sample relationships is a major advantage of SNF, as compared to alternative data integration methods that cluster gene relationships which can be sensitive to differing normalization methods between data types[20].

The sample loadings along the first principal component of each differential molecular level recapitulated known regulatory relationships, with differential acetylation (R = 0.70) and differential miRNA expression (R = 0.51) being highly correlated to differential mRNA expression, whereas differential methylation was less correlated with expression (R = 0.13)[21,22]

(**Figure 1b**). Using this similarity network, samples divided into two distinct clusters (**Figure 1c**), one of which (SNF Group 2) consisted entirely of ASD samples that loaded strongly onto the differential transcriptomic and epigenomic signatures. Therefore, we grouped these samples together as the ASD "Convergent" Subtype. The other cluster (SNF Group 1) consisted of ASD samples that did not load onto the initial differential signatures and were indistinguishable from controls. Therefore, we grouped these samples together as the ASD "Disparate" Subtype. We built a logistic regression classifier to assign the 61 ASD and 61 control samples that were not used in at least one of the four classification datasets into either the Convergent or Disparate subtypes (**Methods, Figure 1d, and Supplementary Figure 4**). Interestingly, 11 of the 43 ASD individuals with samples from both frontal and temporal cortex were classified into different molecular subtypes in the 2 cortical regions (**Supplementary Figure 4d**), a significant difference in comparison to only 2 of the 33 control individuals ($p = 0.032$, Fisher's Exact Test). This finding is consistent with potential molecular heterogeneity across different cortical regions in ASD.

SNF subtype assignments were robust to clustering methodology and comprehensive leave one out cross validation (**Supplementary Figure 3c-d**). We further tested clustering robustness by leaving out each dataset and performing SNF clustering and logistic regression classification using the remaining three datasets. We found the resultant sample subtype assignments were highly concordant (range = 0.89-0.92) with the subtype assignments identified using the entire four dataset collection (**Figure 1e and Supplementary Figure 5**). Additionally, the ASD subtype assignments were not correlated with, or driven by, biological or technical covariates, including age, sex, RNA quality, cell fraction, and post mortem interval (**Supplementary Figure**

6). Finally, to compare with the sample classifications above, which were generated using only differential features in each dataset, we attempted to cluster the samples with SNF using all features in the transcriptomic and epigenomic datasets. We find no clear separation between ASD and control samples (**Supplementary Figure 3a-b**), demonstrating that molecular differences between ASD and control brains are restricted to specific differentiating features and are not a general genome-wide phenomenon.

Subtype-specific mRNA expression differences in ASD

To leverage the increased power of analyzing a more homogenous set of cases, we performed differential mRNA expression analyses separately for ASD Convergent and Disparate subtypes against control samples for each gene (**Methods and Supplementary Figure 7a-b**). For the ASD Convergent subtype, we observed 4599 differentially expressed genes at an FDR < 5%, 1999 of which were upregulated and 2600 downregulated in ASD (**Supplementary Table 2**). We reproduce 94% of the differentially expressed genes from the previous study[9] and identify an additional 3525 genes (**Figure 2a-b**), demonstrating the utility of this subtype-specific approach. The top gene ontology enrichments are very similar to those previously identified, showing an upregulation of genes involved in immune response and a downregulation of genes involved in synaptic transmission and neuronal ion transport (**Supplementary Figure 7c-d**). In contrast, for the ASD Disparate subtype, we found no differentially expressed genes.

Next, we identified mRNA co-expression modules that were differentially associated between ASD and control individuals in the cortical co-expression network defined previously[9]. For each gene module, we tested whether the two ASD subtypes and control samples had differences in

their association with the module eigengene, a summary measure of module expression level (**Supplementary Figure 7e**). For the ASD Convergent subtype, we found 13 differentially associated co-expression modules at an FDR < 5%, including the 6 ASD-associated modules identified previously[9] and 7 newly associated modules (**Figure 2c**). In contrast, for the ASD Disparate subtype, we did not find any differentially associated co-expression modules.

Of the 7 newly identified ASD-associated co-expression modules in this study, 4 were downregulated in ASD: mRNA.M1, a module representing neurogenesis, mRNA.M3, a module representing mitochondrial function in neurons which has been previously implicated in ASD[23], mRNA.M7, a module with no functional enrichments, and mRNA.M17, a module representing synaptic signaling and vesicle transport in neurons (**Figure 2d-e**). To gain insight into neuronal downregulation in ASD at a finer resolution, we compared ASD downregulated modules to neuronal cell type specific markers identified from single-nuclei RNA sequencing of post-mortem human cortex[24]. ASD downregulated modules are significantly enriched with markers of both inhibitory and excitatory neurons (**Figure 2f**). The strongest enrichments are inhibitory neuron subtypes expressing *SST* or *PVALB*, derived from the medial ganglionic eminence as well as deep layer excitatory neurons expressing *RORB* or *FEZF2* (**Figure 2f**), suggesting that the number and/or activity of these cells is heavily decreased in ASD, consistent with a recent single cell analysis of post mortem ASD brain[25].

Three of the newly identified ASD-associated modules were upregulated in ASD: mRNA.M15, a module representing metabolic processes and transcriptional regulation in glia, mRNA.M21, a module representing ribosomal translational, and mRNA.M23, a module enriched with astrocyte

markers. Module mRNA.M15 (**Figure 2g-h**) was particularly interesting, because one of its top hub genes is *REST*, a transcriptional repressor with critical roles in repressing neural genes in non-neural cells[26]. Although module mRNA.M15 is enriched with microglial markers (**Supplementary Figure 7g**), it exhibits a markedly different transcriptional profile[27] than the previously identified ASD upregulated microglial module mRNA.M19 (**Figure 2i**). Module mRNA.M19 is specifically enriched with genes marking microglial activation[28], suggesting that it is directly related to neural-immune response. In contrast, module mRNA.M15 is enriched with markers of juvenile or aging glia, suggesting it may be related to glial growth and maturation. In general, cellular processes underlying broad categories of immuno-glial cell types are upregulated in ASD.

Subtype-specific miRNA expression differences in ASD

We conducted differential miRNA expression analyses for ASD Convergent and Disparate subtypes against control samples for each mature miRNA transcript (**Methods and Supplementary Figure 8a-b**). For the ASD Convergent subtype, we identified 44 differentially expressed miRNAs at an FDR < 5%, 28 upregulated and 16 downregulated in ASD that highly overlapped with the previous study (52%; **Supplementary Figure 8c-d and Supplementary Table 3**)[12]. We analyzed differentially associated miRNA co-expression modules in the miRNA co-expression network defined previously[12]. For the ASD Convergent subtype, we found the same 3 miRNA co-expression modules differentially associated at an FDR < 5% as the previous study: miRNA.brown, which is downregulated in ASD as well as miRNA.magenta and miRNA.yellow, which are upregulated (**Supplementary Figure 8e**). We used TargetScan[29] to predict mRNA targets for the top hubs of each differentially associated miRNA co-expression

module (**Supplementary Table 3**). We found a slight enrichment of genes in the upregulated mRNA.M19 module within the predicted targets of the miRNA.brown module, as well as genes in the downregulated mRNA.M16 module within the predicted targets of both miRNA.brown and miRNA.yellow (**Supplementary Figure 8f**), suggesting a moderate influence of differential miRNA expression on differential gene expression in ASD. In contrast, for the ASD Disparate subtype, we did not find any differentially expressed miRNAs or differentially associated miRNA co-expression modules. Overall, subtype-specific analyses of miRNA expression largely recapitulated findings from previous work[12].

Subtype-specific DNA methylation differences in ASD

We conducted differential DNA methylation analyses for ASD Convergent and Disparate subtypes against control samples for gene promoters and gene bodies (**Methods**). For the ASD Convergent subtype, we identified 4187 differentially methylated gene promoters at an FDR < 5%, 3221 hypermethylated and 966 hypomethylated in ASD (**Supplementary Figure 9b and Supplementary Table 4**). ASD hypermethylated gene promoters are enriched in RNA processing genes, while ASD hypomethylated gene promoters are enriched in chemical sensory receptor genes (**Supplementary Figure 9d-e**). We identified 2415 differentially methylated gene bodies at an FDR < 5%, 1146 hypermethylated and 1269 hypomethylated in ASD (**Supplementary Figure 10b and Supplementary Table 4**). ASD hypermethylated gene bodies are enriched in RNA processing genes, while ASD hypomethylated gene bodies are enriched in keratinization and bile acid transport genes (**Supplementary Figure 10d-e**). We assigned the previously identified differentially methylated probes[11] to their corresponding promoter or gene body annotation and now identify seventy and twenty-fold more differentially methylated

promoters (**Supplementary Figure 9h-i**) and gene bodies (**Supplementary Figure 10h-i**), respectively. In contrast, for the ASD Disparate subtype, we did not find any differentially methylated gene promoters or gene bodies (**Supplementary Figures 9c and 10c**). Genes with differentially methylated gene promoters were largely distinct from those with differentially methylated gene bodies although there was a greater overlap in hypermethylated genes due to their shared biological enrichments (**Figure 3a-b**). There were 711 and 412 genes that were both differentially expressed as well as differentially methylated at gene promoters and gene bodies, respectively. As expected, there was a negative correlation between differential expression and differential methylation for both gene promoters and gene bodies (**Figure 3c-d**), although surprisingly gene body methylation was more strongly negatively correlated with expression than promoter methylation.

We generated co-methylation networks for both promoters and gene bodies (**Supplementary Figures 9a and 10a**) which recapitulated many aspects of the probe-level co-methylation network that was previously built[11] (**Supplementary Figure 11a-b**). We identified 8 ASD-associated promoter co-methylation modules, 4 hypermethylated and 4 hypomethylated in ASD (FDR < 0.05; **Supplementary Figure 9f**). The hypermethylated promotor modules were: Prom.midnightblue, representing coenzyme A biosynthesis, Prom.pink, a module with no functional enrichments, Prom.tan, enriched with oligodendrocyte cell markers, and Prom.turquoise, enriched with astrocyte cell markers and representing RNA processing. The hypomethylated modules were: Prom.brown and Prom.greenyellow, two modules representing sensory perception, as well as Prom.lightcyan and Prom.lightgreen, two modules representing immune processes.

