Molecular Psychiatry
https://doi.org/10.1038/s41380-018-0049-x

ARTICLE

An integrative analysis of non-coding regulatory DNA variations associated with autism spectrum disorder

Sarah M. Williams[1,2] • Joon Yong An[2,3,4] • Janette Edson[2] • Michelle Watts[2] • Valentine Murigneux[1] • Andrew J. O. Whitehouse[5,6] • Colin J. Jackson[7] • Mark A. Bellgrove[8] • Alexandre S. Cristino[1] • Charles Claudianos[2,9]

Received: 15 October 2016 / Revised: 16 January 2018 / Accepted: 19 February 2018

Abstract

A number of genetic studies have identified rare protein-coding DNA variations associated with autism spectrum disorder (ASD), a neurodevelopmental disorder with significant genetic etiology and heterogeneity. In contrast, the contributions of functional, regulatory genetic variations that occur in the extensive non-protein-coding regions of the genome remain poorly understood. Here we developed a genome-wide analysis to identify the rare single nucleotide variants (SNVs) that occur in non-coding regions and determined the regulatory function and evolutionary conservation of these variants. Using publicly available datasets and computational predictions, we identified SNVs within putative regulatory regions in promoters, transcription factor binding sites, and microRNA genes and their target sites. Overall, we found that the regulatory variants in ASD cases were enriched in ASD-risk genes and genes involved in fetal neurodevelopment. As with previously reported coding mutations, we found an enrichment of the regulatory variants associated with dysregulation of neurodevelopmental and synaptic signaling pathways. Among these were several rare inherited SNVs found in the mature sequence of microRNAs predicted to affect the regulation of ASD-risk genes. We show a paternally inherited miR-873-5p variant with altered binding affinity for several risk-genes including *NRXN2* and *CNTNAP2* putatively overlay maternally inherited loss-of-function coding variations in *NRXN1* and *CNTNAP2* to likely increase the genetic liability in an idiopathic ASD case. Our analysis pipeline provides a new resource for identifying loss-of-function regulatory DNA variations that may contribute to the genetic etiology of complex disorders.

Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder with heterogeneous genetic origins. Recent genome sequencing studies have identified many risk genes from the loss-of-function protein-coding variants, which has

These authors contributed equally: Sarah M. Williams, Joon Yong An

These authors jointly supervised this work: Alexandre S. Cristino, Charles Claudianos

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41380-018-0049-x) contains supplementary material, which is available to authorized users.

✉ Alexandre S. Cristino
a.cristino@uq.edu.au

✉ Charles Claudianos
charles.claudianos@anu.edu.au

[1] University of Queensland Diamantina Institute, University of Queensland, Brisbane, Australia

[2] Queensland Brain Institute, University of Queensland, Brisbane, Australia

[3] Department of Psychiatry, University of California San Francisco, San Francisco, USA

[4] Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, USA

[5] Telethon Kids Institute, University of Western Australia, Perth, Australia

[6] Cooperative Research Centre for Living with Autism, Brisbane, Australia

[7] Research School of Chemistry, Australian National University, Canberra, Australia

[8] Monash Institute of Cognitive and Clinical Neuroscience, Monash University, Melbourne, Australia

[9] Centre for Mental Health Research CMHR, Australian National University, Canberra, Australia

Published online: 27 April 2018

driven a move toward analysis of convergent risk pathways [1–4]. However, less is known about the contribution to risk by the variants in non-coding regulatory regions, which hold the potential to disrupt the finely tuned biological pathways involved in brain development, as demonstrated by a recent study in the influence of the 3′ untranslated regions (3′UTR)-regulatory variants in language impairment [5]. Nevertheless, progress is being made by using whole-genome sequencing of the ASD families. In a recent landmark-autism study, non-coding de novo DNA variations were found to be enriched in the untranslated regions of genes, gene boundaries, and DNase I hypersensitive regions [6]. Similarly, a number of previous exome-sequencing studies including our own [7] have shown that there are many DNA variations that occur in the non-coding regulatory regions of genes, albeit of unknown functional significance.

The functional impact of the non-coding variants has been difficult to interpret due to the lack of specific knowledge concerning their contribution to the regulatory regions that control and modulate gene transcription. Even in cases where a variant in a non-coding region is found to be associated with a disorder through linkage analyses, it is a challenge to assess whether such variations have gene regulatory functions, or whether the risk lies elsewhere in the linkage region [8]. The problem is compounded with the de novo mutations, which are unlikely to be represented in public databases, despite their significant contribution to ASD [6, 7, 9–11]. Much of this difficulty arises from the lack of data integration and functional sequence analysis surrounding transcription control, the molecular interactions, and transactions, of which are poorly understood when compared to the classic set of rules (also known as genetic code) by which genetic information encodes the protein sequences. In this regard, the identification and contribution of the non-coding regulatory variants to ASD-associated biological pathways remains unknown.

The availability of large, genome-wide regulatory resources such as FANTOM5 [12], ENCODE [13], and the BrainSpan developmental brain tissue expression datasets [14] have been instrumental in developing modern-systems-biology-approaches to evaluate the regulatory loss-of-function mutations associated with human disorders. Moreover, these novel approaches involve predicting whether a variant has the potential to change the dynamics of gene networks by disrupting the interactions between key regulatory molecules, such as miRNAs and transcription factors, and their target functional pathways [8, 15, 16]. There is now an increased focus on ASD-associated mutations within the regulatory regions, particularly near-known ASD-risk genes [17, 18], with studies examining the enhancer regions [19] and microRNA targeting [5]. Together these data provide a vision of how we might assess the potential impact of non-coding regulatory DNA variations associated with ASD.

In this study, we developed a systems-based analysis to identify the non-coding regulatory variants using the previously published whole-exome sequencing (WES) data of the ASD families [7], and focusing on the microRNA genes and the gene-proximal cis-regulatory variants. Single nucleotide variants (SNVs) in the regulatory regions were filtered by rarity and functional score. We compiled a genome-wide resource of regulatory regions covering microRNA genes and their putative target sites, transcription start sites, and transcription-factor-binding sites (TFBS) in promoters, 5′ and 3′ UTRs, to evaluate the relationship of these regulatory SNVs to the synaptic pathways. We further characterized these regulatory mutations using ASD-related databases and functional network analyses to show that rare heterozygous loss-of-function coding and non-coding variations combine as the probable causal elements. We specifically highlight the functional significance of a SNV found in microRNA-873, which affects target binding affinity and regulation of the key ASD-risk genes.

Materials and methods

Annotation and filtering of whole-exome sequencing data

The WES data used in this study was from our previous study of Australian ASD families ($n = 128$; 48 ASD cases, and 80 parents) [7], and we considered the single nucleotide variants (SNVs) for this analysis as there were no large copy number variations (CNVs) detected in this sample (data not shown). The participants were recruited from the Western Australian Autism Biological Registry at the Telethon Kids Institute in Perth, Australia [20] and the blood samples were collected and analyzed with the approval by the Ethics Committee's at the Princess Margaret Hospital for Children, University of Western Australia (1845/EP) and the University of Queensland (2012000269). Informed written consent was obtained from next of kin, caretakers, or guardians on behalf of the minors/children enrolled in our study. In this study, we have mapped our WES data (NimbleGen SeqCap version 3) to the hg19-reference genome with the BWA-MEM algorithm [21]. Duplicate reads were removed using Picard tools v2.60. Indel realignment and base quality score recalibration were performed using GATK v3.5 [22]. The germline SNVs and indels were called using the GATK HaplotypeCaller tool and only high quality SNVs were selected for further analysis (RMSMappingQuality, $MQ > 40$; QualByDepth, $QD > 2$). We further selected only SNVs at genomic positions covered by more than 20 reads in at least 90% of the

128 samples. Annotations were made using the ANNOVAR package (2017Jul16) [23] with additional databases (dbSNP150, gnomAD exome and genome collections, CADD version 1.3). The Genome Aggregation Database (gnomAD) was used to obtain minor allele frequencies (MAF) for all quality-passed SNVs (NFE or Non-Finnish European children were used as background population) [24].

