Network Modeling Analysis in Health Informatics and Bioinformatics (2022) 11:33
https://doi.org/10.1007/s13721-022-00377-z

ORIGINAL ARTICLE

Integrated in silico functional analysis predicts autism spectrum disorders to be burdened by deleterious variations within CHD8 core domains and its CHD7-binding motif

Ashitha S. Niranjana Murthy[1] · Suryanarayanan Thangalazhi Balakrishnan[1] · Ramachandra Nallur B.[1]

Received: 10 February 2022 / Revised: 17 July 2022 / Accepted: 26 August 2022 / Published online: 21 September 2022

Abstract

Autism spectrum disorder (ASD) is a neurodevelopmental disorder presenting with social and communication deficits, restricted, repetitive behaviours and interest. Several recurrently mutated genetic risk factors have been implicated in ASD manifestation. *Chromodomain helicase remodeller* (*CHD8*) is one such master regulator mediating the expression of genes controlling neuron functions. We collected 8124 exonic SNPs in *CHD8* from four databases representing the general and ASD populations and subjected them to multi-layered analyses on > 25 computational tools. We observed that nsSNPs were common in the general population. Contrastingly, the ASD population recorded significantly higher incidences of truncating SNPs than the general population ($P < 0.0001$). Distinct hotspots for truncating and nsSNPs were identified within exons encoding CHD8's N and C terminals, respectively. The evolutionarily conserved CHD8 core domains—helicase C-terminal, helicase ATP-binding and SNF2_N domains—recorded the lowest density of SNPs that were predicted to be severely damaging. Conversely, the evolutionarily variable regions—CHD7-binding and BRK domains—that hosted the highest aggregate of SNPs were largely benign. Post-translational modifications (PTMS) occurred frequently on residues outside the CHD8 domains ($P < 0.01$); i.e. on non-conserved regions including the N and C terminals, which were also predicted to be intrinsically disordered protein regions with nine molecular recognition feature sites. ASD SNPs frequently occurred within core domains, were severely damaging and accounted for > 30% of all ASD variations. The CHD7-DNA-binding motif, with most PTMs, recorded the highest recurring truncating ASD SNPs. The CHD8 protein–protein interactions recapitulated the clinical phenotypes presented by children with *CHD8* mutations. 11/13 (84.6%) interacting molecules were intrinsically disordered proteins. We identified nine *CHD8* nsSNPs that produced the strongest long-range disturbances, altering the modelled protein's global conformational dynamics.

Keywords Autism spectrum disorders (ASD) · *Chromodomain helicase DNA-binding protein 8* (CHD8) · Intrinsically disordered protein (IDP) · Molecular recognition features (MoRFs) · Protein–protein interaction (PPI) networks · Conformational dynamics

Abbreviations

CHD	*Chromodomain helicase DNA-binding protein (genes are italicized)*
CHD	Protein CHD8
Amino acid	Amino acid
ASD	Autism spectrum disorders
ATP	Adenosine triphosphate
BLAST	Basic Local Alignment Search Tool
BRK	Brahma and Kismet
CLS	Cytoplasmic localization sequence
CONDEL	CONsensus DELeteriousness
D	Deleterious and/or destabilizing
dbSNP	Database of SNPs
DEG	Differentially expressed genes

✉ Ramachandra Nallur B.
nallurbr@gmail.com

Ashitha S. Niranjana Murthy
ashitha@zoology.uni-mysore.ac.in

Suryanarayanan Thangalazhi Balakrishnan
suryantb1995@gmail.com

[1] Genetics and Genomics Laboratory, Department of Studies in Genetics and Genomics, University of Mysore, Manasagangotri, Karnataka 570006, India

