THE UNIVERSITY OF TEXAS AT DALLAS
McDermott Library

TREASURES
@UT Dallas

School of Natural Sciences and Mathematics

Long Genes Linked to Autism Spectrum Disorders Harbor Broad Enhancer-Like Chromatin Domains

UT Dallas Author(s):

Yoon Jung Kim
Tae Hoon Kim

Rights:

CC BY-NC 4.0 (Attribution-NonCommercial)
©2018 The Authors

Citation:

Zhao, Y. -T, D. Y. Kwon, B. S. Johnson, M. Fasolino, et al. 2018. "Long genes linked to autism spectrum disorders harbor broad enhancer-like chromatin domains." Genome Research 29(8): 933-942, doi:10.1101/gr.233775.117

This document is being made freely available by the Eugene McDermott Library of the University of Texas at Dallas with permission of the copyright owner. All rights are reserved under United States copyright law unless specified otherwise.

Long genes linked to autism spectrum disorders harbor broad enhancer-like chromatin domains

Ying-Tao Zhao,[1] Deborah Y. Kwon,[1] Brian S. Johnson,[1] Maria Fasolino,[1] Janine M. Lamonica,[1] Yoon Jung Kim,[2] Boxuan Simen Zhao,[3,4] Chuan He,[3,4] Golnaz Vahedi,[1,5] Tae Hoon Kim,[2] and Zhaolan Zhou[1]

[1]Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA; [2]Department of Biological Sciences, The University of Texas at Dallas, Richardson, Texas 75080, USA; [3]Department of Chemistry, Department of Biochemistry and Molecular Biology, Institute for Biophysical Dynamics, University of Chicago, Chicago, Illinois 60637, USA; [4]Howard Hughes Medical Institute, University of Chicago, Chicago, Illinois 60637, USA; [5]Institute for Immunology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA

Genetic variants associated with autism spectrum disorders (ASDs) are enriched in genes encoding synaptic proteins and chromatin regulators. Although the role of synaptic proteins in ASDs is widely studied, the mechanism by which chromatin regulators contribute to ASD risk remains poorly understood. Upon profiling and analyzing the transcriptional and epigenomic features of genes expressed in the cortex, we uncovered a unique set of long genes that contain broad enhancer-like chromatin domains (BELDs) spanning across their entire gene bodies. Analyses of these BELD genes show that they are highly transcribed with frequent RNA polymerase II (Pol II) initiation and low Pol II pausing, and they exhibit frequent chromatin–chromatin interactions within their gene bodies. These BELD features are conserved from rodents to humans, are enriched in genes involved in synaptic function, and appear post-natally concomitant with synapse development. Importantly, we find that BELD genes are highly implicated in neurodevelopmental disorders, particularly ASDs, and that their expression is preferentially down-regulated in individuals with idiopathic autism. Finally, we find that the transcription of BELD genes is particularly sensitive to alternations in ASD-associated chromatin regulators. These findings suggest that the epigenomic regulation of BELD genes is important for post-natal cortical development and lend support to a model by which mutations in chromatin regulators causally contribute to ASDs by preferentially impairing BELD gene transcription.

[Supplemental material is available for this article.]

Autism spectrum disorders (ASDs) are a heterogeneous group of disorders with a strong genetic component. Thus far, many mutations and genetic variants have been identified and implicated in these disorders (Abrahams et al. 2013; Hu et al. 2014; Chen et al. 2015a; Geschwind and State 2015; Gandal et al. 2018). While many of the identified ASD risk genes encode synaptic proteins that are predominantly expressed in neurons and are critical for synaptic development and function (Geschwind and State 2015), recent human genetic studies have also identified another group of ASD risk genes that are involved in chromatin regulation (Neale et al. 2012; O'Roak et al. 2012; Sanders et al. 2012; Talkowski et al. 2012; De Rubeis et al. 2014; Iossifov et al. 2014; Iwase et al. 2017; Redin et al. 2017). However, the mechanism by which mutations in chromatin regulators contribute to ASD risk is poorly understood.

Chromatin regulators typically include "writer proteins" that establish and maintain epigenetic marks, "eraser proteins" that remove modifications, and "reader proteins" that bind and interpret epigenetic information. Together, these proteins establish, maintain, and interpret epigenetic information that is essential for many biological processes (Allis and Jenuwein 2016). Numerous modifications on histone tails are known to be associated with various genomic features: Histone 3 lysine 4 monomethylation (H3K4me1) marks poised and active enhancers; histone 3 lysine 27 acetylation (H3K27ac) marks regulatory elements such as promoters and active enhancers; histone 3 lysine 79 dimethylation (H3K79me2) occurs preferentially at the 5′ ends of gene bodies of actively transcribed genes; histone 3 lysine 36 trimethylation (H3K36me3) is found preferentially at the 3′ ends of gene bodies of actively transcribed genes; and histone 3 lysine 4 trimethylation (H3K4me3) marks transcriptional start sites (The ENCODE Project Consortium 2012). Notably, in addition to these canonical annotations, recent studies have found the broad domain of H3K4me3 often marks cell identity genes and the expansion of H3K27ac underlies the etiology of midline carcinoma (Alekseyenko et al. 2015; Benayoun et al. 2015; Chen et al. 2015b), indicating different distribution patterns of histone modifications may also play a critical functional role.

Neurons are terminally differentiated post-mitotic cells that carry unique epigenomic and transcriptional features compared with other somatic cells, such as high levels of non-CG DNA methylation (Guo et al. 2014; Zhao et al. 2016), high levels of 5-hydroxymethylcytosine (5hmC) (Kriaucionis and Heintz 2009), and overrepresented expression of long genes >100 kb (Gabel et al. 2015; Zylka et al. 2015; Johnson et al. 2017). Notably, long gene

Corresponding author: zhaolan@pennmedicine.upenn.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.233775.117.

© 2018 Zhao et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

transcription is preferentially impaired in mouse cortical neurons carrying mutations in methyl CpG binding protein 2 (MECP2) (Gabel et al. 2015; Johnson et al. 2017), mutations of which cause the ASD, Rett syndrome (Amir et al. 1999). While the underlying mechanism responsible for this preferential impairment of long gene transcription remains elusive, several lines of evidence indicate that chromatin features may underlie the length-dependent regulation of long gene transcription (King et al. 2013; Gabel et al. 2015; Johnson et al. 2017).

To investigate the functional relationship between chromatin modifications and the transcription of neuronal long genes, we performed chromatin immunoprecipitation followed by sequencing (ChIP-seq) against five different types of histone modifications, whole-genome bisulfite sequencing (WGBS) and Tet-assisted bisulfite sequencing (TAB-seq) for DNA methylation (5mC) and hydroxymethylation (5hmC) status (Yu et al. 2012), and assay for transposase-accessible chromatin using sequencing (ATAC-seq) for regions of accessible chromatin (Buenrostro et al. 2013). To uncover the transcriptional features of neuronal long genes, we also performed global run-on sequencing (GRO-seq) (Core et al. 2008) and nuclear RNA sequencing (RNA-seq) to obtain gene transcriptional status. Our findings support a mechanistic model by which the transcriptional sensitivity of a unique class of long genes, bearing broad enhancer-like chromatin domains (BELDs), to alterations of chromatin regulators underlies the genetic linkage of these chromatin factors to ASD risks, thus uncovering a new avenue to develop therapeutics for ASDs.

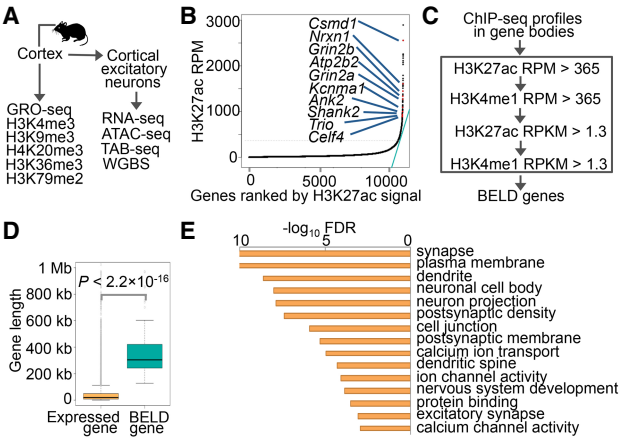

Figure 1. BELD genes in the mouse cortex. (*A*) Diagram of genomic profiling in this study. The RNA-seq, ATAC-seq, TAB-seq, and WGBS were performed from cortical excitatory neurons, while the GRO-seq and ChIP-seq were performed from the cortex. The cortical excitatory neurons and the cortex were obtained from 6-wk-old male mice. (*B*) Distribution of gene body H3K27ac signal (RPM) across expressed genes in the cortex. The green line indicates the tangent line with a slope of one. Ten examples of BELD genes with a known role in ASD risk are highlighted. (*C*) A workflow diagram demonstrating the process used to identify BELD genes. (*D*) Boxplots of gene lengths of expressed genes and BELD genes in the cortex. *P* indicates *P*-value, one-tailed *t*-test. (*E*) The top 15 enriched Gene Ontology terms of the BELD genes.