As per the previous analysis of the coding variants in the same ASD cohort [7], we used a two-stage filtering process to identify the most interesting SNVs. First a Minor Allele Frequency filter (MAF < 0.01) was applied to select the rare variants, which were thought to have a larger effect in complex disorders such as ASD [7, 25, 26], then a 'functional' score (CADD score ≥ 15) [27] was used to prioritize the potential loss-of-function SNVs.

Computational analysis of the regulatory regions

We defined the regulatory SNV set as those situated in promoters, transcriptional start sites, enhancers, microRNA genes, putative microRNA regulatory elements (MREs) on the 3′UTRs, and TFBS on the promoters, 5′UTRs and 3′UTRs of the protein-coding genes.

Promoter regions were defined using FANTOM5 phase 1 TSS (transcription start site) CAGE data (Cap Analysis of Gene Expression) as regions 1000 bp upstream and 300 bp downstream of the annotated CAGE peaks for Human Genome Nomenclature Consortium (HGNC) gene records [12]. The enhancer region coordinates were downloaded from the FANTOM5 phase 1 database [28]. Genomic annotations of the non-coding regions are from the University of California Santa Cruz (UCSC) 'knownGene' annotation (downloaded 21 February 2013) on the hg19-human reference genome including all transcript variants.

For TFBS prediction, we built a database of 1289 DNA motifs collected from different studies: the FANTOM5 project [12], JASPAR [29], Cristino, Williams [30], and Jolma and Yan [31], as well as data from several stem-cell-focused transcription factor studies [32–40]. We searched for instances of these DNA motifs (at least 70% similarity) on 3′UTRs, 5′UTRs, and promoter regions of genes, using the TAMO package (Version 21 March 2012) [41].

We predicted MREs across 3′UTRs for all human microRNAs (miRBase version 19) [42]. The microRNA-target sites predicted by miRanda [43], (default parameters, free energy ≤ −18), having a 75% overlap with RNAhybrid [44] (-b 2000, -e −18, -s 3utr_human, then filtered free energy ≤ −25) were used for further analysis.

We also searched for SNVs falling within the microRNA-mature-sequence regions themselves. These variants were filtered by population frequency (MAF < 0.01) and CADD score (CADD ≥ 15), and SNVs interpreted on the basis of the mature sequence (as annotated in miRbase19) and seed regions (2–8nt at 5′-end of mature sequence). To identify the conserved microRNA match motifs as a proxy for seed regions, we ran MEME (v4.9.0, options: -mod zoops -nmotifs 1-maxw 15-maxsize 10,000,000) over each microRNA's predicted MREs [45]. Regions of motifs with low information density were trimmed out with TAMO (trimming threshold 0.2) [41]. To estimate the effects of the microRNA gene variants in terms of loss or gain of targets, we used miRanda [43] (default parameters, free energy < = −18) to predict the target sites for microRNA sequences incorporating their respective variants across all known 3′UTR sequences available in the TargetScan database v7.0 [46].

Enrichment analyses of the regulatory SNVs

To test for a general association of the regulatory SNVs with ASD, enrichment was calculated against several ASD-associated gene sets from different sources; the Simons Foundation Autism Research Initiative (SFARI) AutDB database (December 2015 update, Scores indicating any support: S,1–4) [47], our previously developed ASD protein-protein interaction network (AXAS-ASD), which uses 'seed' genes from public resources and their first-degree neighbors [30] and Module 13 (M13-Li2014)—a module of protein–protein interactions from Li, Shi [48]—which is associated with ASD risk. In addition three control datasets were retrieved from the GWAS catalog (https://www.ebi.ac.uk/gwas): (1) coronary artery disease (GWAS-CAD), (2) Crohn's disease (GWAS-Crohn), (3) and autism spectrum disorders (GWAS-ASD).

We also calculated enrichment of the regulatory SNVs (via hypergeometric distribution one-tailed test in R) in co-expression modules constructed and characterized by Parikshak et al. [2].

Biotinylated microRNA-mRNA pulldown transcriptome analysis

Biotinylated microRNA-mRNA pulldown experiments were used to identify bound mRNA transcripts of the miR-873-5p wild-type (WT) and the variant miR-873-5p (Mut). Briefly, biotinylated microRNAs were transfected into the undifferentiated SH-SY5Y cells (a human neuroblastoma cell line), and captured along with their targeted transcripts to magnetic streptavidin beads (Invitrogen Dynabeads M-280).

The pulldown protocol was primarily drawn from Wani and Cloonan [49], as well as Cristino and Barchuk [50], with the following modifications. The samples were transfected in a 75 cm [2] flask with 560 pmol of biotinylated microRNA with lipofectamine 2000 (Thermo Fisher), and

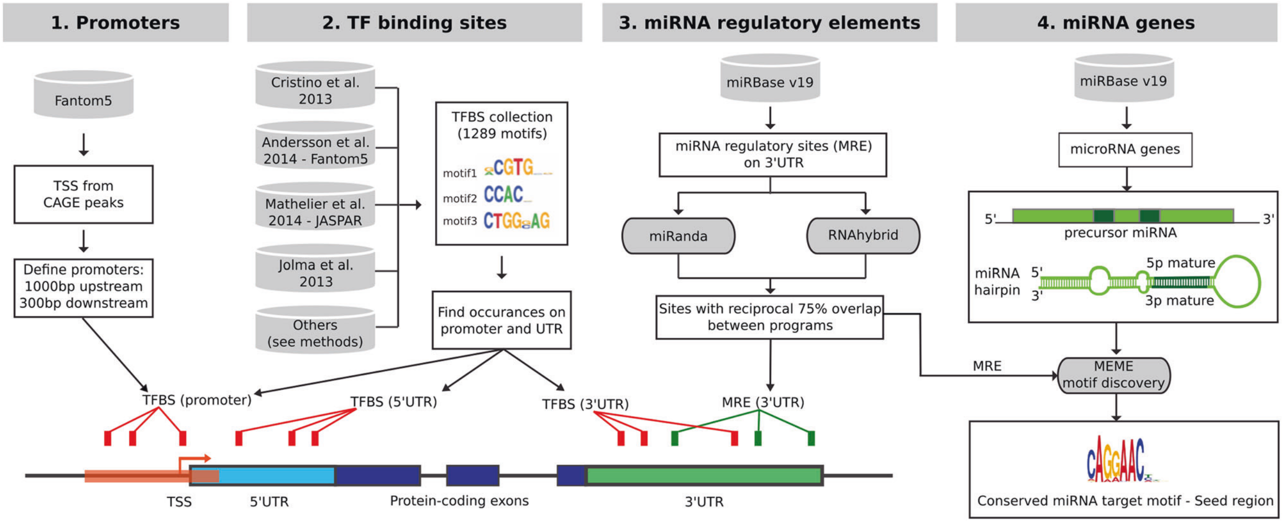

Fig. 1 An overview of the computational pipeline and databases used to identify the putative regulatory regions. Identification of the regulatory regions to check for variants in (1) promoters and transcription start site (TSS), (2) TFBS, (3) MREs, and (4) miRNA genes

grown for 24 h before pulldown. Four biological replicates, from four independent transfection and pulldown experiments, were made for each microRNA (WT and Mut). RNA was purified with RNAeasy columns (QiAGEN), and sequencing libraries were prepared from 100-ng input RNA with Illumina TruSeq Stranded mRNA kit. Libraries were sequenced across two lanes using v4 SBS chemistry on an Illumina HiSeq2000. Cells were grown in DMEM F12 with HEPES media with penicillin/streptomycin at 37 °C with 5% CO_2. The synthetic microRNA duplexes with biotin tags on the 3′ end of the mature microRNA were ordered from IDT (Supplementary Table 1).