DEPICTER	DisorderEd PredictIon CenTER
DNA	Deoxyribonucleic acid
ENCoM	Elastic Network Contact Model
EVS	Exome Variant Server
ExAC	Exome Aggregate Consortium
FATHMM	Functional analysis through hidden Markov models
GI	Gastrointestinal
GMQE	Global Model Quality Estimate
gnomAD	Genome Aggregation Database
GSEA	Gene-set enrichment analysis
hNPC	Human neuronal progenitor cells
ID	Intellectual disability
IDP	Intrinsically disordered protein
IDPRs	Intrinsically disordered protein regions
IPA	Ingenuity pathway analysis
i-Stable	Integrated predictor for protein stability change upon single mutation
IUPred2A	Intrinsically unstructured/disordered proteins prediction
LoF	Loss-of-function
LOF	Loss of function
M	Motif
MA	Mutation assessor
MDS	Molecular dynamics simulation
ModPred	Modification prediction
MoRFs	Molecular recognition features
MPQS	ModPipe quality score
Mupro	MutationsProtein
NCBI	National Center for Biotechnology Information
NLS	Nuclear localization sequence
NMA	Normal mode analysis
nsSNPs	Nonsynonymous single nucleotide polymorphisms
PANTHER	Protein ANalysis THrough Evolutionary Relationships
PBD	PDZ-binding domain
PBM	PIP2-binding motif
PDB	Protein data bank
PDB ID	Protein data bank identification
PEST domain	Proline (P), glutamic acid (E), serine (S), and threonine (T)
PhD-SNPg	Predicting human deleterious SNPs in human genome
Pmut	Pathology of mutations
PolyPhen-2	Polymorphism phenotyping v2
PPI	Protein–protein interaction
ProjectHOPE	Project Have Our Protein Explained
PROVEAN	Protein Variation Effect Analyzer
PTMs	Post-translational modification
QMEAN	Qualitative model energy analysis
QSQE	Quaternary structure quality estimate
RCSB	Research Collaboratory for Structural Bioinformatics
RNA	Ribonucleic acid
SANT	Switching-defective protein 3, adaptor 2, nuclear receptor co-repressor, transcription factor IIIB
SAV	Splice affecting variants
SFARI	Simons Foundation Autism Research Initiative
SIFT	Sorting intolerant from tolerant
SNAP2	Screening for non-acceptable polymorphisms 2
SNF2	Sucrose NonFermentable2
SNP	Single nucleotide polymorphism
SNPs&GO	Single nucleotide polymorphism database and gene ontology
TF	Transcription factor
TFBS	Transcription factor-binding sites
UTRs	3´, 5´ Untranslated regions

1 Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by social and communication deficits with repetitive, restricted behaviours and interests. The genetic aetiology of ASD is significantly influenced by rare de novo and common inherited variants (Krumm et al. 2014; Michaelson et al. 2012). Several studies have accumulated strong evidences on the genetic burden of ASD, leading to the identification of recurrently mutated high-risk-conferring ASD genes. One such gene with the highest de novo loss-of-function (LoF) mutation rates in ASD encodes the gene *chromodomain helicase DNA-binding protein 8* (*CHD8*) protein that regulates gene expression through chromatin remodelling (O'Roak et al. 2012, 2014; Guo et al. 2018; Wade et al. 2019; Satterstrom et al. 2020). Mutations in gene *CHD8* produced a broad range of phenotypes, including ASD, macrocephaly, facial deformities, intellectual disability (ID), gastrointestinal (GI) disorders and cancers (Barnard et al. 2015).

Chromatin remodelling enzymes are crucial for the accurate organization of genomic DNA within chromatin. There are two classes of enzymes: one that mediates post-translational histone modifications and the other utilizes the energy derived from ATP hydrolysis to alter the histone–DNA contacts within the nucleosome (Marfella and Imbalzano 2007). The family of ATP-dependent chromatin remodellers is characterized by two signature sequence motifs: the tandem chromodomains in the N-terminal end that enables histone binding (Wade et al. 2019) and sucrose nonfermentable2 (SNF2)-like ATP-dependant helicase (ATPase) domain (Micucci et al. 2015). Protein CHD8 belongs to subfamily III (CHD6-CHD9) with additional functional motifs—Brahma and Kismet (BRK) domains, a switching-defective protein 3, adaptor 2, nuclear receptor co-repressor, transcription factor IIIB (SANT-like) domain, helicase C-terminal and a CHD7-binding motif (Marfella and Imbalzano 2007). The DNA-binding SANT and SLIDE domain functions as a histone-binding module and confers nonspecific DNA binding, particularly to the linker DNA between nucleosomes (Micucci et al. 2015).

Expression studies revealed that *CHD8* gene mutations indirectly down-regulated gene expression in pathways involving neurodevelopment (Sugathan et al. 2014). Mouse knockdown models of gene *CHD8* resulted in defective neuronal progenitor cell (NPC) proliferation and differentiation, causing abnormal neuronal morphology and behaviours in adult mice. Gene *CHD8* disrupted the expression of key transducers in Wnt signalling pathway, crucial for the correct balance between NPC proliferation and differentiation (Durak et al. 2016). Gene *CHD8* is highly expressed in neurons, but at low levels in glial cells of humans and mice, and plays an essential role in dendritic and axon development and migration of cortical neurons (Xu et al. 2018). Reduced *CHD8* expression led to profound alterations in both excitatory and inhibitory synaptic transmission, leading to impaired excitatory:inhibitory balance (Ellingford et al. 2020). Thus, these multi-layered evidences have rightly prompted the categorization of gene *CHD8* as a master regulator of the foundational pathways in neurodevelopment and ASD (Barnard et al. 2015).