At the gene body level, we identified 10 ASD-associated co-methylation modules, 4 hypermethylated and 6 hypomethylated in ASD (FDR < 0.05; **Supplementary Figure 10f**). The hypermethylated modules were: GB.blue, representing RNA processing, GB.cyan, enriched in neuron cell markers and representing the unfolded protein response, GB.darkred, enriched in neuron cell markers and representing mitochondrial activity, and GB.royalblue, a module with no functional enrichments. The hypomethylated modules were: GB.black, GB.darkgreen, GB.lightcyan, and GB.salmon, 4 modules representing immune processes, as well as GB.green, a module representing glucuronidation, and GB.yellow, representing bile acid ion transport. In contrast, for the ASD Disparate subtype, we did not find any ASD-associated promoter or gene body co-methylation modules.

Overall, the ASD-associated co-methylation modules did not show significant global overlap with the ASD-associated co-expression modules at the gene level (**Supplementary Figure 11c-d**), suggesting that differential methylation is not a prominent driver of differential gene expression. The largest overlaps are between hypomethylated co-methylation modules and upregulated co-expression modules involved in immune processes. In particular, Prom.lightgreen overlaps significantly with mRNA.M19 and GB.darkgreen overlaps significantly with mRNA.M15, suggesting that the ASD-associated upregulation in immune activation is, in part, regulated by a decrease in DNA methylation at promoters and gene bodies (**Figure 3e-j**).

Subtype-specific histone acetylation differences in ASD

We next performed differential histone acetylation analyses for ASD Convergent and Disparate subtypes against control samples for H3K27ac peaks in the genome (**Methods**). For the ASD Convergent subtype, we identified 6894 differentially acetylated peaks at an FDR < 10%, 3577 hyperacetylated and 3317 hypoacetylated in ASD (**Supplementary Figure 12a and Supplementary Table 5**). There was a strong overlap with differentially acetylated peaks identified in a previous study (**Supplementary Figure 12c-d**)[10]. In contrast, for the ASD Disparate subtype, we did not find any differentially acetylated peaks (**Supplementary Figure 12b**). Using GREAT to assess ontology enrichments for genes closest to each differentially acetylated peak[30], we find that genes proximal to ASD hyperacetylated peaks are enriched in GABA receptor activity while genes proximal to ASD hypoacetylated peaks are enriched in neurogenesis and brain development (**Figure 4a-b**). H3K27ac is known to mark active promoters[31], and as expected, differentially acetylated peaks in gene promoters were strongly positively correlated (R = 0.43) with differential expression (**Figure 4d**). Surprisingly, we found that the relationship between differential promoter acetylation and differential expression was cell type specific, with hyperacetylated promoters associated with upregulated microglial genes and downregulated neuronal genes, whereas hypoacetylated promoters were associated with upregulated astrocyte genes and downregulated oligodendrocyte genes (**Figure 4e**).

In addition to marking promoters, H3K27ac also marks distal enhancers up to 1 MB away from gene transcription start sites (TSS)[32]. We utilized expression quantitative trait loci (eQTL) and chromatin conformation capture (Hi-C) datasets from bulk adult brain tissue[33], as well as Hi-C data using sorted neuronal and glial cells from adult brain tissue (**Methods**) to link distal differential H3K27ac peaks with their cognate genes (**Figure 4c; Methods**). As a baseline, we

analyzed all differentially acetylated peaks within 1 MB of a differentially expressed gene TSS and found that the overall correlation between differential acetylation and differential expression was minute (R = 0.012, p = 0.023) (**Supplementary Figure 12e**). The correlations improved when linking differentially acetylated regions to differentially expressed genes using eQTL (R = 0.061, p = 0.0012) (**Supplementary Figure 12f**) and bulk Hi-C (R = 0.067, p = 0.015) (**Supplementary Figure 12g**), but remained small genome wide. However, using cell type specific Hi-C data we were able to see that the correlation was driven largely by glial specific interactions (neuronal Hi-C: R = 0.046, p = 0.054; glial Hi-C: R = 0.12, p = 5e-5) (**Supplementary Figures 12h-i**), reflecting the greater number of glia as compared with neurons in the cerebral cortex[34]. The correlations were further improved when H3K27ac peaks were linked to genes using a combination of eQTL and cell type specific Hi-C datasets (neuronal: R = 0.14, p = 0.038; glial: R = 0.2, p = 0.0031) (**Figure 4f-g)**. However, the combination of eQTL and bulk Hi-C was substantially less correlated (R = 0.037, p = 0.27) (**Supplementary Figure 12j**), suggesting that chromatin contacts display remarkable cell-type specificity and this distinction is attenuated when looking at bulk tissue.

We provide a listing of potential cognate genes linked to differentially acetylated peaks using the eQTL, neuronal Hi-C, and glial Hi-C linkages, emphasizing the highest confidence interactions based on those identified in both a Hi-C dataset and the eQTL data (**Supplementary Table 5**). Using this list of cognate genes, we identified gene co-expression modules potentially regulated by ASD-associated acetylation changes (**Figure 4h**). We find an enrichment of ASD hypoacetylation at promoters in mRNA.M9, an astrocyte module that is upregulated in ASD. Additionally, we find an enrichment of ASD hyperacetylation at distal enhancers in mRNA.M16,

a neuron module that is downregulated in ASD. Surprisingly, in both of these cases, acetylation changes are negatively correlated with expression changes. This suggests, for genes in these modules, that ASD-associated acetylation changes are not causal, but compensatory to gene expression changes, with components of hypoacetylation linked to upregulated astrocyte processes and hyperacetylation linked to downregulated neuronal processes (**Figure 4i**).

Enrichments of ASD heritability in dysregulated genomic regions

Differential gene expression or epigenetic changes may either be contributory to, or a consequence of disease. To provide a causal anchor, we used stratified LD score regression[35] to partition heritability of ASD risk variants from genome wide association studies[4,36] into regions of the genome that are differentially expressed, methylated, or acetylated. We found a significant enrichment of heritability in genomic regions that are hyperacetylated in ASD brains (**Figure 5a**). Specifically, this enrichment is found at distal enhancer regions and not at gene promoters (**Figure 5b**), highlighting the importance of noncoding regulatory elements. Cognate genes linked to hyperacetylated enhancers are enriched within module mRNA.M16 (**Figure 4h**), an ASD-downregulated neuronal module representing genes involved in learning, memory, and behavior (**Figure 5c-d**). This finding supports previous observations that common genetic risk variants for ASD are enriched in regulatory regions of neuronal genes[4,13]. Among the genes within module mRNA.M16, three of its top hub genes are linked to elements hyperacetylated in ASD: *NSF, PRKCE,* and *SCN8A* (**Figure 5e-g**), suggesting that an increase in acetylation is attempting to compensate downregulation of key driver genes of this module. Next, we looked for ASD heritability enrichments in the co-expression and co-methylation network modules (**Supplementary Figure 13a-c**). We found a significant enrichment of heritability within the

promoter co-methylation module Prom.green, which represents genes involved in neurogenesis (**Supplementary Figure 13d-e**), further strengthening the observation that ASD risk variants reside within regulatory regions of genes involved in neuronal function and neurogenesis.

Taken together (**Figure 6**), our findings imply a model whereby ASD risk variants perturb regulatory elements controlling genes in co-expression module M16, leading to the overall downregulation of synaptic signaling and neuronal ion transport observed. This in turn, leads to the transcriptional upregulation of astrocyte and microglial mediated immune processes through a concomitant decrease in DNA methylation and decrease in expression of associated miRNAs. Finally, as a response to compensate these transcriptional changes, there is a decrease of histone acetylation at the promoters of upregulated astrocyte genes as well as an increase of histone acetylation linked to downregulated neuronal genes at the same regulatory elements that were initially impacted by the casual genetic variants.

Discussion

We integrate mRNA expression, miRNA expression, DNA methylation, and histone acetylation datasets to identify a subtype of ASD brain samples with convergent dysregulation across the epigenome and transcriptome. By focusing on this convergent ASD subtype, we identify a three-fold expansion in differentially expressed mRNAs and co-expression modules encompassing the major processes of neuronal downregulation and immune upregulation. We identify thousands of differentially methylated gene promoters and gene bodies, but only a small proportion of the methylation changes, specifically hypomethylation of immune genes, seem to influence gene expression regulation. In contrast, histone acetylation is a strong positive regulator of mRNA

expression[21] and we identify thousands of differentially acetylated regions across the genome and furthermore assign them to cognate genes using eQTL and chromosome conformation datasets. We find differentially acetylated regions enriched within dysregulated astrocyte and neuronal co-expression modules. Surprisingly the acetylation changes in these regions are negatively correlated with expression changes, implying that these changes are compensatory, as previously suggested[10].

Over 50% of ASD genetic liability is carried by small effect size common variants which are mostly noncoding[3]. One of the major challenges in characterizing these common variants is the ability to link noncoding regulatory regions with their cognate genes, which are often dynamically regulated across different cell types and developmental stages. In this study, we find an enrichment of ASD genetic risk within hyperacetylated regions of the genome, specifically those linked to downregulated neuronal genes. While enrichment of ASD risk variants in brain regulatory elements, including those marked by H3K27ac, has been observed before[4,37], this is the first study reporting that ASD risk variants are enriched in differentially regulated enhancers in ASD brains, providing a potential mechanistic understanding of how non-coding ASD risk variants impact gene regulation. We find that neuron-specific and glial-specific Hi-C datasets perform better than a bulk tissue Hi-C dataset in linking distal regulatory elements with genes (**Figures 4f-g and Supplementary Figure 12j**), demonstrating the critical need for cell-type specific regulatory maps of the noncoding genome.