Results

Epigenomic and transcriptomic profiling in the mouse cortex

To investigate the molecular mechanisms by which chromatin regulators contribute to ASD risk, we carried out a series of high-throughput profiling experiments that focus on transcriptional and epigenomic features of genes expressed in the mouse cortex (Fig. 1A; Supplemental Fig. S1). We performed GRO-seq, nuclear RNA-seq, WGBS, TAB-seq, ATAC-seq, and ChIP-seq against different histone modifications. GRO-seq and all ChIP-seq experiments were performed using whole-cortical tissues, while RNA-seq, WGBS, TAB-seq, and ATAC-seq were performed using cortical excitatory neurons, which account for ~85% of all neurons in the cortex (Johnson et al. 2017). We obtained 3.77 billion reads in total from these profiling experiments, and the data show high quality and reproducibility between replicates (Supplemental Fig. S1). We also incorporated publicly available histone ChIP-seq data sets from cortical excitatory neurons (Mo et al. 2015), such as H3K27ac, H3K4me1, and H3K27me3, in our overall analysis.

Identification of genes harboring BELDs

We next identified all expressed genes from the mouse cortex, defined by GRO-seq reads per million uniquely mapped reads per kilo base (RPKM) >0.5. Given that neuronal genes and genes linked to neurodevelopmental disorders tend to be long in length (Supplemental Fig. S2A,B; Zylka et al. 2015), we divided expressed genes into four groups based on gene length. We then analyzed chromatin features of these genes by focusing on genomic regions surrounding the transcription start site (TSS) and the gene body. We found that chromatin accessibility and histone modifications are similarly distributed at the TSS among long and short genes (Supplemental Fig. S2C). However, in contrast to TSS, long genes show a tendency of enriched levels of H3K27ac, H3K4me1, and ATAC-seq signals in their gene bodies (Supplemental Fig. S2C–E).

To further explore this gene body feature of chromatin modification, we utilized and modified a previously described methodology to identify genomic regions with distinct chromatin modifications (Whyte et al. 2013). We used both H3K27ac and H3K4me1 profiles to identify genes that are enriched with these two histone modifications in their gene bodies. To establish a reads per million uniquely mapped reads (RPM) cutoff value for these two modifications, we first calculated the gene body H3K27ac and H3K4me1 RPM values for all expressed genes, ranked them by gene body H3K27ac signal, and then constructed a tangent line with a slope of one to identify the cutoff value for RPM (Fig. 1B). We found that the ranking of RPM values biases toward long genes. Thus, to avoid the length bias of RPM and to correct for gene length, we also used additional cutoffs of RPKM. By using these criteria for both H3K27ac and H3K4me1 (Fig. 1C), we identified 185 genes with enriched H3K4me1 and H3K27ac signals across their gene bodies, which we termed BELD genes (Supplemental Table S1). Notably, the gene body H3K27ac signals are highly correlated with gene body H3K4me1 signals and ATAC-seq signals (Supplemental Fig. S2F,G).

We found that BELD genes, showing a median length of 303 kb and a mean length of 364 kb, are significantly longer than other genes expressed in the cortex (Fig. 1D). To characterize potential features of BELD genes but exclude any bias due to gene length, we further generated a control set of 185 non-BELD genes with a median length of 262 kb and a mean length of 272 kb, which we termed non-BELD control genes (Supplemental Table S2). For all subsequent analyses, BELD genes were compared directly with non-BELD control genes.

Importantly, by using publicly available ChIP-seq data sets from the mouse cortex (Xie et al. 2012; Nord et al. 2013; Halder et al. 2016), we also observed similar BELD features in the

same group of genes (Supplemental Fig. S3), independently supporting the presence of these features in the neuronal epigenome.

BELD genes are enriched in synaptic functions and are expressed specifically in the brain

To assess the biological functions of BELD genes, we performed Gene Ontology enrichment analysis and found that BELD genes are enriched in synaptic signaling and ion channel activity in the nervous system (Fig. 1E). In contrast, non-BELD control genes are not enriched in any Gene Ontology terms (FDR > 0.05). Given the high enrichment of BELD genes in synaptic functions, we next examined the distribution of BELD genes in the coexpression network of the developing brain (Parikshak et al. 2013). Notably, we found that BELD genes are enriched in module M16 of the coexpression network (Supplemental Fig. S4A), a module that is critical for early synaptic development and also demonstrates the highest enrichment of ASD risk genes (Parikshak et al. 2013). Lastly, to determine whether BELD genes are specifically expressed in the brain, we investigated BELD gene expression patterns among 13 different tissues from the ENCODE Project (Shen et al. 2012) and found that BELD genes are specifically expressed in brain tissues, such as the cortex, cerebellum, and olfactory bulb, but not in other tissues (Supplemental Fig. S4B). Together, these results demonstrate that BELD genes are enriched in functions related to the synapse and synaptic development and are expressed specifically in the brain.

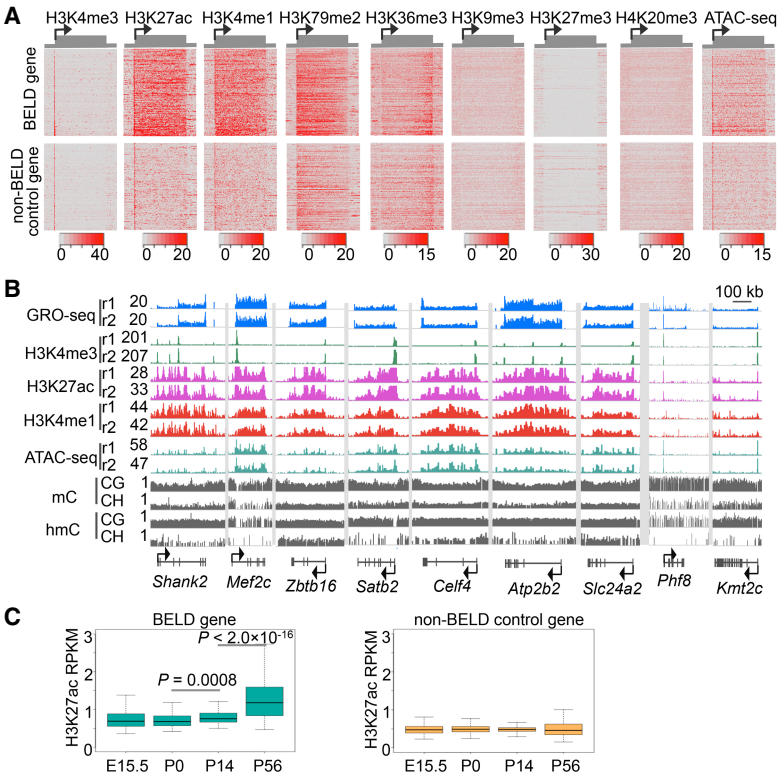

Figure 2. Chromatin modifications in gene bodies of BELD genes. (*A*) Heatmaps of normalized signals (reads per 30 million uniquely mapped reads [RP30M]) of ChIP-seq and ATAC-seq across the gene bodies and surrounding regions (±100 kb) of the BELD genes and non-BELD control genes. (*B*) Browser representations of GRO-seq, ChIP-seq, ATAC-seq, mC, and hmC signals at nine long genes. The first seven genes are BELD genes linked to ASDs, while the last two genes are non-BELD genes. (*C*) Boxplot representations of H3K27ac RPKMs in the gene bodies of BELD genes and non-BELD control genes at four developmental time points. *P* indicates *P*-value, pairwise *t*-test adjusted by the Benjamini-Hochberg method.