The efficiency of transfection was checked quantitatively through staining. Then cells were fixed in 4% PFA, rinsed with PBS, and blocked with 2% BSA in 0.3% Triton X-100 (TX) for 30 min at room temperature for 2 h. The cells were incubated with Alexa Flour 594 streptavidin (Illumina) in BSA/PBS/TX mix for 1 h at room temperature on rotation, followed by 10 min of DAPI incubation.

Sequence data quality was examined with FastQC (v0.11.3). Illumina-adapter sequences were removed and reads quality trimmed with Trimmomatic [51]. Sequences were aligned to the human-GRCh38-Ensembl-release-83 transcriptome with TopHat2 [52]. The gene-level counts were generated from paired reads in HTseq-count (Parameters: -s reverse -m union). The differential expression analysis to evaluate the enrichment of targets in pulldown samples (paired comparisons), compared to matched whole-transcriptome of transfected cells (controls), were done using DEseq2 [53]. Gene ontology hypergeometric enrichment was calculated with GOStats (v2.40.0, parameters, conditional = FALSE) [54]. The sequence data is available via GEO accession: GSE98088.

Bioinformatics, data analysis tools, and structural modeling

Manipulation of sequence, annotation, and the SNV files was undertaken using VCFtools (v0.1.12a) [55], tabix (v0.2.6) [56], ANNOVAR (version: 2017Jul16) [23], SAMtools (v0.1.18) [57], and bedtools (v2.17.0) [58]. Cytoscape (v3) [59] was used for network visualization, and the ClueGO plugin (v2.3.2) [60] used for functional annotation enrichment. The crystal structure of the extracellular domain of NRXN 1 alpha was from *Bos taurus* (PDB: 3QCW [61]), and the resolution of the structure was used as the model was 2.65 Å. The Pro436Ser mutation was modeled using FoldX version 3.0b4 (http://foldxsuite.crg.eu/), with default settings.

Results

Annotation, and filtering the genetic variants in the regulatory regions

In this study, we have used the previously published WES data [7] to investigate putative regulatory genetic variation in 128 individual genomes from an Australian ASD cohort (48 affected cases and 80 parents). To examine the captured non-coding regions, we developed a computational pipeline to systematically analyze WES data in non-coding regulatory contexts: proximal promoter regions, transcription start site (TSS), untranslated regions (5′ and 3′UTRs), TFBS, MRE, and miRNA genes (Fig. 1).

We generated overlapping predictions of TFBSs (4.8 million in 5′UTR/22.4 million in 3′UTR from 1289 DNA

Table 1 Summary of regulatory region predictions, exome-sequencing coverage, and SNVs in the regulatory regions

Regulatory region	Number of features	Cumulative region size	SNV-callable size (% of total size)	Average number of SNVs per individual
Promoters	235917	74.57 Mb	9.7 Mb (13%)	148
TFBS - 5′UTRs	4770730	7.67 Mb	1.9 Mb (24%)	53
TFBS - 3′UTRs	22447998	22.75 Mb	4.8 Mb (21%)	63
MRE - 3′UTRs	2624853	18.08 Mb	3.5 Mb (20%)	35
microRNAs	2042	0.05 Mb	0.02 Mb (48%)	0.07
Enhancers	43011	12.39 Mb	0.06 Mb (0.5%)	0.1

Note that, feature counts include redundant features (e.g., target sites of similar miRNAs) and overlapping regions (e.g., promoters from alternate). Feature sizes and coverage are calculated using non-redundant genomic regions. SNV-callable coverage is defined as genomic regions with 20 TSS × read depth (sufficient for SNV-calling) in at least 90% of the samples. The raw variants are all quality passed SNV calls, and the filtered variants are those having passed the MAF (< 0.01) and functional score (CADD score ≥ 15) filtering, described in the methods. For the promoters, only SNVs in TFBS and TSS (CAGE_peak) were used to estimate the average number of filtered variants per individual

TFBS Transcription factor binding site. *MRE* miRNA regulatory elements

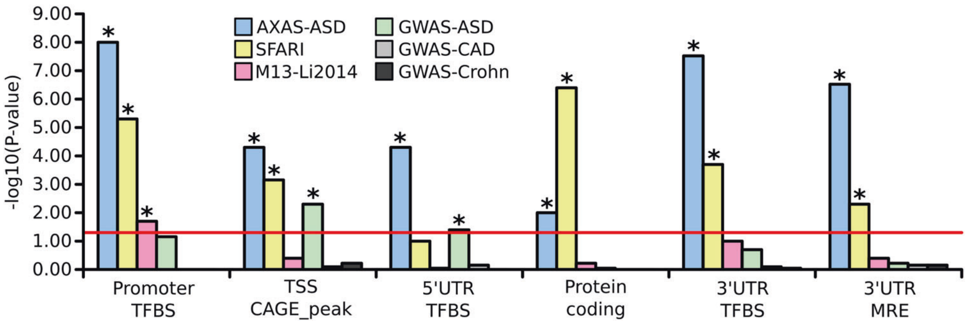

Fig. 2 Enrichment of the filtered regulatory variants and the protein-coding variants from ASD cases in different ASD-associated datasets AXAS-ASD ($n = 2664$) [30], SFARI ($n = 397$) [47], M13 - Li 2014 ($n = 115$) [48], and GWAS-ASD (ASD associated genes as described in the NHGRI-EBI GWAS Catalog). Two unrelated disease control datasets, coronary artery disease (GWAS-CAD), and Crohn's disease (GWAS-Crohn) were also retrieved from the NHGRI-EBI GWAS Catalog. The bar graph shows –log10(*P*-values) for the Z-scores (Supplementary Table 3). Only genes found in the Human Genome Nomenclature Consortium (HGNC) database and the whole protein-protein interaction network [30] were considered as the background datasets used for Z-score calculations (Supplementary Table 3)

motifs) and MREs (2.6 million from 2042 mature miRNAs) across 59,133 untranslated regions of 19,414 genes, and used public databases of miRNA genes (miRbase v19) [42], transcription start sites and promoters (FANTOM5) [12] and enhancers (FANTOM5) [12] (Table 1).

We assessed the WES coverage across the different non-coding regions (Supplementary Figure 1). There was sufficient read depth for the call-variants across 20–24% of all predicted TFBS and MREs in the untranslated regions (Table 1). Fortuitously, despite the limitations of exome capture design in the non-coding regions, the actual sequence coverage is highest in the regions proximal to exome-capture gene targets, where the most potentially interesting cis-regulatory elements would be expected (Supplementary Figure 1). The MiRNA genes were well-covered (48%), and although the promoters had less callable coverage (13%), they still yielded valuable data (Table 1). The enhancers had very low callable coverage (1%) due to their distance from the captured exonic regions, and were omitted from this analysis (Table 1).

After filtering the SNVs for rare population frequency (Minor Allele Frequency ≤1% based on gnomAD exome and genome collections) and functional annotation (CADD score ≥15), each individual case had an average of 288 protein-coding and 299 regulatory variations (Supplementary Figure 2; Supplementary Table 2). Greater numbers of the SNVs were identified in protein-coding regions and TFBSs in the promoter regions, followed by UTR TFBSs and 3′UTR MREs, reflecting the relative sizes of these regions (Table 1).

The genome-wide annotation of regulatory regions, and TFBS PWM (Position Weight Matrix) definitions in TAMO format [41] are available at figshare (https://doi.org/10.6084/m9.figshare.2198773.v1).