A major unanswered question is what molecular processes are involved in the ASD Disparate subtype samples. We were unable to find any consensus molecular dysregulation in these

samples across all the datasets (**Supplementary Figures 7b, 8b, 9c, 10c, 12b**). We also checked whether they were misdiagnosed, but the available clinical records for these individuals are sparse and we found no samples exhibiting differential expression signatures for four other neuropsychiatric disorders (**Supplementary Figure 14**). We noticed that 11 of the ASD subjects were classified into different molecular subtypes when comparing frontal and temporal cortex (**Supplementary Figure 4d**), suggesting that the extent of molecular dysregulation may vary across regions of the cortex in these individuals. As an initial assessment of regional heterogeneity, we analyzed RNA-seq data derived from four additional cortical areas in a subset of individuals in this study[23], including portions of the parietal and occipital lobes, and find considerable heterogeneity in gene expression dysregulation across the cortex (**Supplementary Figure 15**). Future studies of ASD post-mortem brain samples with larger sample sizes, assessing more brain regions, and availability of comprehensive phenotypic data will be needed to further characterize the regional specificity of molecular changes in ASD.

Online Methods

Initial processing of datasets

All molecular datasets came from a cohort of 48 ASD and 45 control brains from the Harvard Autism Tissue Program (https://hbtrc.mclean.harvard.edu/) and NIH Neuro Brain Bank (http://www.medschool.umaryland.edu/btbank/). Initial analyses from the transcriptomic and epigenomic datasets have been previously published[9-12]. We restricted this study to only those samples originating from the frontal or temporal cortex. In the epigenomic studies[10,11], samples from the Oxford and MRC London Brain Banks were also present, however we removed these samples because we did not have transcriptomic data for them.

For mRNA expression, we used the quantification of log2 RPKM values for 16310 genes in 82 ASD samples and 74 control samples from 47 ASD and 44 control brains from Parikshak et al[9] (https://github.com/dhglab/Genome-wide-changes-in-lncRNA-alternative-splicing-and-cortical-patterning-in-autism). These RPKM values were normalized for gene length and GC content using CQN[38], but not adjusted for technical or biological covariates. We downloaded the differential mRNA gene expression summary statistics and mRNA co-expression network module definitions. We calculated principal components (SeqStatPCs) to summarize the following sequencing statistics: log10(TotalReads.picard), log10(Aligned.Reads.picard), log10(HQ.Aligned.Reads.picard), log10(PF.All.Bases.picard), log10(Coding.Bases.picard), log10(UTR.Bases.picard), log10(Intronic.Bases.picard), log10(Intergenic.bases.picard), Median.CV.Coverage.picard, Median.5prime.Bias.picard, Median.3prime.Bias.picard, Median.5to3prime.Bias.picard, AT.Dropout.picard, GC.Dropout.picard, and PropExonicReads.HTSC.

For miRNA expression, we used the quantification of log2 read counts mapping to 699 mature miRNAs in 60 ASD samples and 42 control samples from 39 ASD and 28 control brains from Wu et al[12]. To balance the case/control cohorts with respect to age, we removed all samples from younger individuals (Age <= 10) leaving us with 49 ASD samples and 42 control samples from 31 ASD and 28 control brains. These read counts were normalized for mature miRNA GC content using CQN[38] and batch effects using ComBat[39], but not adjusted for technical or biological covariates. We downloaded the differential miRNA expression summary statistics and miRNA co-expression network module definitions.

For DNA methylation, we used the quantification of methylation beta values for 417460 CpG probes across the genome in 74 ASD samples and 42 control samples from 42 ASD and 27 control brains from Wong et al[11]. To balance the case/control cohorts with respect to age, we removed all samples from younger individuals (Age <= 10) leaving us with 56 ASD samples and 41 control samples from 33 ASD and 26 control brains. Probe quantifications were normalized using wateRmelon as previously described[40]. For each sample, the CET score was calculated[41] which is the proportion of neuronal vs glial cells. For samples where we had expression and/or acetylation data but did not have methylation data, we assigned their CET score as the average CET score. We collapsed the probe level measurements to gene promoters and gene bodies by taking the average methylation level of all probes mapping onto gene promoters (2KB upstream of TSS to TSS) and gene bodies (TSS to transcription end site (TES)) using gencode v19 annotations[42]. We only kept gene promoters or gene bodies that contained 2 or more CpG probes. We downloaded the list of differentially methylated probes and co-methylation network identified from the cross-cortex analysis of Wong et al[11]. These probes were assigned to gene promoters and gene bodies as described above.

For histone acetylation, we downloaded fastq files for 257 H3K27ac ChIP-seq samples as well as input samples from frontal cortex, temporal cortex, and cerebellum from Sun et al[10] (https://www.synapse.org/#!Synapse:syn8104916). We mapped reads from each sample onto the hg19 reference genome using BWA-MEM[43] with default parameters. We removed duplicate reads using Picard tools (http://broadinstitute.github.io/picard/). For each sample, we identified H3K27ac gapped peaks using MACS2[44] callpeak with parameters --broad, -g hs, -q 0.05, and –c

the relevant tissue input sample. We filtered peaks to those with a fold change > 3 and qvalue < 0.01. We looked at the size distribution of all peaks and removed large outlier peaks > 10,503 bp in size (third quartile + 2.5 * interquartile range). We also removed peaks from the ENCODE blacklist regions[45] (http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/hg19-human/wgEncodeHg19ConsensusSignalArtifactRegions.bed.gz). As a Q/C check, we used phantompeakqualtools[18] to calculate ChIP-seq cross-correlation statistics. We removed samples failing Q/C: those with total reads < 10,000,000, read alignment fraction < 75%, read duplication fraction > 30%, called peaks < 7,500, normalized strand coefficient (NSC) < 1.03, relative strand correlation (RSC) < 0.5, or fraction of reads in peaks < 3%. We also removed all cerebellum samples and samples originating from the Oxford or MRC London Brain Banks. This left us with a dataset of 56 ASD samples and 48 control samples from 35 ASD and 33 control brains. We identified 50773 consensus H3K27ac peaks. Each consensus peak overlapped at least one peak called in at least 20 samples. The boundary of each consensus peak was set as the union of all overlapping peaks in the individual samples. We quantified the levels of each consensus peak in each sample by counting the number of overlapping reads and dividing by the library size (in millions of reads). The log2 normalized peak quantifications were used in further analyses.

Similarity network fusion to identify molecular subtypes

For 30 ASD and 17 control samples present in all four datasets, we used SNF[17] to cluster samples together based on their relationships across the four data types. Before running SNF, we adjusted the datasets to remove the influence of technical and biological covariates. For mRNA expression, we fit a linear model for each gene: Expression ~ Diagnosis + Age + Sex + Region + RIN + Brain bank + Sequencing batch + seqStatPC1 + seqStatPC2 + seqStatPC3 + seqStatPC4 +

seqStatPC5 and regressed out the effect of all covariates except Diagnosis. For miRNA expression, we fit a linear model for each miRNA transcript: Expression ~ Diagnosis + Age + Sex + Region + RIN + Brain bank + Proportion of reads mapping to exons + log10(Sequencing depth) + PMI and regressed out the effect of all covariates except Diagnosis. For DNA methylation, we fit a linear model for each gene promoter and gene body: Methylation ~ Diagnosis + Age + Sex + Region + Brain bank + Batch + CET and regressed out the effect of all covariates except Diagnosis. For histone acetylation, we fit a linear model for each H3K27ac peak: Acetylation ~ Diagnosis + Age + Sex + Region + Brain bank + CET + Number of peaks in sample + Fraction of reads in peaks + Duplicate read fraction + Aligned read fraction and regressed out the effect of all covariates except Diagnosis.

To identify ASD molecular subtypes, we restricted each dataset to its differential features between ASD and control samples. For mRNA expression, we restricted the genes to 2591 differentially expressed genes at an FDR < 10% from the idiopathic ASD vs control analysis of Parikshak et al[9]. For miRNA expression, we restricted the miRNA transcripts to 92 differentially expressed miRNAs at an FDR < 10% from Wu et al[12]. For DNA methylation, we ran an initial differential methylation analysis looking at all ASD vs control samples (for details, see below *ASD vs control differential molecular analyses*) and restricted the genes to 2578 differentially methylated promoters at an FDR < 10%. For histone acetylation, we ran an initial differential acetylation analysis looking at all ASD vs control samples (for details, see below *ASD vs control differential molecular analyses*) and restricted the peaks to 274 differentially acetylated peaks at an FDR < 20%.

We ran SNF on the four adjusted and restricted datasets using the SNFtool package in R (https://cran.r-project.org/web/packages/SNFtool/index.html). For each dataset, we normalized the values of each feature using the standardNormalization() function in SNF. For each dataset, we calculated a sample-sample Euclidean distance using the dist2() function and used this distance to calculate a sample-sample affinity matrix using the affinityMatrix() function with parameters: K = 20 and alpha = 0.5. We generated a fused affinity matrix combining all 4 affinity matrices using the SNF() function with parameters: K = 20 and T = 15 and then used the spectralClustering() function to demarcate the fused affinity matrix into 2 clusters of samples. Alternatively, we also used the symnmf_newton() function in Matlab from the symNMF package[46] to cluster the fused affinity matrix.

Classification of samples into the two SNF clusters

For 61 ASD and 61 control samples that were missing from at least one of the four datasets, we built logistic regression models to classify them into one of the two clusters identified by SNF. For each of the 4 datasets, we calculated the sample loadings on the first principal component (PC1) of its differential features used as input to SNF. These PC1 loadings were transformed into Z-scores and used as predictors in the models.

We built a total of twelve different logistic regression models. Three models were for samples present in three datasets: mRNA/miRNA/DNA methylation, mRNA/miRNA/histone acetylation, and mRNA/DNA methylation/histone acetylation. Six models were for samples present in two datasets: mRNA/miRNA, mRNA/DNA methylation, mRNA/histone acetylation, miRNA/DNA methylation, miRNA/histone acetylation, and DNA methylation/histone acetylation. Finally,

three models were for samples present in one dataset: mRNA, DNA methylation, and histone acetylation. There were no samples that were present in miRNA/DNA methylation/histone acetylation or miRNA only datasets.