Epigenomic features of BELD genes

We found that, compared with non-BELD control genes, BELD genes demonstrate significantly higher levels of H3K4me1 and H3K27ac signals in their gene bodies (Supplemental Fig. S4C,D). Given this unique feature of BELD domains, we further analyzed six additional histone modifications. We found that as with H3K27ac and H3K4me1, H3K79me2 signals are also enriched across the gene bodies of BELD genes compared with non-BELD control genes (Fig. 2A,B). In contrast, the other five types of histone modifications do not show notable enrichment in BELD genes compared with non-BELD control genes (Fig. 2A,B). Notably, ATAC-seq signals are also enriched in BELD genes compared with non-BELD control genes (Supplemental Fig. S4E), suggesting that the chromatin in BELD genes is more accessible than that in non-BELD control genes. Finally, given recent studies that link gene body DNA methylation to gene expression levels (Stroud et al. 2017), we specifically analyzed 5mC and 5hmC profiles at these genes. We found that BELD genes contain lower levels of 5mC in both CG and CH contexts than non-BELD control genes, but their 5hmC levels are comparable to those of non-BELD control genes (Supplemental Fig. S4F). Together, these findings demonstrate that BELD genes are selectively demarcated with gene body chromatin features associated with transcriptional activity, including high levels of H3K27ac, H3K4me1, H3K79me2, and ATAC-seq signals and low levels of genic 5mC.

Given the feature of BELD domains in neuronal long genes and their association with neurodevelopment, we next examined the appearance of BELD domains at different developmental time points. We systematically analyzed the H3K27ac ChIP-seq profiles of forebrain tissues at 10 developmental time points from The ENCODE Project and a recent study (The ENCODE Project Consortium 2012; Stroud et al. 2017). We found that the BELD domains are absent at embryonic time points and post-natal day (P) 0, begin to appear at P14, and are fully established at P56 (Fig. 2C; Supplemental Fig. S4G), suggesting that the BELD domains are established post-natally, concomitant with the timing of synapse development.

BELD domains are associated with high transcriptional activity

Given the known effects of enhancer-related chromatin modifications on gene transcription, we next examined whether the broad enhancer-like domains in the gene body of BELD genes are indicative of distinct transcriptional status. We found that BELD genes are expressed at a significantly higher level than non-BELD control

genes (Fig. 3A). Upon analyzing GRO-seq data to measure transcriptional initiation frequency by calculating polymerase II (Pol II) bindings along the gene bodies and to measure Pol II pausing by comparing Pol II densities between TSS and gene bodies (Core et al. 2008), we found an increased frequency of transcriptional initiation at BELD genes, which is about twofold higher than that of non-BELD control genes (Fig. 3B). We also observed a notable decrease in GRO-seq signals around the TSS of BELD genes (Fig. 3C), as well as significantly lower pausing indices (Fig. 3D), a quantitative measurement of Pol II pausing status (Core et al. 2008). This suggests that there is a reduction in Pol II pausing at the promoters of BELD genes (Fig. 3D; Supplemental Fig. S5A–D). Furthermore, we observed higher levels of Pol II occupancy (Fig. 3E), higher levels of H3K79me2 (Fig. 3F), and increased production of eRNA-like transcripts (Supplemental Fig. S5E,F) in the gene body of BELD genes than non-BELD control genes, indicating higher levels of transcriptional activity in BELD genes that is consistent with higher levels of expression (Fig. 3A). Thus, the BELD domains in these genes are associated with distinct transcriptional features, which likely contribute to the high transcriptional activity of BELD genes.

BELD domains are associated with frequent chromatin–chromatin interactions

To investigate a potential mechanism by which broad enhancer-like domains may facilitate gene transcription, we next analyzed the formation of higher-order chromatin–chromatin interactions in BELD genes. We utilized chromatin contact maps generated by Hi-C profiling in the cortex of 8-wk-old male mice (Shen et al. 2012). We found that BELD genes exhibit higher numbers of chromatin–chromatin interactions than non-BELD control genes (Fig. 4A). To further investigate the special localization of these chromatin–chromatin interactions, we divided the gene bodies of BELD genes and non-BELD control genes into 1000 equal-size bins and examined the average interaction number of each bin. We found that BELD genes demonstrate higher levels of chromatin–chromatin interactions compared with non-BELD control genes across their gene bodies (Fig. 4B). Notably, these chromatin–chromatin interactions in BELD genes tend to form within the gene bodies (Fig. 4C; Supplemental Fig. S5G) in contrast to enhancer–promoter interactions that are typically outside of the enhancers (Supplemental Fig. S5G).

Frequently interacting regions (FIREs) are genomic loci that exhibit a high frequency of local chromatin interactions (Schmitt et al. 2016). We therefore obtained FIREs of mouse cortex (n = 3167, FIRE score > 3) and examined the percentage of BELD genes containing FIREs, finding that the percentages of BELD genes that overlap with FIREs are significantly higher than those of non-BELD control genes (Fig. 4D,E). Furthermore, we quantified the percentage of the gene body region defined as FIREs for both BELD genes and non-BELD control genes, which we found to be significantly higher among BELD than non-BELD control genes (Fig. 4F). Thus, the broad enhancer-like domains appear to associate with local higher-order chromatin–chromatin interactions, suggesting that the gene body of BELD genes is likely organized into a highly connected chromatin unit (Fig. 4B–F; Supplemental Fig. S5G). Together, these results indicate that BELD genes contain high levels of gene body–restricted chromatin–chromatin interactions.

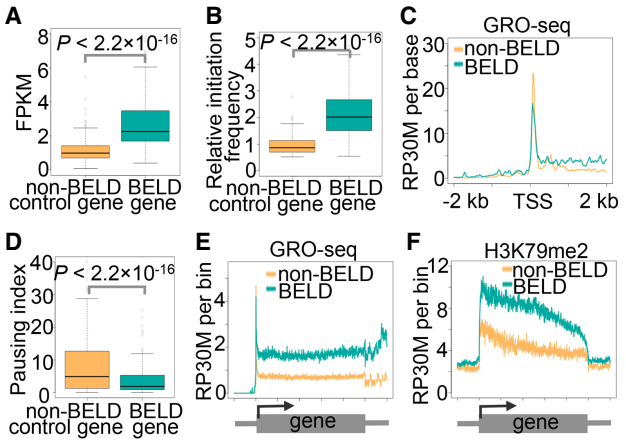

Figure 3. BELD genes display unique transcriptional features. (*A*) Boxplot representation of the expression levels of BELD and non-BELD control genes. *P* indicates *P*-value, one tailed *t*-test. (*B*) Boxplot representation of the relative transcription initiation frequency of non-BELD control genes and BELD genes. The ratio of transcriptional initiation frequency was measured by calculating Pol II bindings along the gene bodies from the GRO-seq data. The average initiation frequency of non-BELD control genes was set to one. *P* indicates *P*-value, one tailed *t*-test. (*C*) Comparison of GRO-seq profile around TSS between non-BELD control genes and BELD genes. (*D*) Boxplot representation of the pausing index of non-BELD control genes and BELD genes. *P* indicates *P*-value, one tailed *t*-test. (*E*) Comparisons of GRO-seq profile between non-BELD control genes and BELD genes in gene body regions and surrounding regions (± 100 kb). (*F*) Comparisons of H3K79me2 ChIP-seq profile between non-BELD control genes and BELD genes in gene body regions and surrounding regions (±100 kb).

BELD genes are enriched in risk genes identified in neurodevelopmental disorders such as ASDs

M16 of the developing brain transcriptome network demonstrates the highest enrichment for ASD risk genes (Parikshak et al. 2013). Given our findings of the enrichment of BELD genes in module M16 (Supplemental Fig. S4A), we next investigated whether BELD genes are associated with risk of neurodevelopmental disorders. We analyzed the enrichment of BELD genes with risk genes identified for ASDs (Werling et al. 2018), developmental disorder (DD) (Deciphering Developmental Disorders Study 2017), intellectual disability (ID) (Lelieveld et al. 2016), and attention deficit hyperactivity disorder (ADHD) (Zhang et al. 2012). We found that BELD genes are significantly and selectively enriched in ASD risk genes (P = 0.0006), relative to DD-associated genes (P = 0.0441) and ADHD risk genes (P = 0.0334), but not in ID risk genes (P = 0.5) (Supplemental Fig. S6A,B). By using the recently reported highly constrained genes identified by de novo loss-of-function mutation analysis and protein truncating variants analysis as targets (Samocha et al. 2014; Lek et al. 2016), we also observed significant enrichment of BELD genes (Supplemental Fig. S6C), consistent with the high fraction of ASD risk genes in these gene lists. In contrast, BELD genes do not show a significant enrichment in genes associated with schizophrenia, diabetes, and height identified by GWAS (Supplemental Fig. S6B; MacArthur et al. 2017). Together, these results demonstrate that BELD genes represent a significantly higher risk for neurodevelopmental disorders, particularly ASDs.