Association of the rare variants with the ASD-risk genes

We used a standardized score (Z-score) of the binomial distribution test to assess whether the genes with rare

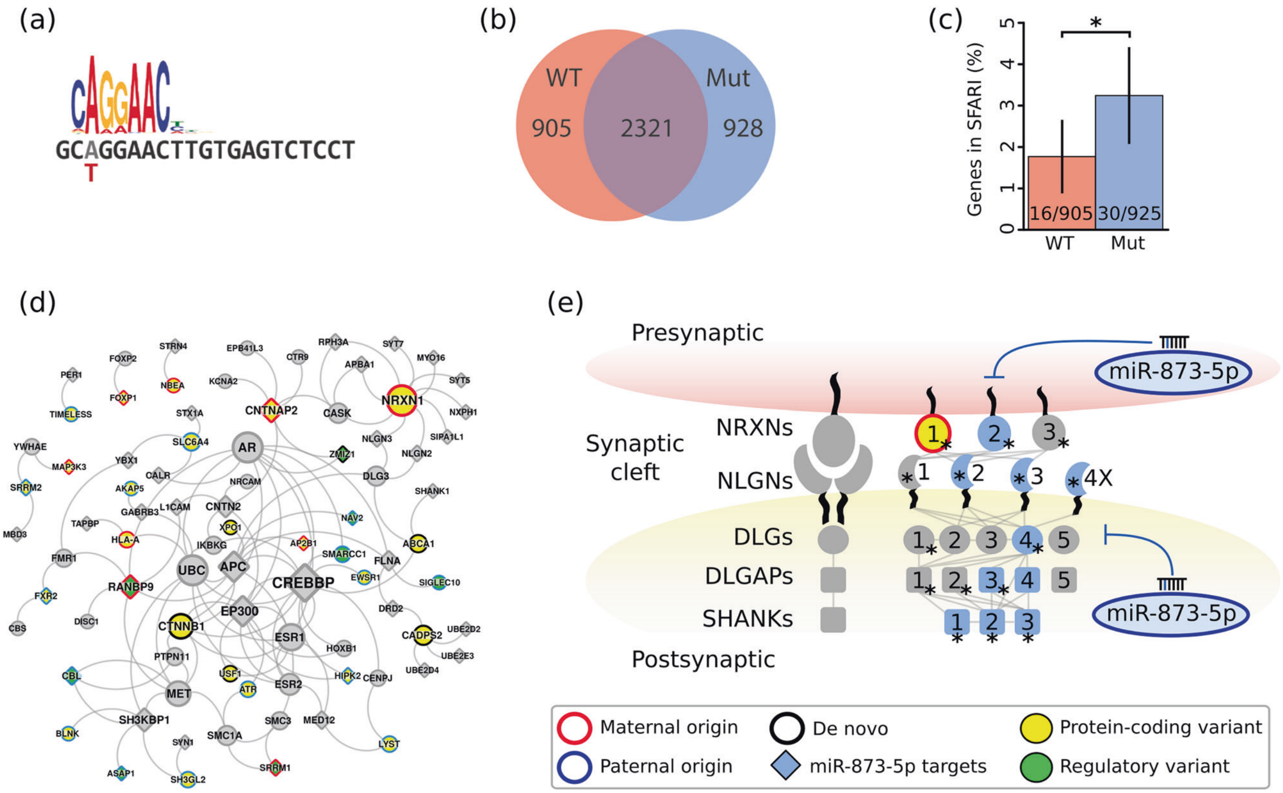

Fig. 3 Biological significance of the miR-873 variant. **a** miR-873-5p showing the variant in individual f34-s1, relative to the conserved seed motif generated from predicted miR-873-5p regulatory elements with MEME [45]. **b** Venn diagram showing the overlap between genes significantly enriched in the miR-875-5p wild-type (WT) and mutant (Mut)-pulldown assays. **c** Percentage of the SFARI genes in genes specifically enriched in the miR-875-5p wild-type (WT) ($n = 905$) or mutant ($n = 925$) pulldowns – genes enriched in both pulldowns are not included. *P*-value calculated with chi-squared test between the groups, *p*-value = 0.044, chi-squared statistic = 4.06, and degrees of freedom = 1. Error bars represent 95% confidence interval. Only genes present in both pulldown assays at high enough levels to have been contrasted (*p*-value calculated) by DEseq2 in both WT and Mutant experiments were included, omitting three genes that were significantly enriched in Mut but uncertain in WT. **d** Rare protein-coding and regulatory variants found in case f34-s1 overlap the AXAS-ASD network [30] and the miR-873-5p target genes (shown as diamonds, see Supplementary Table 6 for more details). The overlapped network shown is enriched in functional pathways associated with synaptic vesicle cycle and synaptic transmission (Supplementary Table 7). **e** Family f34 case study showing direct interaction of several synaptic proteins in the neurexin-neuroligin axis, in the context of the maternally inherited putative loss-of-function *NRXN1* coding and the paternally inherited miR-873-5p variants in individual f34-s1. Putative target genes of miR-873-5p are shown in blue. (*) SFARI genes

protein-coding and non-coding regulatory SNVs found in our ASD cohort were over-represented in the ASD-associated genes. In a previous study [7], we have shown that the rare protein-coding SNVs identified in our ASD cohort were significantly enriched in ASD-associated pathways (AXAS-ASD) [30]. Herein, we have modified our pipeline to identify the putative loss-of-function SNVs in both protein-coding and non-coding regions, and also to reduce filtering steps biases (MAF < 0.01 and CADD score ≥15; Supplementary Table 2). Strikingly, the regulatory SNVs identified in ASD cases were significantly enriched in the ASD-associated datasets (AXAS-ASD, SFARI, M13-Li2014, and GWAS-ASD; Fig. 2 and Supplementary Table 3), but not in the control datasets—coronary artery disease (GWAS-CAD) and Crohn's disease (GWAS-Crohn; Fig. 2 and Supplementary Table 3). Interestingly, the SNVs identified in the protein-coding regions showed the highest enrichment in SFARI gene list, while the SNVs in the non-coding regulatory regions are highly enriched in the AXAS-ASD-gene network (Fig. 2; Supplementary Table 3).

Having seen an overall ASD association, we then examined whether the regulatory mutations were over-represented in particular pathways or regions during neurodevelopment. Parikshak et al. [2] have previously constructed and characterized a set of co-expressed gene modules with respect to ASD. We saw significant enrichment of genes having regulatory mutations in ASD cases against two modules associated with regulation of gene expression (M3) and early synaptic development (M16; Supplementary Table 4), at late fetal and early-postnatal timescales [2].

Effect of the rare microRNA variants in ASD-associated gene networks

We identified a total of four microRNAs with rare SNVs within their mature sequences, which could globally disrupt their mRNA binding specificity (Supplementary Table 5). Of these, one variant was in the conserved “seed” regions of the microRNAs, where there is greater potential for disruption due to the exact target complementarity required in canonical microRNA-mRNA binding (Supplementary Table 5). We hypothesize that miRNA mutations within the mature sequences will affect the microRNA-mRNA-binding affinity and change the regulation of several target genes. We found miR-873 gene had a SNV within the seed region of the mature miR-873-5p (Fig. 3a) sequence. Computational predictions of miRNA binding suggest that the mutant miR-873 loses 76% of the predicted target genes, while also gaining 74% potential new target genes. This is the greatest predicted loss- or gain-of-binding of any of the miRNA variants (Supplementary Table 5). In contrast, the impact of variants that lie in less-conserved positions outside the seed regions (e.g., miR-411-5p, miR-668-3p and miR-323b-5p; Supplementary Table 5) would be less disruptive.

We propose that this rare miR-873-5p seed mutation could be relevant to ASD-associated genes and pathways, due to miR-873’s expression and genomic context. The miR-873 gene is embedded in the intronic region of *the LINGO2* gene (second intron of *LINGO2* variant 2). Examples of *LINGO2* copy number variation (CNV) have been seen previously in ASD cases [62-65], including two instances where CNV knocks out the 5′ region of *LINGO2* variant 2 including the miR-873 gene [66, 67]. Furthermore, *LINGO2* has been shown to exhibit a neural-specific expression pattern in early development in mice [68]. Although the host gene expression is not an absolute indicator of embedded microRNA context, it may be concordant, and miR-873-5p mature sequence itself has neuronal expression in the human brain (Supplementary Figure 3) [69].