For each model, the training set was the 47 sample assignments identified by SNF. In training the model, the response variable was either 0 (for SNF group 1) or 1 (for SNF group 2) and the predictors were the sample differential Z-scores. We performed exhaustive leave one out validation on the training set by leaving each sample out, training the model with the remaining samples and predicting the response of the held-out sample. For each model, we chose a cutoff to distinguish between the two groups that maximized the cross-validation accuracy. We classified the test samples by running them through the model and applying the chosen cutoff.

Outlier removal

Before running differential analyses and WGCNA, we removed outlier samples from each of the 4 datasets. For each dataset, we calculated sample-sample correlations and removed samples with a signed Z-score > 3, as previously described[47].

ASD vs control differential molecular analyses

For mRNA expression, we ran differential expression analysis by fitting a linear mixed effect model for each gene: Expression ~ Diagnosis + Age + Sex + Region + RIN + Brain bank + Sequencing batch + seqStatPC1 + seqStatPC2 + seqStatPC3 + seqStatPC4 + seqStatPC5 as fixed effects, and brainID as a random effect. For miRNA expression, we ran differential expression analysis by fitting a linear mixed effect model for each miRNA transcript: Expression ~

Diagnosis + Age + Sex + Region + RIN + Brain bank + Proportion of reads mapping to exons + log10(Sequencing depth) + PMI as fixed effects, and brainID as a random effect. For DNA methylation, we ran differential methylation analysis by fitting a linear mixed effect model for each promoter and each gene body: Methylation ~ Diagnosis + Age + Sex + Region + Brain bank + Batch + CET as fixed effects, and brainID as a random effect. For histone acetylation, we ran differential acetylation analysis by fitting a linear mixed effect model for each H3K27ac peak: Acetylation ~ Diagnosis + Age + Sex + Region + Brain bank + CET + Number of peaks in sample + Fraction of reads in peaks + Duplicate read fraction + Aligned read fraction as fixed effects, and brainID as a random effect. For all differential analyses, we ran them separately for ASD Convergent subtype vs control samples and ASD Disparate subtype vs control samples. For DNA methylation and histone acetylation, we also ran an initial analysis looking at all ASD vs control samples before running SNF and sample classification.

Co-expression /co-methylation network analysis (WGCNA)

For mRNA and miRNA expression, we used the co-expression networks defined in the previous studies[9,12]. To identify ASD-associated modules, we fit a linear mixed effect model for each co-expression module. For mRNA modules: module eigengene ~ Diagnosis + Age + Sex + Region + RIN + Brain bank + Sequencing batch + seqStatPC1 + seqStatPC2 + seqStatPC3 + seqStatPC4 + seqStatPC5 as fixed effects, and brainID as a random effect. For miRNA modules: module eigengene ~ Diagnosis + Age + Sex + Region + RIN + Brain bank + Proportion of reads mapping to exons + log10(Sequencing depth) + PMI as fixed effects, and brainID as a random effect. We ran these analyses separately for ASD Convergent subtype vs control samples and ASD Disparate subtype vs control samples.

For DNA methylation, we generated co-methylation networks separately for both gene promoters and gene bodies. First, we set a linear model for each gene promoter and gene body as: Methylation ~ Diagnosis + Age + Sex + Region + Brain bank + Batch + CET and regressed out the effect of Brain bank. We generated networks with robust consensus WGCNA (rWGCNA)[48] using the WGCNA package in R[49]. We used a soft threshold power of 9 for gene promoters and 8 for gene bodies. We created 100 topological overlap matrices (TOMs) using 100 independent bootstraps of the samples with parameters: type = signed and corFnc = bicor. The 100 TOMs were combined edge-wise by taking the median of each edge across all bootstraps. The consensus TOM was clustered hierarchically using average linkage hierarchical clustering (using 1 – TOM as a dissimilarity measure). The topological overlap dendrogram was used to define modules using the cutreeHybrid() function with parameters: mms = 100, ds = 4, merge threshold of 0.1, and negative pamStage. To identify ASD-associated modules, we fit a linear mixed effect model for each co-methylation module: module eigengene ~ Diagnosis + Age + Sex + Region + Brain bank + Batch + CET as fixed effects, and brainID as a random effect. We ran these analyses separately for ASD Convergent subtype vs control samples and ASD Disparate subtype vs control samples.

Prediction of miRNA target genes

We predicted mRNA target genes for each miRNA using TargetScan v7.2[29]. We downloaded 3' UTR sequences of human genes and miRNA family information from the TargetScan database (http://www.targetscan.org/cgi-bin/targetscan/data_download.vert72.cgi). For miRNAs that were identified in the previous publication[12] and not present in the TargetScan default predictions, we

manually curated their family conservation by visually inspecting the multiz 46-way vertebrate alignment at their genomic locus in the hg19 human assembly of the UCSC genome browser[50]. For each putative miRNA-UTR target site, TargetScan calculates a context++ score which takes into account both evolutionary conservation and targeting efficiency. These context++ scores were weighted based on affected isoform ratios. We took the top weighted context++ score for each unique miRNA-UTR target pair.

To assess enrichment of targets within miRNA co-expression modules, we first filtered for the top 25% miRNAs by connectivity (module hubs) within each module (kME ≥ 0.84, 0.82, and 0.57 for the brown, magenta, and yellow modules, respectively) and identified the strongest targets with a context++ score ≤ -0.05 for these hub genes.

Assignment of H3K27ac regions to cognate gene

We assigned H3K27ac regions within promoter regions (2KB upstream of TSS to TSS) to their proximal gene. For H3K27ac regions that did not lie within a gene promoter, we assigned them to their cognate gene using adult brain eQTL data and Hi-C data from bulk adult brain tissue[33] as well as Hi-C data from sorted NeuN+ and NeuN- cells from adult brain tissue (Synapse accession number: syn10248174 for NeuN-, syn10248215 for NeuN+). For eQTL data, we assigned a H3K27ac region to a gene if the eSNP resided within the H3K27ac peak. For Hi-C data, we assigned a H3K27ac region to a gene if the promoter of that gene physically interacted with a region containing the H3K27ac peak at an FDR < 1%.

Enrichment analyses

We downloaded post-mortem brain single nucleus gene expression data from Hodge et al[24]. The count data were normalized using log2(CPM + 1). To identify markers of neuronal cell types, we ran differential expression analyses for a particular cell cluster against all other clusters when restricting the dataset to inhibitory neurons or excitatory neurons separately. Differential expression analyses were run in R using a linear model: expression ~ cluster membership. For each cluster, we identified markers as those genes with an FDR corrected P-value < 0.05 and a log2(fold change) > 0.75.

We downloaded cell type markers for neurons, astrocytes, oligodendrocytes, endothelial cells, and microglia from Zhang et al[51]. We downloaded microglial cell-type specific markers (fold change >= 1) from Hammond et al[27]. We downloaded markers of microglial activation from Hirbec et al[28]. For module cell type enrichments, enrichments of co-expression vs co-methylation modules, and enrichment with orthogonal gene lists we used logistic regression to test whether gene set 1 ~ gene set 2 using a background set of genes shared between study 1 and study 2.

For expression and methylation gene ontology enrichments, we used the g:Profiler[52] package in R with parameters: correction_method = fdr, max_set_size = 1000, and hier_filtering = moderate. We performed ordered queries with genes ordered by fold change for differential expression and methylation or by connectivity to the module eigengene (kME) for co-expression and co-methylation modules. For acetylation gene ontology enrichments, we used GREAT[30] (http://great.stanford.edu/public/html/index.php) with the default basal plus extension association rule setting.

We defined genic regions using the gencode v19 annotations[42] and downloaded epigenomically defined chromatin states from the Roadmap Epigenomics project for adult brain cortex[19] (https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state). To test for H3K27ac peak enrichment in genic regions or chromatin states, we calculated the enrichment using the formula: [(number of bases in region or state AND H3K27ac peaks) / (number of bases in genome)] / [(number of bases in H3K27ac peaks) / (number of bases in genome) x (number of bases in region or state) / (number of bases in genome)] as previously described[19].

Partitioned heritability

We ran stratified LD-score regression[35] to test for enrichment of common variant heritability from GWAS studies of ASD[4,36], Alzheimer's disease[53], and Inflammatory bowel disease[54] in genomic regions of interest. We downloaded the full baseline model of 53 functional categories (https://github.com/bulik/ldsc/wiki/Partitioned-Heritability) and included them with each calculation of partitioned heritability. For differentially expressed, differentially methylated, and co-expression/co-methylation modules, we defined their genomic regions as each gene body +/- 10 KB. For differentially acetylated regions, we defined the genomic region as each H3K27ac peak +/- 1 KB.

Comparison to transcriptomic signatures of other neuropsychiatric disorders

We corrected mRNA expression data from ASD Disparate subtype and control samples for technical and biological covariates by fitting a linear model for each gene: Expression ~ Diagnosis + Age + Sex + Region + RIN + Brain bank + Sequencing batch + seqStatPC1 + seqStatPC2 + seqStatPC3 + seqStatPC4 + seqStatPC5 and regressing out the effect of all

covariates except Diagnosis. We downloaded differentially expressed genes at an FDR < 5% for ASD, Schizophrenia, Bipolar disorder, Major depressive disorder, Alcoholism, and Inflammatory bowel disease from a cross-disorder analysis of neuropsychiatric disorders[23]. For each disorder, we calculated the first principal component on the corrected expression when restricting to genes differentially expressed in that disorder. We checked for differential loading between ASD Disparate subtype samples and control samples using a two-sided Mann-Whitney U test.