BELD genes are transcriptionally sensitive to the dysfunction of chromatin regulators

Recent studies have reported that ASD risk genes are enriched in genes encoding chromatin regulators in addition to synaptic

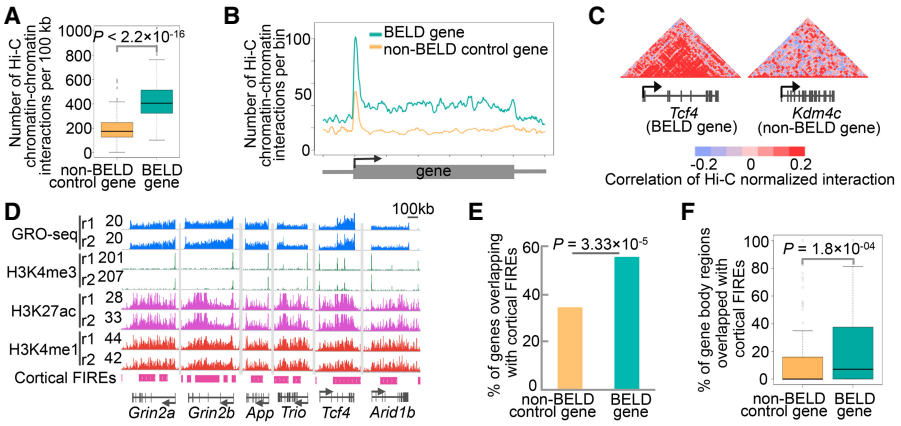

Figure 4. Chromatin–chromatin interactions in BELD genes. (*A*) Boxplots of Hi-C chromatin–chromatin interactions per 100 kb in non-BELD control genes and BELD genes. *P* indicates *P*-value, one-tailed *t*-test. (*B*) Comparisons of Hi-C chromatin–chromatin interactions per bin in the cortex between non-BELD control genes and BELD genes in gene body regions and surrounding regions (±100 kb). The values of *y*-axes are normalized to gene length. (*C*) Heatmaps of correlations of Hi-C normalized interactions in gene body regions and surrounding regions (±100 kb) of *Tcf4* and *Kdm4c*. (*D*) Browser representations of GRO-seq, ChIP-seq, and cortical FIREs profiles at six BELD genes linked to ASD. (*E*) Percentage of genes overlapping with cortical frequently interacting regions (FIREs). *P* indicates *P*-value, Fisher's exact test. (*F*) Boxplot of percentages of gene body regions that overlap with cortical FIREs. *P* indicates *P*-value, one-tailed *t*-test.

proteins (De Rubeis et al. 2014; Iossifov et al. 2014). The high levels of gene body chromatin modifications observed in BELD genes led us to postulate that the high transcriptional activity of BELD genes is sensitive to the disruption of chromatin regulators implicated in ASDs. To test this possibility, we investigated transcriptome changes associated with alternations in chromatin regulators that are genetically implicated in ASDs (Abrahams et al. 2013), such as MECP2, lysine-specific demethylase 6b (KDM6B), DNA topoisomerase 1 (TOP1), chromodomain-helicase-DNA-binding protein 8 (CHD8), and lysine-specific demethylase 5c (KDM5C).

MECP2, KDM6B, and TOP1 are chromatin regulators that are known to facilitate gene transcription (King et al. 2013; Wijayatunge et al. 2014; Johnson et al. 2017). MECP2, mutations of which are responsible for Rett syndrome (Amir et al. 1999), is known to bind methylated DNA and modulate gene transcription (Lyst and Bird 2015). We profiled transcriptional activity in the cortex of the MECP2 R106W–mutant mice that carry a Rett syndrome mutation using GRO-seq and found that BELD genes are significantly more down-regulated than non-BELD control genes (Fig. 5A). KDM6B is a histone lysine demethylase for H3K27 methylation (Mosammaparast and Shi 2010). After analyzing the transcriptome changes in *Kdm6b* knockdown cortical neurons (Wijayatunge et al. 2014), we found that BELD genes are significantly more down-regulated than non-BELD control genes upon *Kdm6b* knockdown (Fig. 5A). TOP1 is a topoisomerase that is implicated in ASDs and can be inhibited by topotecan. When we analyzed the transcriptome changes of topotecan-treated cortical neurons (King et al. 2013), we observed a greater decrease of gene expression for BELD genes, in contrast to non-BELD control genes in response to the treatment (Fig. 5A).

In contrast, CHD8 and KDM5C, an ATP-dependent DNA helicase and a histone H3K4 methylation demethylase, respectively, are chromatin regulators that show repressive activity toward gene transcription and are also linked to ASDs (Mosammaparast and Shi 2010; Abrahams et al. 2013; Ronan et al. 2013; Sugathan et al. 2014). We examined the transcriptome changes in *Chd8* knockdown cortical cells (Durak et al. 2016) and in cortical transcriptome profiles from *Kdm5c* knockout mice (Iwase et al. 2016). We found a greater increase in gene expression for BELD genes relative to non-BELD control genes in both studies (Fig. 5B). Furthermore, upon examination of H3K4me1 ChIP-seq from *Kdm5c* knockout mice (Iwase et al. 2016), we found that H3K4me1 signals are significantly elevated in *Kdm5c* knockout mice selectively for BELD genes than non-BELD control genes as well (Fig. 5C), suggesting that the greater increase in gene expression of BELD genes is associated with the increased H3K4me1 modifications in their gene bodies.

To further corroborate our findings, we next tested the sensitivity of BELD gene expression to the disruption of chromatin regulators in cultured Neuro 2a cells. We first carried out GRO-seq, RNA-seq, H3K4me3, and H3K27ac ChIP-seq in the Neuro 2a cells

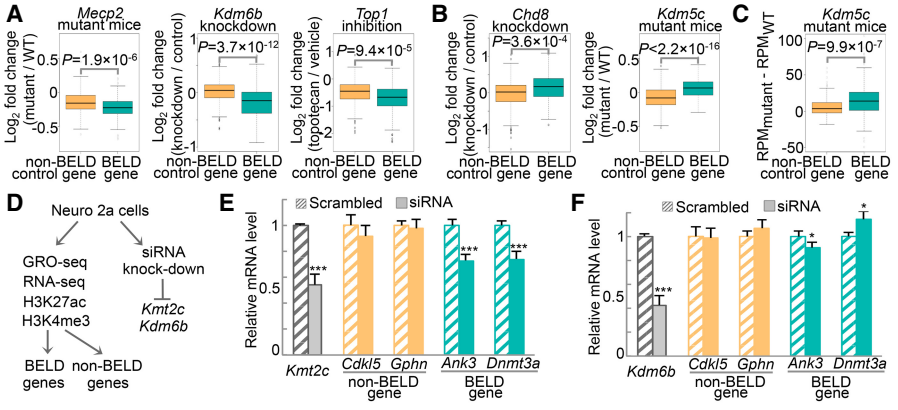

Figure 5. Expression of BELD genes show vulnerability to disruption of ASD-associated chromatin regulators. (*A*) Boxplots of expression changes of non-BELD control genes and BELD genes in response to MECP2 R106W mutation (mouse cortex), *Kdm6b* knockdown (cortical neurons), or TOP1 inhibition by topotecan (cortical neurons). *P* indicates *P*-value, one-tailed *t*-test. (*B*) Boxplots of expression changes of non-BELD control genes and BELD genes in response to *Chd8* knockdown in cortical neurons or *Kdm5c* knockout in cortical tissues. *P* indicates *P*-value, one-tailed *t*-test. (*C*) Boxplots of H3K4me1 changes in gene body regions of non-BELD control genes and BELD genes in the cortex of *Kdm5c* knockout mice. *P* indicates *P*-value, one-tailed *t*-test. (*D*) Diagram of genomic profiling and siRNA knockdown in Neuro 2a cells. (*E*,*F*) The expression changes of two BELD genes and two non-BELD genes in response to *Kmt2c* knockdown (*E*) or to *Kdm6b* knockdown (*F*) in Neuro 2a cells. (***) $P < 3.76 \times 10^{-7}$, one-tailed *t*-test; (*) $P < 8.81 \times 10^{-3}$, one-tailed *t*-test.