Biotinylated microRNA-mRNA pulldown experiments were performed to identify the transcriptome-wide binding profiles of wild-type and mutant miR-873-5p (Fig. 3b) with 3226 and 3249 significantly enriched targeted genes identified respectively (Supplementary Table 6). By contrasting the relative pulldown assay enrichment of the wild-type and mutant miR-873-5p, we identified 2321 overlapping genes putatively targeted by both wild-type and mutant miR-873-5p, while 905 and 928 genes were specifically enriched in the wild-type and mutant microRNA-mRNA pulldown, respectively (Fig. 3b).

We assessed functional associations of the miR-873-5p target genes by calculating gene ontology enrichment of their significantly enriched genes (Supplementary Table 6). Notably, the target genes of both wild-type and mutant miR-873-5p were both significantly enriched in the synapse genes (p-value = 2.88E-09). Moreover, 109 SFARI candidate genes (27%, SFARI categories 1-4,S) overlapped with miR-873-5p targets, identified in our miRNA-mRNA pulldown experiment (in either wild-type or mutant). Interestingly, the mutant-specific miR-873-5p targets are more enriched (compared to ‘wild-type’ miR-873-5p) with many of the ASD-risk genes documented in the SFARI database (chi-squared test, p-value = 0.04, df = 1) (Fig. 3c).

Members of the neurexin gene family have been implicated in cognitive disorders including schizophrenia and ASD [70]. Two neurexin-family members, *NRXN2* and *CNTNAP2* (also known as *NRXN4*), have been found significantly enriched in our pulldown assay. *NRXN2* has been found significantly enriched in both wild-type (log2 fold-change = 1.8, p-value = 4.03E-50) and mutant miR-873-5p (log2 fold-change = 1.1, p-value = 1.61E-20). However, the mutation may have caused a partial loss-of-binding to *NRXN2*, since its enrichment levels are smaller in mutant miR-873-5p-pulldown data (1.5 fold-change decrease in enrichment). *CNTNAP2* was found significantly enriched in mutant (log2 fold-change = 0.75, p-value = 2.5E-04) and wild-type miR-873-5p (log2 fold-change = 0.58, p-value = 0.03), but suggests partial gain-of-binding to *CNTNAP2* for the miR-873-5p variant. Intriguingly, miR873’s host gene, *LINGO2,* is enriched in the wild-type miR-873-5p-pulldown assay (log2 fold-change = 0.72, p-value = 6.76E-05) but not in the mutant microRNA (log2 fold-change = 0.14, p-value = 0.49), indicating loss-of-binding to *LINGO2* mRNA. Another relevant ASD-risk gene, the transcription factor *FOXP1* [71], has also been found to be significantly enriched only in the mutant miR-873-5p pulldown assay (log2 fold-change = 0.53, p-value = 3.7E-05) indicating a complete gain-of-binding, specifically associated with the variant miR-873-5p (Supplementary Table 6).

Convergence of the rare regulatory and the protein-coding variants in synaptic pathways

We chose to follow-up on the effect of the SNV within miR-873-5p found in individual f34-s1 due to its location in the seed region (Fig. 3a), the microRNA’s robustly expressed profiles across all regions of the human brain (Supplementary Figure 3), and functional enrichment analysis of its predicted target genes pointing toward a role in the regulation of nervous system development (Fisher’s exact test with Bonferroni step down corrected P-value = 0.002) and signaling (Fisher’s exact test with Bonferroni step down corrected $P = 0.0003$) [60].

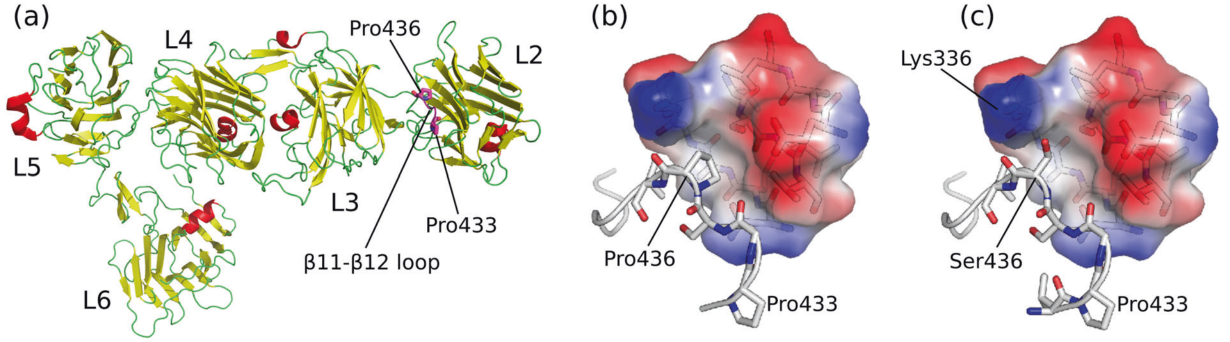

Fig. 4 Structural modeling of the Pro429Ser NRXN1α mutation using the crystal structure of bovine NRXN1α (PDB: 3QCW). **a** The β11-β12 loops comprise of a large part of the L2-L3-LNS domain interfaces in NRXN1α. In the L2 LNS domain, the loop is characteristically bent in a horseshoe type conformation, with two proline residues Pro433 and Pro436 (magenta: equivalent to human Pro426 and Pro429) stabilizing the tightly kinked structure. **b** The packing of this loop in the interface involves insertion of Pro436 into a hydrophobic pocket of NRXN1α. **c** The Pro436Ser replacement (equivalent to Pro429Ser in humans) analyzed with the FoldX force field revealed this mutation is moderately (1.81 kcal/mol) to severely (3.18 kcal/mol) destabilizing, depending on whether Ser436 can hydrogen bond to a disordered Lys336

However, the miR-873-5p SNV is heterozygous and inherited from an unaffected father, so although not causative on its own, we hypothesized that it could contribute additively through interactions with other small-effect mutations inherited from parents, as well as *de novo* mutations. In our published whole-exome screen, the family members were assessed for a 'broader autism phenotype' (BAP), reflecting a subclinical expression of ASD phenotypes [7, 72]. Among parents of this proband, the mother was recorded as having a BAP [7]. In this study, we focused on the maternally inherited *NRXN1* variant, as it is one of the most highly connected proteins in the case f34-s1-gene network (Fig. 3d) and a key ASD-risk gene involved in synaptic transmission [70]. Notably, in case f34-s1, few variant genes overlapped in both the AXAS-ASD and SFARI datasets and are well known players in the synaptic pathways (i.e., *NRXN1*, *CADPS2*, *CNTNAP2*) and regulation of gene expression (i.e., *CTNNB1*, *FOXP1*) previously associated with neurological disorders including autism [71, 73]. Three of these gene variants (*NRXN1*, *CNTNAP2,* and *FOXP1*) are inherited from the BAP mother, while two are *de novo* mutations (*CADPS2* and *CTNNB1*). Nevertheless, there will be additional contribution of several other small effects inherited and *de novo* variations in both protein-coding and regulatory regions of genes involved in the functional pathways associated with synaptic vesicle cycle and transmission (Fisher's exact test with Bonferroni step down corrected $P = 0.001$ [60]; Supplementary Table 7).