Assessment of transcriptome in other cortical regions

We previously sequenced 87 samples in 4 additional cortical regions (BA4-6, BA7, BA17, BA38) from the individuals in this study[23] (Synapse accession number syn11242290). We mapped sequencing reads onto the hg19 genome using STAR[55] and calculated RNA-seq quality control metrics using PicardTools (http://broadinstitute.github.io/picard/). We quantified gene expression using RSEM[56] with gencode v25 annotations[42]. We corrected the expression data for technical and biological covariates by fitting a linear mixed model for each gene: Expression ~ Region + Batch + Age + Sex + Diagnosis + Ancestry_Genotype + PMI + RIN + picard_rnaseq.PCT_CORRECT_STRAND_READS + picard_rnaseq.PCT_MRNA_BASES + picard_gcbias.AT_DROPOUT + star.multimapped_percent + picard_alignment.PCT_CHIMERAS + star.multimapped_toomany_percent + picard_insert.MEDIAN_INSERT_SIZE + picard_rnaseq.PCT_INTERGENIC_BASES + picard_rnaseq.MEDIAN_5PRIME_BIAS + picard_rnaseq.PCT_UTR_BASES + star.num_ATAC_splices + star.num_GCAG_splices + star.num_splices + star.avg_mapped_read_length + Age_sqd +

picard_alignment.PCT_PF_READS_ALIGNED_sqd + picard_rnaseq.PCT_CORRECT_STRAND_READS_sqd + star.avg_mapped_read_length + star.num_ATAC_splices_sqd + star.num_annotated_splices_sqd + star.num_GCAG_splices_sqd as fixed effects and subject as a random effect. We regressed out the effect of all technical covariates which created an expression dataset containing the effects of only biological covariates (subject, diagnosis, region, age, age squared, sex, and ethnicity).

For each region, we calculated the sample loadings across the first principal component (PC1) of gene expression for the 2591 differentially expressed genes at an FDR < 10% from the idiopathic ASD vs control analysis of Parikshak et al[9]. These PC1 loadings were then transformed into Z-scores.

Acknowledgements

We thank Neelroop Parikshak and Ye Emily Wu for assistance in obtaining datasets, William Pembroke and Damon Polioudakis for assistance with figures, as well as members of the Geschwind lab for stimulating discussions. Funding for this work was provided by grants to D.H.G (NIMH R01MH110927 and R01MH094714), G.R. (NIMH 1F32MH114620), H.W. (NIMH R00MH113823 and NARSAD Young Investigator Award), M.J.G (SFARI Bridge to Independence Award), S.P. (Core funds from A*STAR Singapore), and J.M. (Medical Research Council R005176 and K013807).

Author contributions

G.R. and D.H.G. planned the analyses and wrote the manuscript with assistance from all authors. The analyses were performed by G.R. with help from H.W., M.J.G., J.H., and J.C.W. H.W. processed all Hi-C datasets. C.C.Y.W. and J.M. generated, normalized, and provided guidance on the DNA methylation dataset. W.S. and S.P. generated and provided guidance on the histone acetylation dataset.

Code availability

R code to run SNF clustering and ASD/Control differential analyses is available at https://github.com/dhglab/ASD-Integration-Subtypes-Manuscript.

References

1. Chen, J.A., Penagarikano, O., Belgard, T.G., Swarup, V. & Geschwind, D.H. The emerging picture of autism spectrum disorder: genetics and pathology. *Annu Rev Pathol* **10**, 111-44 (2015).

2. Ramaswami, G. & Geschwind, D.H. Genetics of autism spectrum disorder. *Handb Clin Neurol* **147**, 321-329 (2018).

3. de la Torre-Ubieta, L., Won, H., Stein, J.L. & Geschwind, D.H. Advancing the understanding of autism disease mechanisms through genetics. *Nat Med* **22**, 345-61 (2016).

4. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet* **51**, 431-444 (2019).

5. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-21 (2014).

6. Leppa, V.M. *et al.* Rare Inherited and De Novo CNVs Reveal Complex Contributions to ASD Risk in Multiplex Families. *Am J Hum Genet* **99**, 540-554 (2016).

7. Weiner, D.J. *et al.* Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nat Genet* **49**, 978-985 (2017).

8. 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).

9. Parikshak, N.N. *et al.* Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* **540**, 423-427 (2016).

10. Sun, W. *et al.* Histone Acetylome-wide Association Study of Autism Spectrum Disorder. *Cell* **167**, 1385-1397 e11 (2016).

11. Wong, C.C.Y. *et al.* Genome-wide DNA methylation profiling identifies convergent molecular signatures associated with idiopathic and syndromic autism in post-mortem human brain tissue. *Hum Mol Genet* **28**, 2201-2211 (2019).

12. Wu, Y.E., Parikshak, N.N., Belgard, T.G. & Geschwind, D.H. Genome-wide, integrative analysis implicates microRNA dysregulation in autism spectrum disorder. *Nat Neurosci* **19**, 1463-1476 (2016).

13. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380-4 (2011).

14. Nardone, S., Sams, D.S., Zito, A., Reuveni, E. & Elliott, E. Dysregulation of Cortical Neuron DNA Methylation Profile in Autism Spectrum Disorder. *Cereb Cortex* **27**, 5739-5754 (2017).

15. Zhang, W. *et al.* Integrating genomic, epigenomic, and transcriptomic features reveals modular signatures underlying poor prognosis in ovarian cancer. *Cell Rep* **4**, 542-53 (2013).

16. Woo, H.G. *et al.* Integrative analysis of genomic and epigenomic regulation of the transcriptome in liver cancer. *Nat Commun* **8**, 839 (2017).

17. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* **11**, 333-7 (2014).

18. Landt, S.G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**, 1813-31 (2012).

19. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).

20. Giorgi, F.M., Del Fabbro, C. & Licausi, F. Comparative study of RNA-seq- and microarray-derived coexpression networks in Arabidopsis thaliana. *Bioinformatics* **29**, 717-24 (2013).

21. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414 e24 (2016).

22. Friedman, R.C., Farh, K.K., Burge, C.B. & Bartel, D.P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19**, 92-105 (2009).

23. Gandal, M.J. *et al.* Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359**, 693-697 (2018).

24. Hodge, R.D. *et al.* Conserved cell types with divergent features between human and mouse cortex. *bioRxiv*, 384826 (2018).

25. Velmeshev, D. *et al.* Single-cell genomics identifies cell type-specific molecular changes in autism. *Science* **364**, 685-689 (2019).

26. Ballas, N., Grunseich, C., Lu, D.D., Speh, J.C. & Mandel, G. REST and its corepressors mediate plasticity of neuronal gene chromatin throughout neurogenesis. *Cell* **121**, 645-657 (2005).

27. Hammond, T.R. *et al.* Single-Cell RNA Sequencing of Microglia throughout the Mouse Lifespan and in the Injured Brain Reveals Complex Cell-State Changes. *Immunity* **50**, 253-271 e6 (2019).

28. Hirbec, H. *et al.* The microglial reaction signature revealed by RNAseq from individual mice. *Glia* **66**, 971-986 (2018).

29. Agarwal, V., Bell, G.W., Nam, J.W. & Bartel, D.P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4**(2015).

30. McLean, C.Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501 (2010).

31. Karlic, R., Chung, H.R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A* **107**, 2926-31 (2010).

32. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol Cell* **49**, 825-37 (2013).

33. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**(2018).

34. Azevedo, F.A. *et al.* Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J Comp Neurol* **513**, 532-41 (2009).

35. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-35 (2015).

36. Autism Spectrum Disorders Working Group of The Psychiatric Genomics, C. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a

novel locus at 10q24.32 and a significant overlap with schizophrenia. *Mol Autism* **8**, 21 (2017).

37. Hannon, E., Marzi, S.J., Schalkwyk, L.S. & Mill, J. Genetic risk variants for brain disorders are enriched in cortical H3K27ac domains. *Mol Brain* **12**, 7 (2019).

38. Hansen, K.D., Irizarry, R.A. & Wu, Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13**, 204-16 (2012).

39. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-27 (2007).

40. Pidsley, R. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* **14**, 293 (2013).

41. Guintivano, J., Aryee, M.J. & Kaminsky, Z.A. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics* **8**, 290-302 (2013).

42. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-74 (2012).

43. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. in *ArXiv e-prints* (2013).

44. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

45. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

46. Kuang, D., Yun, S. & Park, H. SymNMF: nonnegative low-rank approximation of a similarity matrix for graph clustering. *Journal of Global Optimization* **62**, 545-574 (2015).

47. Oldham, M.C., Langfelder, P. & Horvath, S. Network methods for describing sample relationships in genomic datasets: application to Huntington's disease. *BMC Syst Biol* **6**, 63 (2012).

48. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* **4**, Article17 (2005).

49. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

50. Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res* (2018).

51. Zhang, Y. *et al.* Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron* **89**, 37-53 (2016).

52. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* **35**, W193-200 (2007).

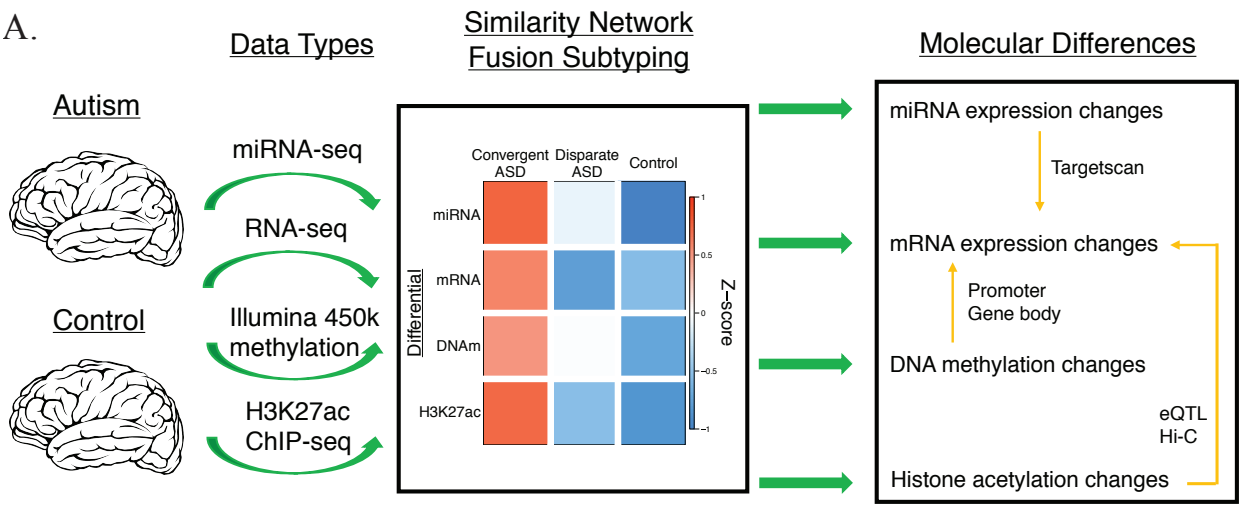
53. Lambert, J.C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* **45**, 1452-8 (2013).