and utilized similar methodology to identify BELD genes and non-BELD genes (Fig. 5D; Supplemental Fig. S7A–C). We then used siRNAs to knockdown the expression of two ASD-associated chromatin regulators that facilitate gene transcription, lysine methyltransferase 2C (KMT2C) and KDM6B, and measured their effect on BELD and non-BELD gene expression by RT-PCR. When *Kmt2c* expression levels are reduced by 50%, we found a significant reduction in BELD gene expression that is not observed in non-BELD genes (Fig. 5E; Supplemental Fig. S7D). Similarly, when *Kdm6b* expression levels are reduced by 60%, we found a significant reduction of *Ank3* expression that is not observed in non-BELD genes (Fig. 5F; Supplemental Fig. S7D). Notably, *Dnmt3a* expression is increased in response to *Kdm6b* knockdown (Fig. 5F; Supplemental Fig. S7D), indicating a rather complex mechanism for the regulation of *Dnmt3a* in Neuro 2a cells. Taken together, these results demonstrate that BELD genes are more sensitive to the functional deregulation of ASD-associated chromatin regulators.

BELD genes exist in human brains and are preferentially affected in autistic individuals

To explore whether BELD domains also exist in humans, we analyzed H3K27ac ChIP-seq data sets derived from human prefrontal cortices (Liu et al. 2015; Vermunt et al. 2016). We first identified the human orthologs of the murine BELD and non-BELD control genes, followed by analysis of H3K27ac signals in these orthologs. Notably, we uncovered that BELD domains do present in the human orthologs of the mouse BELD genes (Fig. 6A–C). We next analyzed the gene expression profiles in the prefrontal cortices of 38 control and 25 autistic individuals (Liu et al. 2016). Notably, we found that the expression of the human BELD orthologs is significantly more affected in the autistic individuals than that of non-BELD control genes (Fig. 6D). Together, these results indicate that BELD genes also exist in humans and are dysregulated in individuals with idiopathic autism.

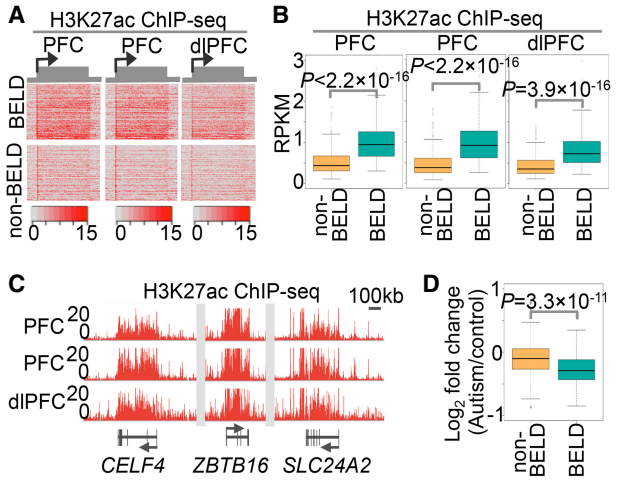

Figure 6. Human orthologs of BELD genes and their expression in autistic individuals. (*A*) Heatmaps of normalized signals of H3K27ac in human orthologs of the mouse BELD (BELD) and non-BELD control genes (non-BELD). (PFC) Prefrontal cortex: (dlPFC) dorsolateral prefrontal cortex. (*B*) Boxplots of H3K27ac RPKMs in gene body regions of non-BELD control genes and BELD genes. *P* indicates *P*-value, one-tailed *t*-test. (*C*) Browser representation of H3K27ac ChIP-seq from human PFC at three BELD genes. (*D*) Boxplots of expression changes of human non-BELD control genes and BELD genes in the PFC of autistic individuals. *P* indicates *P*-value, one-tailed *t*-test.

Discussion

Neurons demonstrate length-dependent transcriptional impairment upon topotecan treatment and MECP2 disruption (King et al. 2013; Gabel et al. 2015; Johnson et al. 2017). While the underlying mechanisms remain unclear, one proposed model is that long genes may harbor distinct chromatin domains and higher-order structures compared with short genes (King et al. 2013), which render them vulnerable to the disruption of chromatin regulators. We found that a unique set of long genes, with a median length of 303 kb, harbors BELDs across their entire gene bodies. These BELD genes also demonstrate specific higher-order chromatin–chromatin interactions that accompany high transcriptional activity. Notably, these BELD genes are transcriptionally sensitive to the disruption of ASD-associated chromatin regulators. Consequently, the transcriptional impairment of BELD genes, which largely encode synaptic proteins, may directly contribute to cellular and functional impairment as observed in ASD individuals who carry genetic defects in chromatin regulators. Notably, long genes without BELD domains do not show this sensitivity, which may explain why different long genes respond differently to topotecan treatment (King et al. 2013).

Gene transcription relies on RNA polymerases that travel along the entire gene body region, making gene length a critical factor for transcription efficiency, especially for long genes. Therefore, long genes may require specific mechanisms to overcome this length constraint in order to maintain comparable and sustained expression levels as short genes. We found that BELD domains in long genes show increased chromatin interactions and higher-order chromatin formations, which may promote the transcriptional cycles of RNA polymerases and then facilitate gene transcription. Profiling RNA polymerase activity and mapping high-resolution chromatin interactions are thus necessary to further illustrate this length-dependent transcriptional regulatory mechanism.

Our analysis showed that BELD genes are enriched in the module M16 of the coexpression network in the developing brain. Module M16 is the earliest up-regulated module during cortical development and likely functions in the development and formation of synaptic structures (Parikshak et al. 2013), indicating BELD genes may play a role in synaptic development. Indeed, BELD domains are absent at P0, begin to appear at P14, and are fully established at P56 in the mouse cortex. Thus, the appearance and establishment of BELD domains are consistent with the timing of cortical synaptogenesis. But, further investigation is required to dissect the functional requirement of BELD domains in this context.

In summary, we report that many of the long genes involved in synaptic development and function harbor BELDs in their gene bodies that correlate with the robust transcription of these genes in the brain. The observation that BELD genes, which are enriched in risk genes for neurodevelopmental disorders, particularly ASDs, are transcriptionally sensitive to the disruption of chromatin regulators implicated in ASDs, highlights an etiological model whereby the disruption in chromatin regulators may contribute to the development of ASDs by preferentially impairing the expression of BELD genes. Our studies reveal the possibility of targeting the transcription of BELD genes as a potential avenue for future therapeutic development for ASDs.

Methods

Animals

All mice used were on the C57BL/6 background, housed in a 12-h light/12-h dark cycle, and fed a standard diet, ad libitum. The Rett syndrome MECP2 R106W–mutant mice were generated as previously reported (Johnson et al. 2017). All experiments were conducted in accordance with the ethical guidelines of the US National Institutes of Health and with the approval of the institutional animal care and use committee of the University of Pennsylvania.

ATAC-seq libraries

Three cortices of 6-wk-old male mice were dissected, and nuclei were isolated for each replicate. Briefly, 1.6 mL of nuclei in PBS was mixed with 372 µL of pelleting buffer by gentle inversion, incubated on ice for 15 min, and centrifuged for 15 min at 5000 rpm at 4°C. Nuclei were resuspended in 60 µL of 1× TD buffer (Illumina catalog no. FC-121-1030) and counted using a hemocytometer; 50,000 nuclei were brought up to 47.5 µL with 1× TD buffer and transposed with 2.5 µL of transposase (Illumina catalog no. FC-121-1030) for 30 min at 37°C. ATAC-seq libraries were constructed as previously described (Buenrostro et al. 2015), with a few exceptions. First, only 2.0 µL of 25 µM customized Nextera PCR primers was used during PCR amplification. Second, the additional number of PCR cycles for each library was kept to three to four cycles. Third, the final amplified library was purified using AMPure XP beads (Beckman A63881) to remove larger fragments (>800 bp) by using 0.5× beads to remove larger fragments, removing the supernatant, adding 1.6× beads to the supernatant, and subsequently following the AMPure XP bead (Beckman A63881) instructions for DNA isolation. Libraries were analyzed on a Bioanalyzer (Agilent) and sequenced on an Illumina HiSeq.