The rare variation in *NRXN1* genes (G ≥ A mutation at chr2:50847195; rs78540316) results in a Pro429Ser structural change in LNS domain 2 of NRXN1α (UniProtKB - Q9ULB1 [74]). The extracellular domain structure of NRXN1α and other alpha neurexins is characterized by the linear assembly of six LNS domains, with the L6 LNS domain extended to form an "L" shape; this assembly has been shown to be important for their function as synaptic organizers [61]. Continuous electron density of bovine NRXN1α (PDB: 3QCW) is visible for L2-L6 LNS domains showing a concatenated arrangement of these domains mediated by β11-β12 loops, which play an important role in the interaction between LNS domains [61] (Fig. 4a). The packing of this loop in the interface of L2-L3 LNS domain involves insertion of Pro436 into a hydrophobic pocket of NRXN1α (Fig. 4b). Pro436 is highly conserved (Supplementary Figure 4), supporting its important role for the structural arrangement of NRXN1α. Analysis of the putative effects of replacing this conserved proline residue with serine on the stability of NRXN1α was undertaken with the FoldX force field, which has been shown to accurately predict the effects of mutations on the free energy of unfolding of proteins [75]. This revealed that the mutation is moderately (1.81 kcal/mol) to severely (3.18 kcal/mol) destabilizing, depending on whether Ser436 can hydrogen bond to the disordered lysine residue (Lys336), due to the loss of these stabilizing hydrophobic interactions (Fig. 4c). Given the importance of Pro436 to the correct packing of the L2-L3 LNS domains, the mutation of this conserved proline residue to serine (equivalent to Pro429Ser mutation in the human *NRXN1*) will most likely result in loss-of-stability of NRXN1α and disruption of the L2–L3–LNS domains interface. Thus, the predicted destabilizing effects of the Pro429Ser mutation provides a plausible structural change that will affect the structural integrity and thus the function of NRXN1α.

Considering this structural evidence, we propose that putative loss of NRXN1α protein function may additively contribute to mutant miR-873-5p dysregulation of the *NRXN2* expression (Fig. 3e) and increase the threshold of genetic liability regarding their function in pre-synaptic transmission and synapse development in case f34-s1.

Moreover, there may be further impact at the synapse, as the miR-873-5p-pulldown assay indicated the capability to bind a number of other ASD-risk genes. MiR-873-5p binds to key genes for synapse formation and function including all three SHANK family members (*SHANK1*, *SHANK2* and *SHANK3)*, three Neuroligin family members (*NLGN2*, *NLGN3* and *NLGN4X*), Discs Large MAGUK Scaffold Protein 4 (*DLG4* also known as *PSD95*), and two DLG associated proteins (*DLGAP3* and *DLGAP4*), which have been linked to ASD by multiple studies [47, 76, 77] (Fig. 3e, Supplementary Table 6).

Discussion

Our analyses provide a novel and comprehensive genome-wide approach for identification and characterization of putative regulatory variations in an exome-sequenced-ASD cohort. By combining our computational TFBS and MRE predictions with cis-regulatory elements resources, we showed there was an overall cohort-level enrichment of the regulatory variants within the ASD-associated genes, derived from several gold-standard ASD datasets (Fig. 2) [30, 47, 48]. We specifically focused on a rare mutation observed in the miR-873 gene, and characterized its potential impact in the regulation of the synaptic genes.

ASDs have complex and heterogeneous genetic causes, and may arise from combinations of inherited and de novo variation in the protein-coding [4] and non-coding regions [78, 79]. To investigate a possible additive contribution of the regulatory SNVs, it therefore makes sense to focus on the clusters of variants, both regulatory and protein-coding, that are in the functional neighborhood of known ASD-associated gene networks and functional pathways [30]. Herein, we identified the rare loss-of-function variants in the regulatory regions potentially disrupting the expression of the "peripheral" genes incurring in non-zero effects on regulation of the "core" genes with more direct and specific roles in disease etiology. Our findings support the omnigenic model of complex disease [80], which proposes that any regulatory variation in the genes expressed in disease-relevant cell-types can contribute to disease etiology with non-linear effects on risk for that disease/disorder.

There have been several studies that have identified the functional gene modules involved in ASD-associated biological processes [2, 3, 48]. Two of the genome-scale co-expression modules constructed by Parikshak et al. [2] were enriched in our regulatory variant set; notably one of the modules was the same one (M16) that they had highlighted for the potential involvement in ASD. It is enriched for ASD relevant GO terms such as "synaptic transmission" and "homophilic cellular adhesion" and more specifically expressed in later fetal development and early infancy [2]. The hypothesis of Parikshak et al. [2], that M16 might be targeted by lower-risk inherited variants, in contrast to other modules involved in more fundamental early developmental processes, is supported by enrichment of our dataset of regulatory inherited mutations.

Our study identified four variants found within the miRNA genes that were predicted to disrupt gene regulation. Notably, we identified a rare variation in the seed region (the major factor in binding specificity) of miR-873-5p and functionally verified the changes in binding affinity of the miR-873-5p variant. MiR-873 is of particular interest as it is located within a neural gene *LINGO2*, which has been investigated with respect to essential tremor and Parkinson's disease [81]. There are also a number of reported CNVs of *LINGO2* found in ASD cases [62, 63, 64, 65], including one which overlays miR-873 [66, 67]. Furthermore, miR-873 promoter expression has also been found to be significantly enriched in human brain cells and tissues (Fantom5 miRNA Atlas; *P*-value = 2e-40) [82]. It is still unknown whether these particular rare SNVs on miR-873-5p (rs777143952; MAF < 6.6e-05) will be statistically over-represented in ASD cases, and only very large whole-genome sequencing (WGS) studies would have the necessary power to assign any statistical association to a specific rare SNV. However, the relevance of the rare variants outside protein-coding regions, for human complex traits and phenotypic variance, has only started to be explored in developmental disorders by recent studies using WGS data from the large samples sets of European ancestry [78, 79].

The binding profile of miR-873-5p measured by the pulldown assay includes several ASD-risk genes—all three members of the SHANK family (*SHANK1*, *SHANK2*, *and SHANK3*), several neuroligins (*NLGN2*, *NLGN3*, and *NLGN4X*), and neurexins (*NRXN2 and CNTNAP2).* This targeting information, combined with the neural expression of miR-873-5p and its host gene *LINGO2*, strongly suggest a potential regulatory role in synapse development and function. We also found that the miR-873-5p variant binds to the transcription factor *FOXP1* transcript, suggesting a dysregulation of FOXP1 expression, which has previously been shown to play a critical regulatory role in CNTNAP2 expression [83], as well as a functional role of striatal pathways important for vocal communication [84]. Despite this miRNA mutation being inherited from an unaffected parent, the measurable loss-of-binding in several ASD-risk genes leads us to postulate that it may contribute additively to ASD risk, particularly when combined with the mutation load from the mother with recorded BAP. Consistent with this idea, is the convergence of functional disruption of neurexins in this individual—specifically an observed partial change in *NRXN2* and *CNTNAP2* binding affinity from the paternally inherited miR-873-5p mutation when combined with the maternally inherited *NRXN1* and *CNTNAP2* coding variants. Moreover, this ASD case (f34-s1) carry two additional de novo protein-coding variations in *CADPS2* gene that are involved in neurotrophin release and interaction with dopamine receptor type 2 [85], and *CTNNB1* gene associated with impaired social interactions and repetitive behaviors [86]. Considering the above evidence concerning pathways and processes associated with ASD (e.g., synapse development), this work showcases the importance of compounded effects of protein-coding and regulatory mutations in a biological systems framework, particularly in the case of de novo mutations, which contribute significantly to ASD risk [10].

We are aware of the limitations of using a human neuroblastoma cell line (SH-SY5Y), in which the transcript levels and alternative splicing isoforms might not necessarily capture the same molecular interactions as CNS neurons, and we are mindful that spatial and temporal gene expressions patterns of various brain cells and tissues have been shown to also vary. Nevertheless, our study is the first to propose an integrated approach to predict and validate a mechanistic hypothesis for a possible causal effect of a single nucleotide variation changing the binding affinity between miRNA and target transcripts. Hence, our approach still represents a significant and novel contribution to systematically evaluate the (under the same SH-SY5Y transcriptome background) loss or gain of affinity effects of the genetic variants in regulatory regions, such as miRNA genes.