54. Liu, J.Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* **47**, 979-986 (2015).

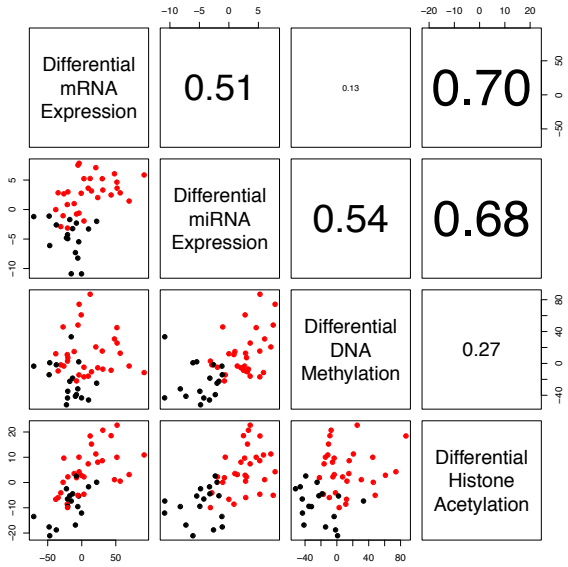
55. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

56. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

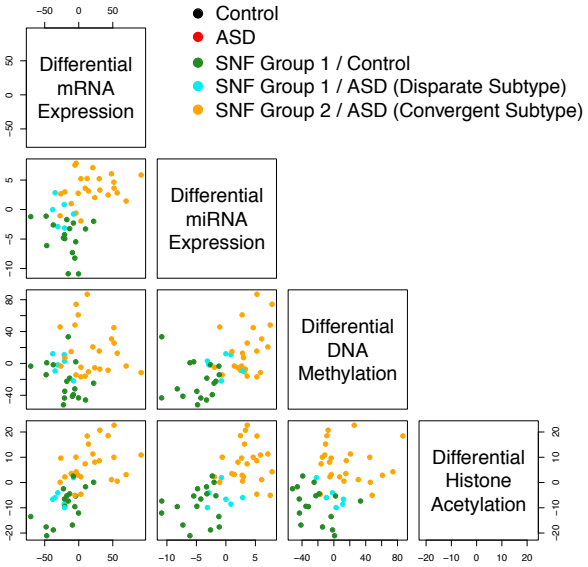
A.



B.



C.



D.

		Sample Numbers	
		Control	ASD
SNF	Group 1	17	6
	Group 2	0	24
LR Classifier	Group 1	53	21
	Group 2	8	40

E.

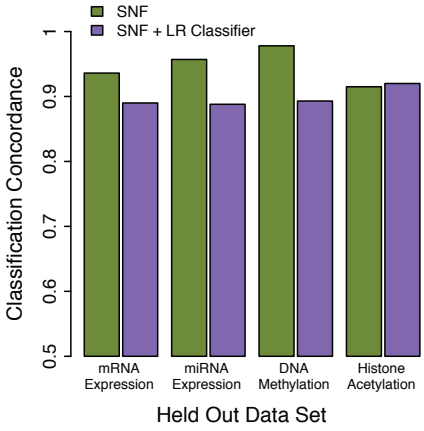


Figure 1. SNF to identify ASD molecular subtypes.

A) Overview of data integration and molecular subtyping to characterize the cascade of molecular changes in ASD.

B) Relationship between sample loadings on the first principal component of differential mRNA expression, miRNA expression, DNA methylation, and histone acetylation.

C) Identification of two sample clusters using SNF: SNF Group 1 and SNF Group 2. ASD samples in SNF Group 1 constitute the Disparate Subtype, whereas ASD samples in SNF Group 2 constitute the Convergent Subtype.

D) Number of samples classified into the two cluster groups using SNF and logistic regression (LR) classifiers.

E) Comparison of SNF clustering and logistic regression (LR) classification assignments when utilizing three out of four datasets (see Supplementary Figure 5). The concordance of sample assignments to those when using the complete dataset are plotted.

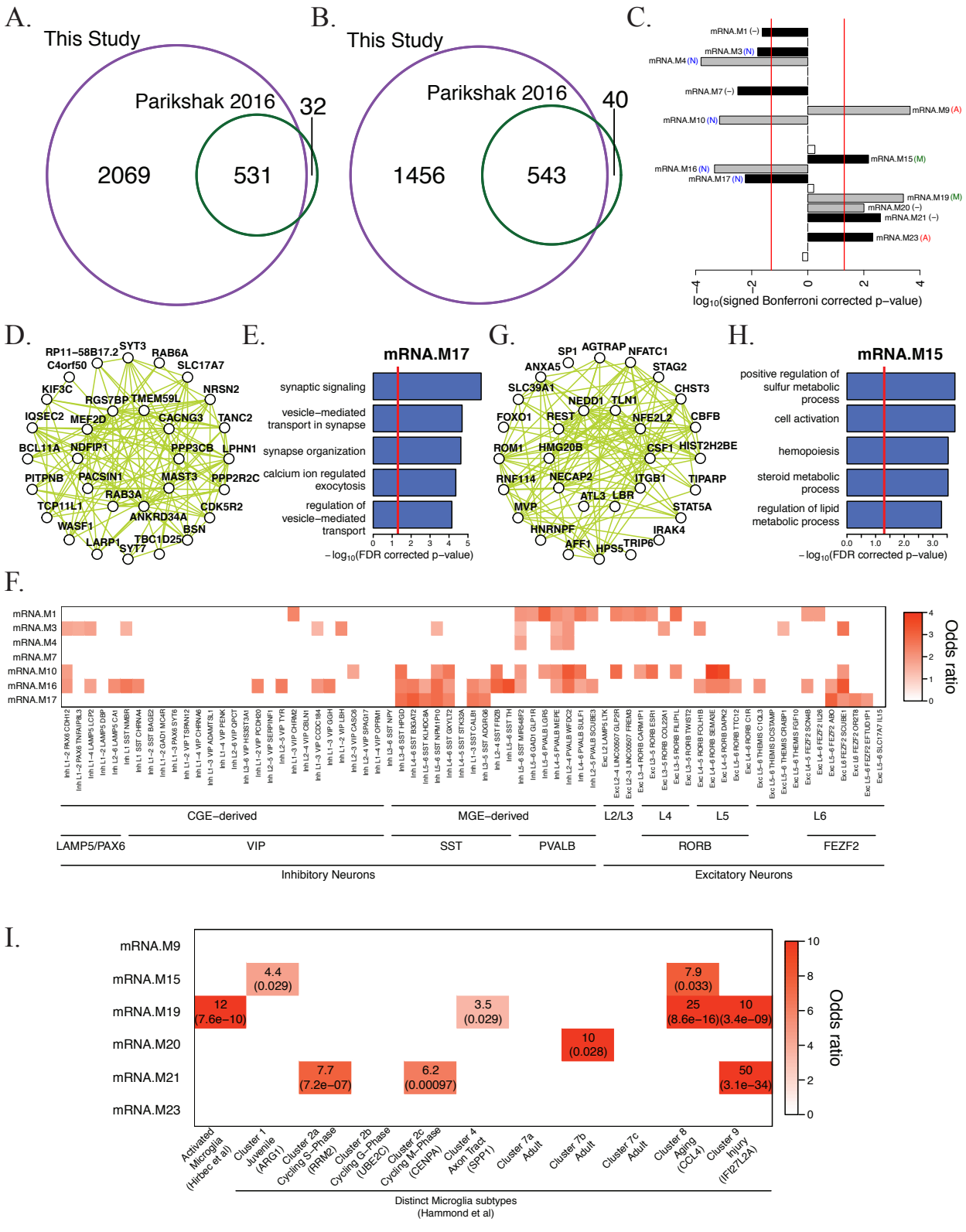


Figure 2. mRNA expression differences in ASD.

A) Overlap in ASD downregulated genes identified in this study against Parikshak et al[9].

B) Overlap in ASD upregulated genes identified in this study against Parikshak et al[9].

C) Signed association of mRNA co-expression module eigengenes with diagnosis (Bonferroni-corrected p-value from an LME model, see Supplementary Figure 7e). Positive values indicate modules with an increased expression in ASD samples. Grey and black bars with labels signify ASD-associated modules identified in Parikshak et al. 2016, and newly identified in this study, respectively. Cell type enrichment for each module is shown in parenthesis: neuron (N), astrocyte (A), microglia (M), and no enrichment (-) (see Supplementary Figure 7g).

D) Top 30 hub genes and 300 connections for co-expression module mRNA.M17.

E) Top gene ontology enrichments for co-expression module mRNA.M17.

F) Enrichment of ASD downregulated neuronal co-expression modules with neuronal cell-type markers identified from single nuclei RNA sequencing[24]. Only enrichments with an FDR corrected p-value < 0.05 are shown.

G) Top 30 hub genes and 300 connections for co-expression module mRNA.M15.

H) Top gene ontology enrichments for co-expression module mRNA.M15.

I) Enrichment of ASD upregulated glial co-expression modules with microglial activated genes[28] and microglial cell-type markers[27]. FDR corrected p-values are shown in parentheses.