ChIP-seq in the mouse cortex

Cortices of 6-wk-old male mice were dissected. The H3K4me3 ChIP-seq was generated from the pooled cortices of three 6-wk-old male mice; 50–100 µg of chromatin were used per IP. Chromatin was precleared with Protein A Dynabeads (Invitrogen) for 2 h, and an aliquot was saved as input. Immunoprecipitation was performed using 32 µL Protein A Dynabeads and antibody. Chromatin was eluted with elution buffer and reverse crosslinked overnight at 65°C, followed by treatment with RNase A for 30 min at 42°C and proteinase K for 3 h at 55°C. DNA was extracted twice with phenol/chloroform and once with chloroform and ethanol precipitated. Ten nanograms of ChIP DNA was used for library preparation. The antibodies used were H3K36me3 (Abcam ab9050), H3K4me3 (Millipore 07-473), H3K79me2 (Abcam ab3594), H3K9me3 (Abcam ab8898), and H4K20me3 (Abcam ab9053). Libraries were generated using NEB enzymatic reagents (end repair, 5′ adenylation, adapter ligation) with Illumina TruSeq adapters according to Illumina ChIP-seq prep kit manufacturer instructions. Libraries were sequenced on a HiSeq 2500 (50SE).

ChIP-seq in Neuro 2a cells

Neuro 2a cells were grown to confluency on 150-mm culture dishes. Cells were crosslinked directly on the dishes for 10 min at room temperature with 1% formaldehyde, followed by quenching with 0.125 M glycine for 5 min. Cells were scraped, pelleted, and lysed in cell lysis buffer for 10 min on ice. Nuclei were collected and lysed in 10 mM Tris (pH 8.0), 1% SDS, 1 mM EDTA, and 1 mM EGTA. DNA shearing was performed on a Bioruptor instrument. Chromatin was precleared with Protein A Dynabeads (Invitrogen) for 2 h, and an aliquot was saved as input. Immunoprecipitation was performed using 32 µL Protein A Dynabeads and 5 µL of H3K4me3 antibody (Millipore; 07-473) or 10 µg H3K27ac antibody (Abcam; ab4729). Chromatin was eluted with elution buffer and reverse crosslinked overnight at 65°C, followed by treatment with RNase A for 30 min at 42°C and proteinase K for 3 h at 55°C. DNA was extracted twice with phenol/chloroform and once with chloroform and ethanol precipitated; 10 ng of ChIP DNA was used for library preparation.

GRO-seq

Nuclei were isolated from the fresh cortex tissue of the 6-wk-old WT and MECP2 R106W–mutant male mice and from the Neuro 2a cells. GRO-seq libraries were generated as previously described (Greer et al. 2015).

TAB-seq and WGBS

Cortices of 6-wk-old male mice were dissected and nuclei were isolated. Genomic DNA was isolated from nuclei of cortical excitatory neurons using an AllPrep DNA/RNA mini kit (Qiagen catalog no. 80204) and subsequently treated with RNase (Roche catalog no. 11119915001). TAB-seq and WGBS libraries were generated from 400 and 300 ng of DNA, respectively, as previously reported (Yu et al. 2012).

Nuclear total RNA-seq in the mouse cortex

Cortices of 6-wk-old male mice were dissected and nuclei were isolated. Nuclear total RNAs were isolated from nuclei of cortical excitatory neurons and RNA-seq libraries were generated as previously reported (Johnson et al. 2017).

RNA-seq in Neuro 2a cells

RNA-seq was performed in Neuro 2a cells using two replicates. RNA was extracted with TRIzol reagent (Invitrogen) and purified using the RNeasy MinElute clean-up kit (Qiagen catalog no. 74204); 2.5 µg of purified RNA was used for library construction using the TruSeq stranded mRNA library prep kit (Illumina RS-122-2101). Indexed libraries were sequenced at the University of Pennsylvania Next-Generation Sequencing Core on a HiSeq 2500 (Illumina).

siRNA knockdown

Predesigned dicer-substrate siRNAs (DsiRNAs) to *Kmt2c* were purchased from Integrated DNA Technologies (mm.Ri.Kmt2c.13). DsiRNAs to *Kdm6b* were purchased from Integrated DNA Technologies (hs.Ri.KDM6B.13.2). Neuro 2a cells were transfected with each DsiRNA to a concentration of 5 nM for 48 h using Lipofectamine 2000 (Life Technologies, no. 11668019) according to the manufacturer's specifications and then retransfected with another 5 nM for another 24 h. Cells were harvested 72 h after the initial transfection.

Quantitative RT-PCR

RNA was extracted with TRIzol reagent (Invitrogen) and purified using the RNeasy MinElute clean-up kit (Qiagen catalog no. 74204). One thousand nanograms of RNA was converted into cDNA using a high-capacity cDNA reverse transcription kit (Applied Biosystems, 4368814), and real-time PCR was performed using TaqMan gene expression assay probes purchased from Applied Biosystems and TaqMan universal PCR master mix (Applied Biosystems, catalog no. 4304437). The following

TaqMan assay primer/probe sets were used for this study: *Gapdh* (Mm99999915_g1), *Hprt* (Mm03024075_m1), *Ank3* (Mm00464776_m1), *Dnmt3a* (Mm00432881_m1), *Cdkl5* (Mm01156815_m1), *Gphn* (Mm00556895_m1), *Kdm6b* (Mm01332680_m1), and *Kmt2c* (Mm01156942_ m1). Results were quantified on an ABI 7900 system. All RNA expression levels were normalized to *Gapdh* or *Hprt*.

Genome annotation

The mouse genome annotation file of Ensembl release 75 was used. Genes annotated as protein_coding, lincRNA, miRNA, snRNA, and snoRNA were included in the analyses.

GRO-seq and RNA-seq mapping and comparison

The raw FASTQ files were mapped to the mouse mm10 genome by STAR (Dobin et al. 2013) using the parameters of "--outFilterMultimapNmax 1 --outFilterMismatchNmax 3." The edgeR (Robinson et al. 2010) was used to perform the comparison (Supplemental Fig. S8).

ATAC-seq mapping

The mate1 and mate2 files of the paired-end sequencing were mapped separately to the mm10 genome by Bowtie (Langmead et al. 2009) using "-v 2 -m 1." If both mates of a read pair were mapped to the opposite strands of the same chromosome and the distance in between was <2 kb, the read pair was included for further analysis.

ChIP-seq mapping and peak calling

The ChIP-seq FASTQ files were mapped to the mouse mm10 genome by Bowtie using "-v 2 -m 1." Only uniquely mapped reads were included for further analysis. The HOMER (Heinz et al. 2010) findPeaks function was used to call peaks from the H3K4me3 data sets using the parameters of "-style histone -minDist 2500."

TAB-seq and WGBS mapping and analysis

TAB-seq and WGBS mapping were performed as previously described (Zhao et al. 2016). Briefly, Trim Galore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) was used to remove the adapter contamination from the raw sequencing reads using "-stringency 2 --length 36." The trimmed reads were mapped to mm10 genome by Bismark (Krueger and Andrews 2011) using parameters "-n 2 -l 40." Clonal reads were excluded to avoid PCR artifacts. Methylation calling was performed by binomial distribution followed by the Benjamini-Hochberg correction. The mC level for a cytosine (C) was calculated as the WGBS methylation levels subtracted the hmC level in that C.

Pausing index calculation

The pausing index of a gene was calculated as the ratio of GRO-seq signal density in the proximal promoter region versus the signal density in the gene body region. The GRO-seq signal density was calculated as the number of reads mapped to the region divided by the region length.

BELD gene identification

The methodology to identify BELD genes was modified from a previous method (Whyte et al. 2013). Briefly, the RPM cutoff was identified by the intersection point between the ranking curve and the tangent line. The RPKM cutoff was defined as the median levels of gene body H3K27ac RPKM values of expressed genes. For the cortex, BELD genes were identified by the criteria of H3K4me1 RPM > 365 and H3K27ac RPM > 365 and H3K4me1 RPKM > 1.3 and H3K27ac RPKM > 1.3. For the Neuro 2a cells, BELD genes were identified by the criteria of H3K27ac RPM > 430 and H3K27ac RPKM > 2.5.