We are also mindful that our analyses are focused on rare loss-of-function heterozygous DNA variations that are represented by a minor allele frequency of $<1\%$ in the human population. We are similarly aware of the ongoing debate as to the relative contribution of the rare inherited alleles to ASD and that there exists an uncertainty associated with current estimates [87]. Much of this uncertainty surrounds the contribution of loss-of-function variations found in non-coding DNA that comprises majority of the genome. Arguably, in the absence of detailed whole-genome analysis, estimates concerning the contribution of the rare inherited non-coding variants to ASD remain speculative. The whole-genome sequencing approaches when combined with rigorous clinical and behavioral profiling of family cohorts and functional characterization of DNA variations are now providing the means to further identify and evaluate rare and de novo non-coding regulatory variations that may affect brain development [4, 6, 88, 89], yet we see that even exome-based data can yield useful information. A major confound regarding estimates of heritability surrounds the impact of epigenetic processes in the penetrance of heterozygous alleles. Although it is difficult to the access neural tissue, there are attempts to interpret global epigenetic events to identify patterns of methylation and other DNA modifying chemistries, and assess the regulatory significance of non-coding DNA variations occurring at these sites. Interestingly, Yuen et al. (2016) recently reported that de novo regulatory variations associated with *DNMT3A* and *ADNP* are likely to more directly affect the global methylation processes. With regard to the epigenetic mechanisms, any new data resource detailing non-coding sequences that control gene regulation or the stability of transposons and repetitive DNA that affect transcription, can logically be incorporated into this analysis pipeline to more completely account for other regulatory layer of DNA variations associated with ASD.

Public genomic resources are constantly improving, larger populations of SNVs will aid the identification of the rare variants and projects such as psychENCODE [90] and CommonMind consortium [91], aimed at generating detailed datasets of regulation in the brain, can only improve interpretations of the variants in the regulatory regions. The approach described here is equally applicable to the whole-genome sequencing data and to the evaluation of the variants identified in the genome-wide association studies; moreover it provides a generalized template for analysis of complex genetic disorders. When combined, more ASD-cases-sequence data, new genomic resources, and integrated systems-based approaches that examine the regulatory variants in ASD-associated genes and pathways, as identified in this study, should elucidate the role of the non-coding regulatory variants and their additive contribution to the risk of ASDs.

Acknowledgements We acknowledge the financial support of Australian Commonwealth Department of Social Services (formerly the Department of Families, Housing, Community Services and Indigenous Affairs). CC was supported by funding from the Australian Research Council (FT110100292) and the National Health and Medical Research Council (APP1008125). AJOW is supported by a Senior Research Fellowship form the National Health and Medical Research Council (APP1077966). JYA was supported by a University of Queensland PhD scholarship and SMW by an Australian Postgraduate Award.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature. 2012;485: 246–50.
2. Parikshak Neelroop N, Luo R, Zhang A, Won H, Lowe Jennifer K, Chandran V, et al. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. Cell. 2013;155:1008–21.
3. Willsey AJ, Sanders Stephan J, Li M, Dong S, Tebbenkamp Andrew T, Muhle Rebecca A, et al. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. Cell. 2013;155:997–1007.
4. An JY, Claudianos C. Genetic heterogeneity in autism: from single gene to a pathway perspective. Neurosci Biobehav Rev. 2016;68:442–53.
5. Devanna P, Chen XS, Ho J, Gajewski D, Smith SD, Gialluisi A et al. Next-gen sequencing identifies non-coding variation disrupting miRNA-binding sites in neurological disorders. Mol Psychiatry. 2017. Advance online publication. https://doi.org/10.1038/mp.2017.30.
6. Yuen RK, Merico D, Cao H, Pellecchia G, Alipanahi B, Thiruvahindrapuram B, et al. Genome-wide characteristics of de novo mutations in autism. NPJ Genom Med. 2016;1:160271–1602710.
7. An JY, Cristino AS, Zhao Q, Edson J, Williams SM, Ravine D, et al. Towards a molecular characterization of autism spectrum disorders: an exome sequencing and systems approach. Transl Psychiatry. 2014;4:e394.
8. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012;22:1748–59.
9. Takata A, Ionita-Laza I, Gogos JA, Xu B, Karayiorgou M. De novo synonymous mutations in regulatory elements contribute to the genetic etiology of autism and schizophrenia. Neuron. 2016;89:940–7.
10. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014;515:216–21.
11. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet. 2011;43:585–9.
12. Forrest ARR, Kawaji H, Rehli M, Kenneth Baillie J, de Hoon MJL, Haberle V, et al. A promoter-level mammalian expression atlas. Nature. 2014;507:462–70.
13. Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
14. Miller JA, Ding SL, Sunkin SM, Smith KA, Ng L, Szafer A, et al. Transcriptional landscape of the prenatal human brain. Nature. 2014;508:199–206.
15. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science. 2013;342:1235587.
16. Sabarinathan R, Wenzel A, Novotny P, Tang X, Kalari KR, Gorodkin J. Transcriptome-wide analysis of UTRs in non-small cell lung cancer reveals cancer-related genes with SNV-induced changes on RNA secondary structure and miRNA target sites. PLoS ONE. 2014;9:e82699.
17. Vaishnavi V, Manikandan M, Munirajan AK. Mining the 3′UTR of autism-implicated genes for SNPs perturbing microRNA regulation. Genom Proteom Bioinf. 2014;12:92–104.
18. Turner Tychele N, Hormozdiari F, Duyzend Michael H, McClymont Sarah A, Hook Paul W, Iossifov I, et al. Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. Am J Human Genet. 2016;98:58–74.
19. Yao P, Lin P, Gokoolparsadh A, Assareh A, Thang MW, Voineagu I. Coexpression networks identify brain region-specific enhancer RNAs in the human brain. Nat Neurosci. 2015;18: 1168–74.
20. Taylor LJ, Maybery MT, Wray J, Ravine D, Hunt A, Whitehouse AJ. Brief report: do the nature of communication impairments in autism spectrum disorders relate to the broader autism phenotype in parents? J Autism Dev Disorders. 2013. Advance online publication. https://doi.org/10.1038/mp.2017.30.
21. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589–95.
22. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
23. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:1–7.
24. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.
25. Anney R, Klei L, Pinto D, Almeida J, Bacchelli E, Baird G, et al. Individual common variants exert weak effects on the risk for autism spectrum disorders. Hum Mol Genet. 2012;21:4781–92.
26. Yuen RK, Thiruvahindrapuram B, Merico D, Walker S, Tammimies K, Hoang N, et al. Whole-genome sequencing of quartet families with autism spectrum disorder. Nat Med. 2015;21: 185–91.
27. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310–5.
28. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507:455–61.
29. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 2014;42:D142–7.
30. Cristino AS, Williams SM, Hawi Z, An JY, Bellgrove MA, Schwartz CE, et al. Neurodevelopmental and neuropsychiatric disorders represent an interconnected molecular system. Mol Psychiatry. 2013;19:294–301.
31. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-binding specificities of human transcription factors. Cell. 2013;152:327–39.
32. Boyer La, Plath K, Zeitlinger J, Brambrink T, Medeiros La, Lee TI, et al. Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature. 2006;441:349–53.
33. Bruce AW, Donaldson IJ, Wood IC, Yerbury Sa, Sadowski MI, Chapman M et al. Genome-wide analysis of repressor element 1 silencing transcription factor/neuron-restrictive silencing factor (REST/NRSF) target genes. Proc Natl Acad Sci USA. 2004; 101**:** 10458–63.
34. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell. 2008;133:1106–17.
35. Gu P, Goodwin B, Chung AC, Xu X, Wheeler Da, Price RR, et al. Orphan nuclear receptor LRH-1 is required to maintain Oct4 expression at the epiblast stage of embryonic development orphan nuclear receptor LRH-1 is required to maintain oct4 expression at the epiblast stage of embryonic development. Mol Cell Biol. 2005;25:3492–505.
36. Jin VX, O'Geen H, Iyengar S, Green R, Farnham PJ. Identification of an OCT4 and SRY regulatory module using integrated computational and experimental genomics approaches. Genome Res. 2007;17:807–17.