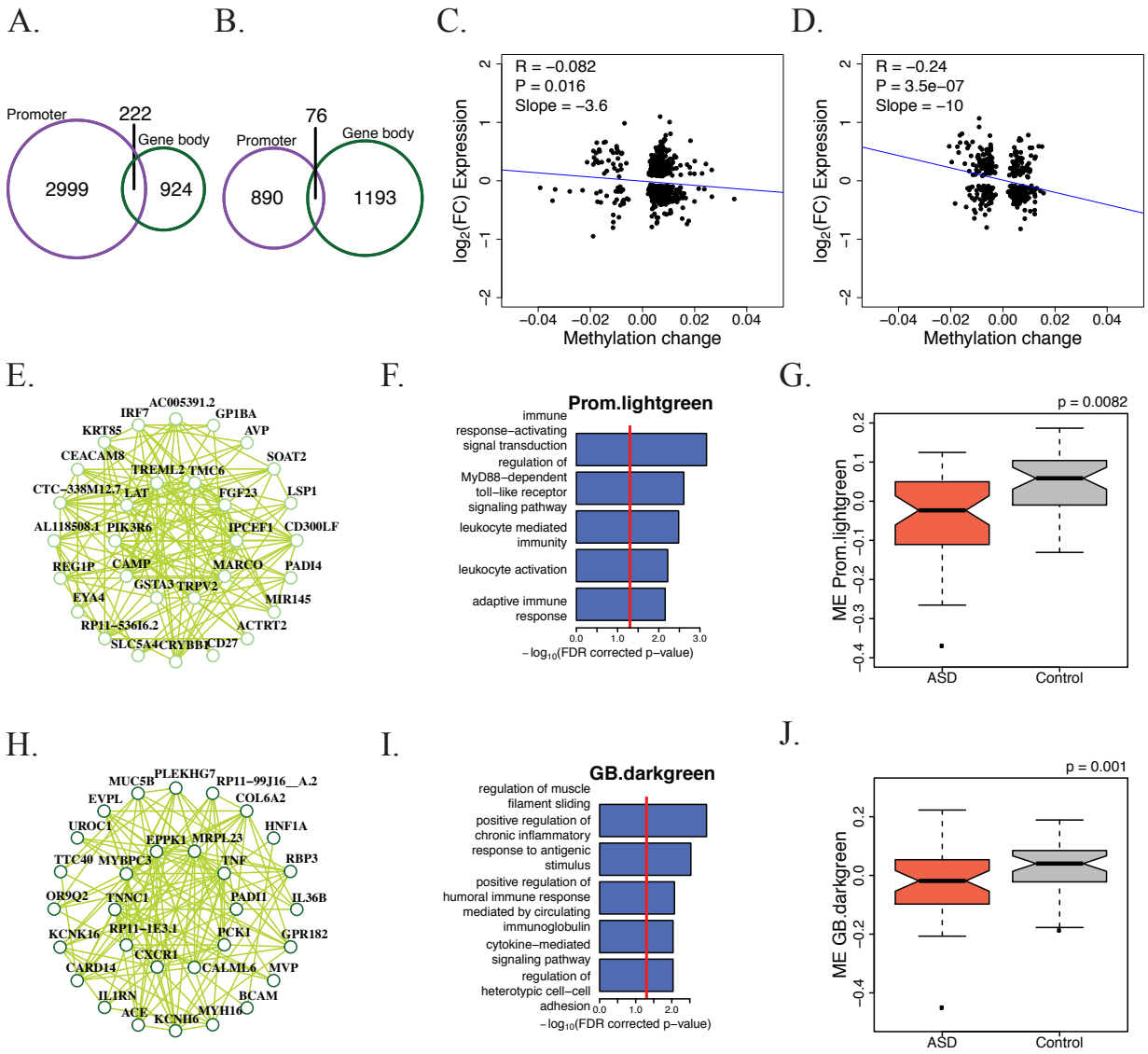


Figure 3. DNA methylation differences in ASD.

A) Overlap in ASD hypermethylated gene promoters and gene bodies.

B) Overlap in ASD hypomethylated gene promoters and gene bodies.

C) Correlation between expression and methylation changes for genes that have differential promoter methylation and are differentially expressed. A linear model was used to correlate differential expression with differential methylation.

D) Correlation between expression and methylation changes for genes that have differential gene body methylation and are differentially expressed. A linear model was used to correlate differential expression with differential methylation.

E) Top 30 hub genes and 300 connections for promoter co-methylation module Prom.lightgreen.

F) Top gene ontology enrichments for promoter co-methylation module Prom.lightgreen.

G) Promoter co-methylation module Prom.lightgreen eigengene values for ASD and control samples. P-value is from a linear mixed effects model (see Supplementary Figure 9f).

H) Top 30 hub genes and 300 connections for gene body co-methylation module GB.darkgreen.

I) Top gene ontology enrichments for gene body co-methylation module GB.darkgreen.

J) Gene body co-methylation module GB.darkgreen eigengene values for ASD and control samples. P-value is from a linear mixed effects model (see Supplementary Figure 10f).

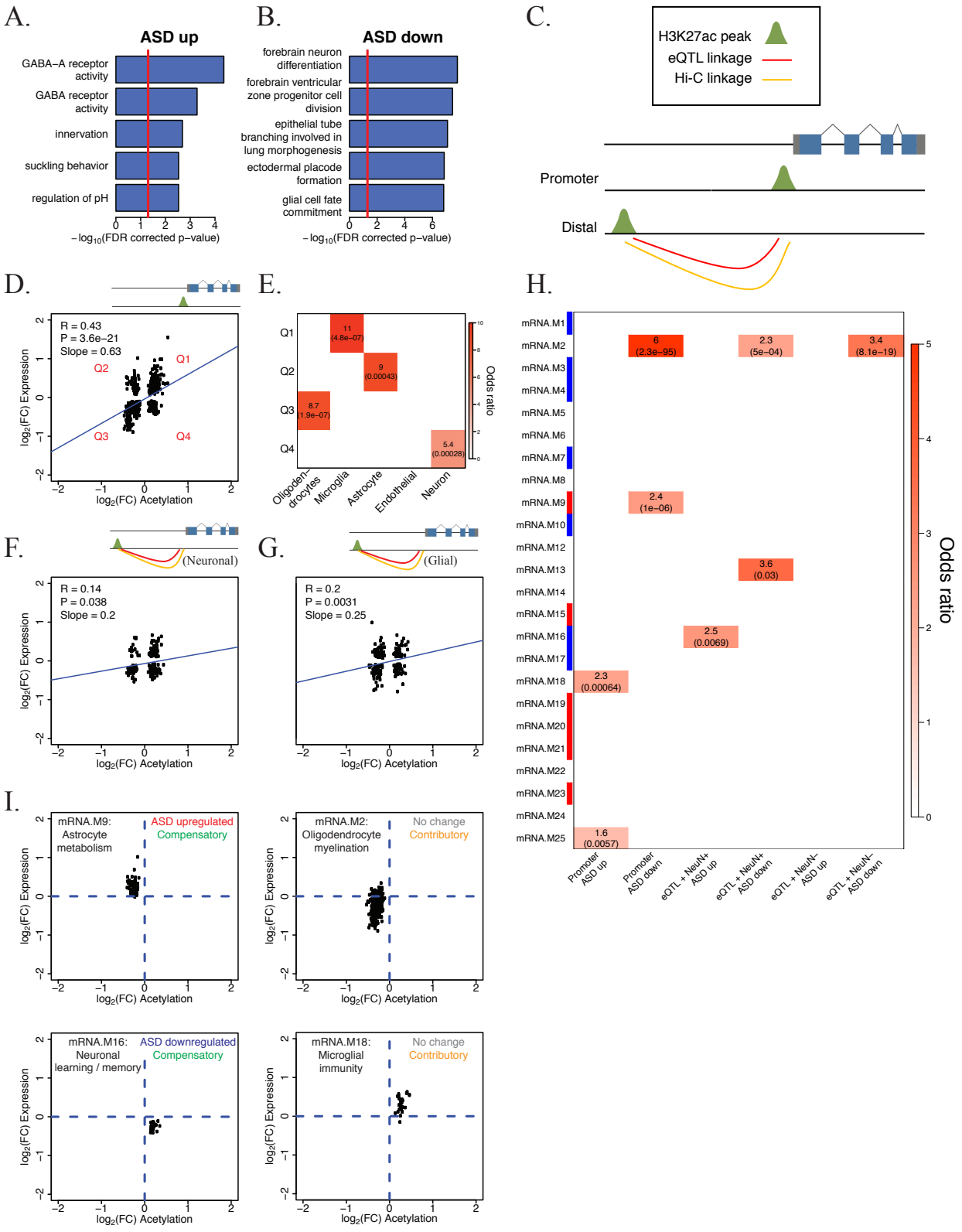


Figure 4. Histone acetylation differences in ASD.

A) Top gene ontology enrichments when linking ASD hyperacetylated regions to proximal genes using GREAT[30].

B) Top gene ontology enrichments when linking ASD hypoacetylated regions to proximal genes using GREAT[30].

C) Schema to link H3K27ac regions with their cognate genes. H3K27ac peaks within promoters were directly assigned to the proximal gene. Distal H3K27ac peaks were assigned to genes using eQTL and Hi-C datasets.

D) Correlation between expression and acetylation changes for genes that have a differentially acetylated region within their promoter and are differentially expressed. A linear model was used to correlate differential expression with differential acetylation. The four separate quadrants are marked.

E) Cell type enrichments for the four quadrants in **D**. FDR corrected p-values are shown in parentheses.

F) Correlation between expression and acetylation changes for differentially acetylated regions linked to differentially expressed genes with both eQTL and neuronal Hi-C evidence. A linear model was used to correlate differential expression with differential acetylation.

G) Correlation between expression and acetylation changes for differentially acetylated regions linked to differentially expressed genes with both eQTL and glial Hi-C evidence. A linear model was used to correlate differential expression with differential acetylation.