Non-BELD gene identification

The non-BELD genes were identified by the criteria of H3K4me1 RPM < 365 and H3K27ac RPM < 365 and H3K4me1 RPKM < 1.3 and H3K27ac RPKM < 1.3. The 185 length-matched non-BELD control genes were selected from all non-BELD genes based on their length distribution. For the Neuro 2a cells, non-BELD genes were identified by the criteria of H3K27ac RPM < 430 and H3K27ac RPKM < 2.5.

eRNA-like transcript identification

The identification of intragenic enhancer RNA-like transcripts was performed as previously described (Meng et al. 2014). HOMER was used to perform the do novo transcript identification using the parameters of "findPeaks -style groseq." Transcripts overlapped with H3K4me3 peaks were excluded from downstream analysis to remove the possible divergent promoter transcripts. Transcripts that located in the gene antisense strand and did not overlap with H3K4me3 peaks were defined as the eRNA-like transcripts.

Hi-C data analysis

The raw files of Hi-C data of the mouse cortex (Shen et al. 2012) were download from NCBI Gene Expression Omnibus (GEO) database and were converted into FASTQ files using fastq-dump.2.5.7 (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software). The FASTQ files were mapped to the mouse mm10 genome by HiCUP (Wingett et al. 2015). The mapped SAM files were then converted into the HOMER Hi-C summary format files using hicup2homer (Heinz et al. 2010). The HOMER findHiCInteractionsByChr.pl (Heinz et al. 2010) was used to identify the significant chromatin–chromatin interactions using the parameters of "-res 2000 -superRes 10000."

Metagene analysis (heatmaps and line plots)

The total numbers of uniquely mapped reads or read-pairs for all sequencing data were normalized to 30 million. The signal represents the RP30M per base pair or per bin. For bin analysis, the gene body, upstream, and downstream regions were divided into 1000 equal bins for each gene, respectively, and the number of sequencing reads for each bin was calculated and normalized (Supplemental Fig. S9). The mean value for each bin was calculated as the sum of signals of all bases in the bin divided by the size of that bin. For the heatmaps, the normalized bin signals were used to generate the heatmaps. For the line plots, the median value of the normalized bin signals of a given bin of that gene group was used to generate the plots.

Relative transcription initiation frequency

The Pol II density and activity were obtained from the intragenic GRO-seq profiles of the 1000 bins. The total initiated Pol II was composed of the paused Pol II, the elongating Pol II, and the terminated Pol II. For a given time period (t) and a given gene, the initiated Pol II equals to the initiation frequency ($F_{(i)}$) multiplying the time period (t), which is $F_{(i)}{}^{*}t$; the paused Pol II equals to the GRO-seq signals ($S_{(i)}$) in intragenic bin 1 multiplying the time period ($\sum_{i=1}^{1} S_{(i)}{}^{*}t$); the elongating Pol II equals to the GRO-seq signals in

intragenic from bin 2 to bin 1000 multiplying the time period ($\sum_{i=2}^{1000} S_{(i)} * t$); the terminated Pol II equals to the GRO-seq signals in intragenic bin 1000 multiplying the time period ($\sum_{i=1000}^{1000} S_{(i)} * t$). The equation for the initiation frequency was calculated as follows: $F_{(i)} = (\sum_{i=1}^{1} S_{(i)} * t + \sum_{i=2}^{1000} S_{(i)} * t + \sum_{i=1000}^{1000} S_{(i)} * t)/t$. The relative transcription initiation frequency was calculated as the ratio of the frequency of a given gene versus the average initiation frequency of non-BELD control genes.

Publicly available data sets

The accession numbers of publicly available data sets used in this study are listed in Supplemental Table S3.

Other bioinformatics analyses

All the bioinformatics analyses were done using in-house Perl programs (see "Data access"). All plotting and statistical analyses were performed in R (R Core Team 2015). Heatmaps were generated by heatmap.2 function in the gplots R package (R Core Team 2015). Gene Ontology enrichment analysis was done using DAVID (Huang da et al. 2009), and all expressed genes in the cortex were used as the background. IGV (Robinson et al. 2011) was used to visualize all the sequencing tracks. The conversion of mouse BELD and non-BELD control genes into human orthologs were performed by Ensembl BioMart (Kinsella et al. 2011).

Data access

The raw sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo) under accession numbers of GSE104576. All scripts used in this study are shown in Supplemental Figures S8, S9 and are also accessible via GitHub (https://github.com/Jerry-Zhao/BELD-long-genes).

Acknowledgments

This work was supported by NIH R01MH091850 and R01NS081054 to Z.Z. D.Y.K. is supported by the T32 Training Program in Neurodevelopmental Disabilities (T32NS007413). B.S.J. is supported by a Cell and Molecular Biology Training Grant (TG32-GM072290) and the UNCF/Merck Graduate Research Dissertation Fellowship. Z.Z. is a Pew Scholar in Biomedical Sciences.

References

Abrahams BS, Arking DE, Campbell DB, Mefford HC, Morrow EM, Weiss LA, Menashe I, Wadkins T, Banerjee-Basu S, Packer A. 2013. SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol Autism* **4:** 36.

Alekseyenko AA, Walsh EM, Wang X, Grayson AR, Hsi PT, Kharchenko PV, Kuroda MI, French CA. 2015. The oncogenic BRD4-NUT chromatin regulator drives aberrant transcription within large topological domains. *Genes Dev* **29:** 1507–1523.

Allis CD, Jenuwein T. 2016. The molecular hallmarks of epigenetic control. *Nat Rev Genet* **17:** 487–500.

Amir RE, Van den Veyver IB, Wan M, Tran CQ, Francke U, Zoghbi HY. 1999. Rett syndrome is caused by mutations in X-linked *MECP2*, encoding methyl-CpG-binding protein 2. *Nat Genet* **23:** 185–188.

Benayoun BA, Pollina EA, Ucar D, Mahmoudi S, Karra K, Wong ED, Devarajan K, Daugherty AC, Kundaje AB, Mancini E, et al. 2015. H3K4me3 breadth is linked to cell identity and transcriptional consistency. *Cell* **163:** 1281–1286.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10:** 1213–1218.

Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. 2015. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol* **109:** 21.29.1-9.

Chen JA, Penagarikano O, Belgard TG, Swarup V, Geschwind DH. 2015a. The emerging picture of autism spectrum disorder: genetics and pathology. *Annu Rev Pathol* **10:** 111–144.

Chen K, Chen Z, Wu D, Zhang L, Lin X, Su J, Rodriguez B, Xi Y, Xia Z, Chen X, et al. 2015b. Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. *Nat Genet* **47:** 1149–1157.

Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322:** 1845–1848.

De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE, Kou Y, Liu L, Fromer M, Walker S, et al. 2014. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515:** 209–215.

Deciphering Developmental Disorders Study. 2017. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542:** 433–438.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21.

Durak O, Gao F, Kaeser-Woo YJ, Rueda R, Martorell AJ, Nott A, Liu CY, Watson LA, Tsai LH. 2016. Chd8 mediates cortical neurogenesis via transcriptional regulation of cell cycle and Wnt signaling. *Nat Neurosci* **19:** 1477–1488.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.

Gabel HW, Kinde B, Stroud H, Gilbert CS, Harmin DA, Kastan NR, Hemberg M, Ebert DH, Greenberg ME. 2015. Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature* **522:** 89–93.

Gandal MJ, Haney JR, Parikshak NN, Leppa V, Ramaswami G, Hartl C, Schork AJ, Appadurai V, Buil A, Werge TM, et al. 2018. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359:** 693–697.

Geschwind DH, State MW. 2015. Gene hunting in autism spectrum disorder: on the path to precision medicine. *Lancet Neurol* **14:** 1109–1120.

Greer CB, Tanaka Y, Kim YJ, Xie P, Zhang MQ, Park IH, Kim TH. 2015. Histone deacetylases positively regulate transcription through the elongation machinery. *Cell Rep* **13:** 1444–1455.

Guo JU, Su Y, Shin JH, Shin J, Li H, Xie B, Zhong C, Hu S, Le T, Fan G, et al. 2014. Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat Neurosci* **17:** 215–222.