37. Loh Y-H, Wu Q, Chew J-L, Vega VB, Zhang W, Chen X, et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat Genet. 2006;38:431–40.
38. Minoguchi S, Taniguchi Y, Kato H, Okazaki T, Strobl LJ, Zimber-Strobl U, et al. RBP-L, a transcription factor related to RBP-Jkappa. Mol Cell Biol. 1997;17:2679–87.
39. Mizugishi K, Aruga J, Nakata K, Mikoshiba K. Molecular properties of Zic proteins as transcriptional regulators and their relationship to GLI proteins. J Biol Chem. 2001;276:2180–8.
40. Zhou Q, Chipperfield H, Melton Da, Wong WH. A gene regulatory network in mouse embryonic stem cells. Proceedings of the National Academy of Sciences of the USA. 2007; 104**:** 16438–43.
41. Gordon DB, Nekludova L, McCallum S, Fraenkel E. TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. Bioinformatics. 2005;21:3164–5.
42. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2011;39:D152–157.
43. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. Genome Biol. 2003;5:R1.
44. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. RNA. 2004;10:1507–17.
45. Bailey TL. Discovering novel sequence motifs with MEME. Curr Protoc Bioinf. 2002;Chapter 2:Unit 2 4.
46. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. eLife 2015;4. https://doi.org/10.7554/eLife.05005.
47. Abrahams BS, Arking DE, Campbell DB, Mefford HC, Morrow EM, Weiss La, et al. SFARIGene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Mol Autism. 2013;4:36.
48. Li J, Shi M, Ma Z, Zhao S, Euskirchen G, Ziskin J, et al. Integrated systems analysis reveals a molecular network underlying autism spectrum disorders. Mol Syst Biol. 2014;10:774–774.
49. Wani S, Cloonan N. Profiling direct mRNA-microRNA interactions using synthetic biotinylated microRNA-duplexes. *BioRxiv*. 2014**:** 0–11.https://doi.org/10.1101/005439.
50. Cristino AS, Barchuk AR, Freitas FC, Narayanan RK, Biergans SD, Zhao Z, et al. Neuroligin-associated microRNA-932 targets actin and regulates memory in the honeybee. Nat Commun. 2014;5:5529.
51. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
52. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14: R36–R36.
53. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
54. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics. 2007;23:257–8.
55. Danecek P, Auton A, Abecasis G, Albers Ca, Banks E, DePristo Ma, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
56. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics. 2011;27:718–9.
57. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
58. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
59. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011;27:431–2.
60. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25:1091–3.
61. Chen F, Venugopal V, Murray B, Rudenko G. The structure of neurexin 1α reveals features promoting a role as synaptic organizer. Structure. 2011;19:779–89.
62. Gai X, Xie H, Perin J, Takahashi N, Murphy K, Wenocur A, et al. Rare structural variation of synapse and neurotransmission genes in autism. Mol Psychiatry. 2011;17:402–11.
63. Gazzellone MJ, Zhou X, Lionel AC, Uddin M, Thiruvahindrapuram B, Liang S, et al. Copy number variation in Han Chinese individuals with autism spectrum disorder. J Neurodev Disord. 2014;6:34–34.
64. Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L, et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am J Hum Genet. 2014;94:677–94.
65. Kousoulidou L, Moutafi M, Nicolaides P, Hadjiloizou S, Christofi C, Paradesiotou A, et al. Screening of 50 cypriot patients with autism spectrum disorders or autistic features using 400 K custom array-CGH. Biomed Res Int. 2013;2013:843027.
66. Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, et al. Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet. 2008;82:477–88.
67. Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. Neuron. 2015;87: 1215–33.
68. Homma S, Shimada T, Hikake T, Yaginuma H. Expression pattern of LRR and Ig domain-containing protein (LRRIG protein) in the early mouse embryo. Gene Expr Patterns. 2009;9:1–26.
69. Allen Brain Institute. BrainSpan: Atlas of the Developing Human Brain. 2011.
70. Südhof TC. Neuroligins and neurexins link synaptic function to cognitive disease. Nature. 2008;455:903–11.
71. Bacon C, Rappold GA. The distinct and overlapping phenotypic spectra of FOXP1 and FOXP2 in cognitive disorders. Human Genet. 2012;131:1687–98.
72. Wheelwright S, Auyeung B, Allison C, Baron-Cohen S. Defining the broader, medium and narrow autism phenotype among parents using the Autism Spectrum Quotient (AQ). Mol Autism. 2010;1:10.
73. Krumm N, O'Roak BJ, Shendure J, Eichler EE. A de novo convergence of autism genetics and molecular neuroscience. Trends Neurosci. 2014;37:95–105.
74. UniProt. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43:D204–12.
75. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Res. 2005;33:W382–8.
76. Chen J, Yu S, Fu Y, Li X. Synaptic proteins and receptors defects in autism spectrum disorders. Front Cell Neurosci. 2014;8:276.
77. Leblond CS, Nava C, Polge A, Gauthier J, Huguet G, Lumbroso S, et al. Meta-analysis of SHANK Mutations in Autism Spectrum Disorders: a gradient of severity in cognitive impairments. PLoS Genet. 2014;10:e1004580.
78. Consortium UK, Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015;526:82–90.
79. Deciphering Developmental Disorders S. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017;542:433–8.

80. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. Cell. 2017;169:1177–86.
81. Vilariño-Güell C, Wider C, Ross Oa, Jasinska-Myga B, Kachergus J, Cobb Sa, et al. LINGO1 and LINGO2 variants are associated with essential tremor and Parkinson disease. Neurogenetics. 2010;11:401–8.
82. de Rie D, Abugessaisa I, Alam T, Arner E, Arner P, Ashoor H, et al. An integrated expression atlas of miRNAs and their promoters in human and mouse. Nat Biotechnol. 2017;35:872–8.
83. Rodenas-Cuadrado P, Ho J, Vernes SC. Shining a light on CNTNAP2: complex functions to complex disorders. Eur J Human Genet. 2014;22:171–8.
84. Araujo DJ, Anderson AG, Berto S, Runnels W, Harper M, Ammanuel S, et al. FoxP1 orchestration of ASD-relevant signaling pathways in the striatum. Genes Dev. 2015;29:2081–96.
85. Bonora E, Graziano C, Minopoli F, Bacchelli E, Magini P, Diquigiovanni C, et al. Maternally inherited genetic variants of CADPS2 are present in autism spectrum disorders and intellectual disability patients. EMBO Mol Med. 2014;6:795–809.
86. Dong F, Jiang J, McSweeney C, Zou D, Liu L, Mao Y. Deletion of CTNNB1 in inhibitory circuitry contributes to autism-associated behavioral defects. Hum Mol Genet. 2016;25:2738–51.
87. Gaugler T, Klei L, Sanders SJ, Bodea CA, Goldberg AP, Lee AB, et al. Most genetic risk for autism resides with common variation. Nat Genet. 2014;46:881–5.
88. Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell. 2012;151:1431–42.
89. McRae JF, Clayton S, Fitzgerald TW, Kaplanis J, Prigmore E, Rajan D, et al. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017;542:433–8.
90. Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, Crawford GE, et al. The PsychENCODE project. Nat Neurosci. 2015;18:1707–12.
91. CommonMind Consortium. CommonMind Consortium Knowledge Portal. https://www.nimhgenetics.org/acknowledgements.php