H) Enrichment of cognate genes linked to differentially acetylated regions within mRNA co-expression modules. Modules with a significant relationship to diagnosis are marked along the y

axes (red: increased expression in ASD; blue: decreased expression in ASD). FDR corrected p-values are shown in parentheses.

I) Correlation between expression and acetylation changes for differentially acetylated peaks linked to gene co-expression modules. The functional annotation for each module is represented in the top left corner. The association of each module to ASD diagnosis is represented in the top right corner as well as whether acetylation changes are contributory or compensatory to changes in expression.

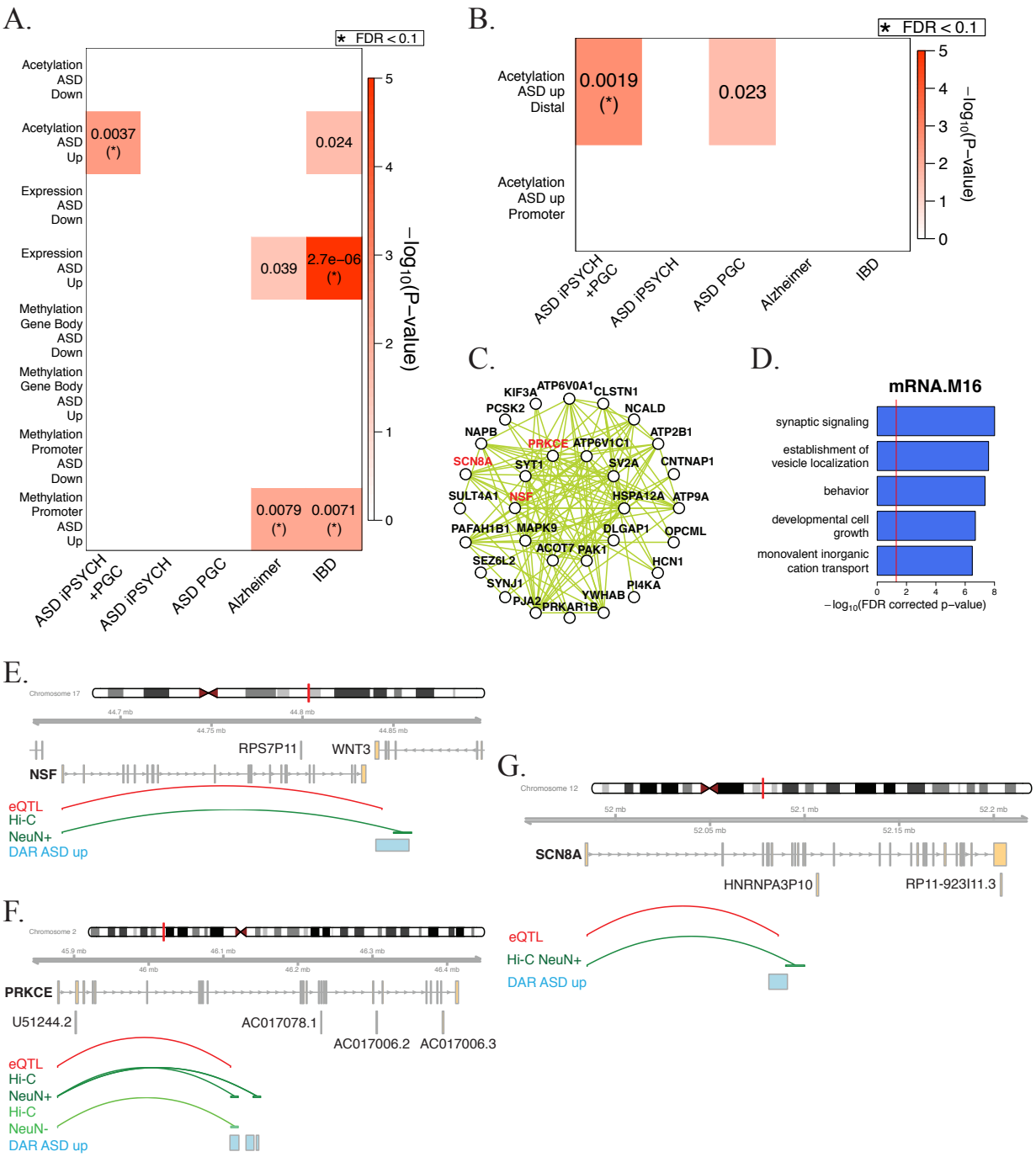


Figure 5. ASD genetic risk variant enrichments.

A) Partitioned heritability enrichments for ASD, Alzheimer's, and IBD GWAS in differentially expression, methylation, or acetylated regions of the genome.

B) Partitioned heritability enrichments for ASD, Alzheimer's, and IBD GWAS in differentially acetylated regions of the genome within, or distal to, gene promoters.

C) Top 30 hub genes and 300 connections for co-expression module mRNA.M16.

D) Top gene ontology enrichments for co-expression module mRNA.M16.

E-G) Genomic region around *NSF* (**E**), *PRKCE* (**F**), and *SCN8A* (**G**). ASD-associated hyperacetylated regions are shown along with eQTL and Hi-C linkages to the gene TSS.

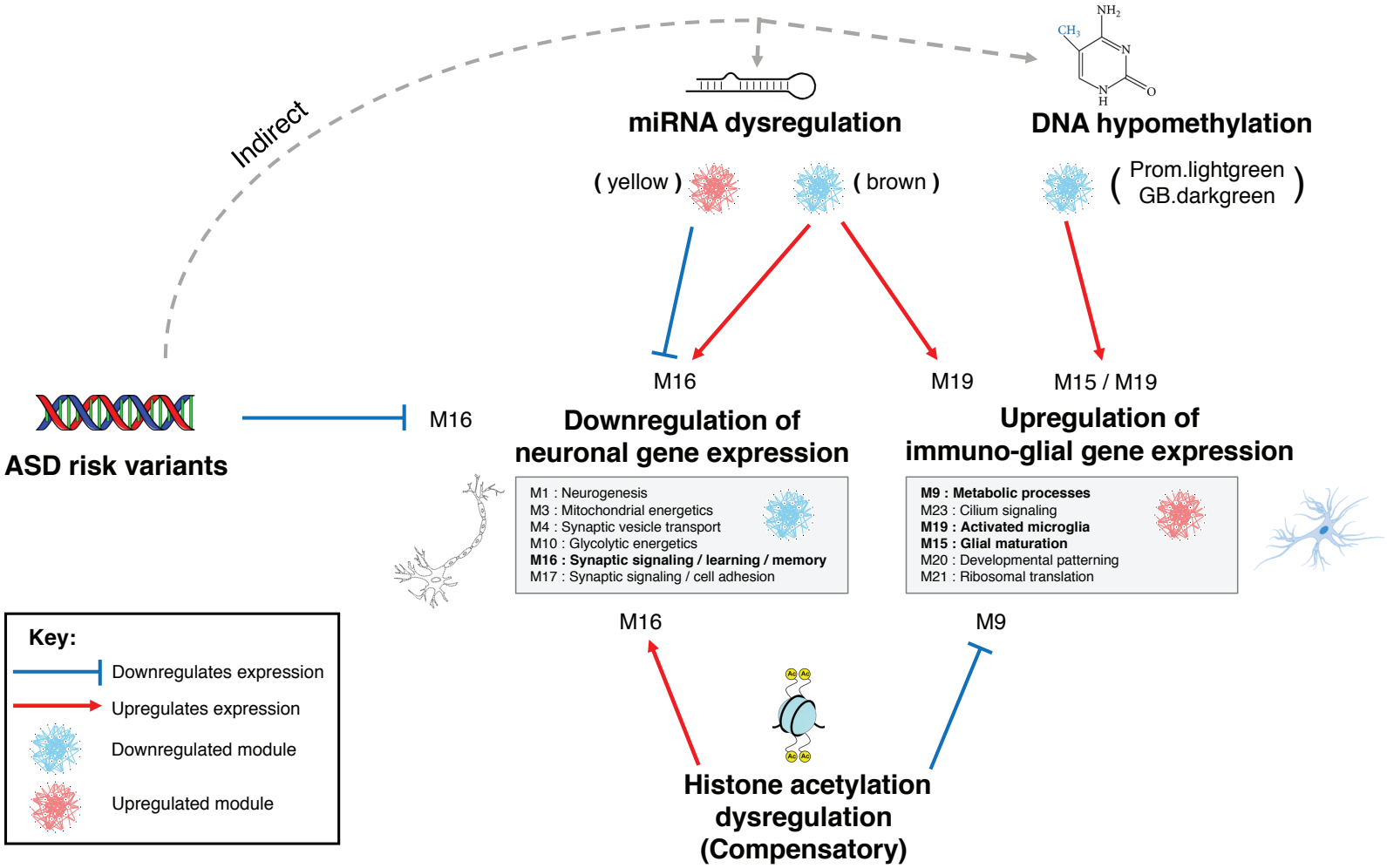

Indirect
miRNA dysregulation
DNA hypomethylation
(yellow)
(brown)
(Prom.lightgreen
GB.darkgreen)
M16
M19
M15 / M19
ASD risk variants
M16
Downregulation of
neuronal gene expression
Upregulation of
immuno-glial gene expression
M1 : Neurogenesis
M3 : Mitochondrial energetics
M4 : Synaptic vesicle transport
M10 : Glycolytic energetics
M16 : Synaptic signaling / learning / memory
M17 : Synaptic signaling / cell adhesion
M9 : Metabolic processes
M23 : Cilium signaling
M19 : Activated microglia
M15 : Glial maturation
M20 : Developmental patterning
M21 : Ribosomal translation
M16
M9
Histone acetylation
dysregulation
(Compensatory)
Key:
Downregulates expression
Upregulates expression
Downregulated module
Upregulated module

Figure 6. Model of molecular dysregulation across genetic, transcriptomic, and epigenomic levels in ASD. Blue bar-headed and red arrows correspond to regulatory mechanisms predicted to decrease and increase gene expression, respectively.