Halder R, Hennion M, Vidal RO, Shomroni O, Rahman RU, Rajput A, Centeno TP, van Bebber F, Capece V, Garcia Vizcaino JC, et al. 2016. DNA methylation changes in plasticity genes accompany the formation and maintenance of memory. *Nat Neurosci* **19:** 102–110.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38:** 576–589.

Hu WF, Chahrour MH, Walsh CA. 2014. The diverse genetic landscape of neurodevelopmental disorders. *Annu Rev Genomics Hum Genet* **15:** 195–213.

Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4:** 44–57.

Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, Stessman HA, Witherspoon KT, Vives L, Patterson KE, et al. 2014. The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature* **515:** 216–221.

Iwase S, Brookes E, Agarwal S, Badeaux AI, Ito H, Vallianatos CN, Tomassy GS, Kasza T, Lin G, Thompson A, et al. 2016. A mouse model of X-linked intellectual disability associated with impaired removal of histone methylation. *Cell Rep* **14:** 1000–1009.

Iwase S, Berube NG, Zhou Z, Kasri NN, Battaglioli E, Scandaglia M, Barco A. 2017. Epigenetic etiology of intellectual disability. *J Neurosci* **37:** 10773–10782.

Johnson BS, Zhao YT, Fasolino M, Lamonica JM, Kim YJ, Georgakilas G, Wood KH, Bu D, Cui Y, Goffin D, et al. 2017. Biotin tagging of MeCP2 in mice reveals contextual insights into the Rett syndrome transcriptome. *Nat Med* **23:** 1203–1214.

King IF, Yandava CN, Mabb AM, Hsiao JS, Huang HS, Pearson BL, Calabrese JM, Starmer J, Parker JS, Magnuson T, et al. 2013. Topoisomerases facilitate transcription of long genes linked to autism. *Nature* **501:** 58–62.

Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, et al. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)* **2011:** bar030.

Kriaucionis S, Heintz N. 2009. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324:** 929–930.

Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27:** 1571–1572.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536:** 285–291.

Lelieveld SH, Reijnders MR, Pfundt R, Yntema HG, Kamsteeg EJ, de Vries P, de Vries BB, Willemsen MH, Kleefstra T, Lohner K, et al. 2016. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat Neurosci* **19:** 1194–1196.

Liu F, Hon GC, Villa GR, Turner KM, Ikegami S, Yang H, Ye Z, Li B, Kuan S, Lee AY, et al. 2015. EGFR mutation promotes glioblastoma through epigenome and transcription factor network remodeling. *Mol Cell* **60:** 307–318.

Liu X, Han D, Somel M, Jiang X, Hu H, Guijarro P, Zhang N, Mitchell A, Halene T, Ely JJ, et al. 2016. Disruption of an evolutionarily novel synaptic expression pattern in autism. *PLoS Biol* **14:** e1002558.

Lyst MJ, Bird A. 2015. Rett syndrome: a complex disorder with simple roots. *Nat Rev Genet* **16:** 261–275.

MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, et al. 2017. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45:** D896–D901.

Meng FL, Du Z, Federation A, Hu J, Wang Q, Kieffer-Kwon KR, Meyers RM, Amor C, Wasserman CR, Neuberg D, et al. 2014. Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. *Cell* **159:** 1538–1548.

Mo A, Mukamel EA, Davis FP, Luo C, Henry GL, Picard S, Urich MA, Nery JR, Sejnowski TJ, Lister R, et al. 2015. Epigenomic signatures of neuronal diversity in the mammalian brain. *Neuron* **86:** 1369–1384.

Mosammaparast N, Shi Y. 2010. Reversal of histone methylation: biochemical and molecular mechanisms of histone demethylases. *Annu Rev Biochem* **79:** 155–179.

Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V, et al. 2012. Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485:** 242–245.

Nord AS, Blow MJ, Attanasio C, Akiyama JA, Holt A, Hosseini R, Phouanenavong S, Plajzer-Frick I, Shoukry M, Afzal V, et al. 2013. Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* **155:** 1521–1531.

O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD, et al. 2012. Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485:** 246–250.

Parikshak NN, Luo R, Zhang A, Won H, Lowe JK, Chandran V, Horvath S, Geschwind DH. 2013. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155:** 1008–1021.

R Core Team. 2015. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Redin C, Brand H, Collins RL, Kammin T, Mitchell E, Hodge JC, Hanscom C, Pillalamarri V, Seabra CM, Abbott MA, et al. 2017. The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nat Genet* **49:** 36–45.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29:** 24–26.

Ronan JL, Wu W, Crabtree GR. 2013. From neural development to cognition: unexpected roles for chromatin. *Nat Rev Genet* **14:** 347–359.

Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnstrom K, Mallick S, Kirby A, et al. 2014. A framework for the interpretation of *de novo* mutation in human disease. *Nat Genet* **46:** 944–950.

Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, et al. 2012. *De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485:** 237–241.

Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, Li Y, Lin S, Lin Y, Barr CL, et al. 2016. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep* **17:** 2042–2059.

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, et al. 2012. A map of the *cis*-regulatory sequences in the mouse genome. *Nature* **488:** 116–120.

Stroud H, Su SC, Hrvatin S, Greben AW, Renthal W, Boxer LD, Nagy MA, Hochbaum DR, Kinde B, Gabel HW, et al. 2017. Early-life gene expression in neurons modulates lasting epigenetic states. *Cell* **171:** 1151–1164.e16.

Sugathan A, Biagioli M, Golzio C, Erdin S, Blumenthal I, Manavalan P, Ragavendran A, Brand H, Lucente D, Miles J, et al. 2014. *CHD8* regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. *Proc Natl Acad Sci* **111:** E4468–E4477.

Talkowski ME, Rosenfeld JA, Blumenthal I, Pillalamarri V, Chiang C, Heilbut A, Ernst C, Hanscom C, Rossin E, Lindgren AM, et al. 2012. Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell* **149:** 525–537.

Vermunt MW, Tan SC, Castelijns B, Geeven G, Reinink P, de Bruijn E, Kondova I, Persengiev S; Netherlands Brain Bank, Bontrop R, et al. 2016. Epigenomic annotation of gene regulatory alterations during evolution of the primate brain. *Nat Neurosci* **19:** 494–503.

Werling DM, Brand H, An JY, Stone MR, Zhu L, Glessner JT, Collins RL, Dong S, Layer RM, Markenscoff-Papadimitriou E, et al. 2018. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat Genet* **50:** 727–736.

Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153:** 307–319.

Wijayatunge R, Chen LF, Cha YM, Zannas AS, Frank CL, West AE. 2014. The histone lysine demethylase Kdm6b is required for activity-dependent preconditioning of hippocampal neuronal survival. *Mol Cell Neurosci* **61:** 187–200.

Wingett S, Ewels P, Furlan-Magaril M, Nagano T, Schoenfelder S, Fraser P, Andrews S. 2015. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* **4:** 1310.

Xie W, Barr CL, Kim A, Yue F, Lee AY, Eubanks J, Dempster EL, Ren B. 2012. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148:** 816–831.

Yu M, Hon GC, Szulwach KE, Song CX, Zhang L, Kim A, Li X, Dai Q, Shen Y, Park B, et al. 2012. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149:** 1368–1380.

Zhang L, Chang S, Li Z, Zhang K, Du Y, Ott J, Wang J. 2012. ADHDgene: a genetic database for attention deficit hyperactivity disorder. *Nucleic Acids Res* **40:** D1003–D1009.

Zhao YT, Fasolino M, Zhou Z. 2016. Locus- and cell type-specific epigenetic switching during cellular differentiation in mammals. *Front Biol (Beijing)* **11:** 311–322.

Zylka MJ, Simon JM, Philpot BD. 2015. Gene length matters in neurons. *Neuron* **86:** 353–355.

Received December 18, 2017; accepted in revised form May 29, 2018.

Long genes linked to autism spectrum disorders harbor broad enhancer-like chromatin domains

Ying-Tao Zhao, Deborah Y. Kwon, Brian S. Johnson, et al.

Genome Res. 2018 28: 933-942 originally published online May 30, 2018
Access the most recent version at doi:10.1101/gr.233775.117

Supplemental Material http://genome.cshlp.org/content/suppl/2018/06/14/gr.233775.117.DC1

References This article cites 64 articles, 16 of which can be accessed free at:
http://genome.cshlp.org/content/28/7/933.full.html#ref-list-1

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

© 2018 Zhao et al.; Published by Cold Spring Harbor Laboratory Press