To date, only one study by An et al. (2020) described the domain-wise mutational landscape of gene *CHD8* across three different populations—SD, cancer and the general population. However, they relied on just one parameter for variant prioritization, i.e. effect prediction score. Considering the immense genetic burden appended by gene *CHD8* on ASD manifestation, we performed a comprehensive computational analysis involving mutational density mapping across protein CHD8, followed by mutational sensitivity analysis through SNP effect prediction, protein stability change prediction, post-translational modification and evolutionary conservation analysis, deleterious mutation cluster analysis, protein homology modelling and protein dynamics study in an attempt to decipher the specific roles of ASD-associated *CHD8* variations.

2 Materials and methods

2.1 Single nucleotide polymorphisms (SNPs) and protein data collection

All *CHD8* SNPs in the general population were retrieved from the database of SNPs (dbSNP), Ensembl, Exome Variant Server (EVS), Exome Aggregation Consortium (ExAC) and Genome Aggregation Database (gnomAD) (Karczewski et al. 2019). ASD-specific genetic variations were extracted from Simons Foundation Autism Research Initiative (SFARI) repository (Banerjee-Basu and Packer 2010). All common and de novo variants were included. Regulatory SNPs (splice-site, 3´ and 5´ UTR SNPs), intronic and inframe SNPs in non-canonical transcripts were excluded; SNPs within coding regions like nonsynonymous SNPs (nsSNPs)/missense SNPs and truncating SNPs (frameshift deletion/insertion, stop gain/loss) were included for analysis. The nsSNPs were subjected to pathogenicity predictions to identify the most damaging nsSNPs, while truncating variations were all considered as loss of function (LOF) SNPs. CHD8 transcripts and corresponding protein IDs were collated using NCBI, Ensemble and UniProt database. The protein domains were predicted using the tool InterPro.

2.2 nsSNP effect and protein stability predictions

For a holistic evaluation of the consequences of nsSNPs on protein function, they were analysed on different prediction tools built on varying principals like evolutionary conservation and structure-based information. A total of ten tools were utilized to determine if an nsSNP was deleterious/damaging (D) or tolerant/benign/neutral (N), which aided in their uniform categorization. Subsequently, only those nsSNPs determined as D by $\geq 90\%$ tools providing results were designated as deleterious nsSNPs and were subsequently subjected to protein stability change prediction on three different tools. A stability change (DDG) value < -1.0 across all tools was used to identify the most destabilizing SNPs (D). In general, if the energy changes $\Delta\Delta G$ value was positive, the mutation increased stability and was classified as neutral. If the $\Delta\Delta G$ value was negative, the mutation was destabilizing and classified as deleterious (Cheng et al. 2006). Detailed descriptions are available in Supplementary Material.

2.3 Evolutionary conservation analysis

ConSurf, a Web-based tool, was used to analyse the evolutionary conservation of amino acid substitutions within a protein. The results were interpreted in the form of normalized conservation score ranging from highly conserved to the least conserved amino acid at a particular position of the protein (Glaser et al. 2003). ConSurf also provided information on the residue's location within the protein as either exposed (e) or buried (b). The total number of conserved and non-conserved residues within each domain and non-domain regions was counted to arrive at the most conserved region of protein CHD8. Two-tailed Fisher's exact test of independence was used to determine the dependencies between the set of conserved and variable residues.

2.4 Post-translational modification (PTM) prediction

Modification Prediction (ModPred) was used to predict 23 different kinds of PTMs on a unified platform (Pejaver et al. 2014). Only those PTMs predicted with high and medium confidence were considered. The total number of PTMs within and outside domains were calculated and tested for its statistical significance as described above.

2.5 Intrinsically disordered protein regions (IDPRs) and molecular recognition features (MoRFs) prediction

Protein structure disorder and disorder function prediction tools DisorderEd PredictIon CenTER (DEPICTER) (Barik et al. 2020) and intrinsically unstructured/disordered proteins prediction tool (IUPred2A) were used. IUPred2A returns a score between 0 and 1 for each residue, corresponding to the probability of the given residue being part of a disordered region (Mészáros et al. 2018). For IDR and binding site predictions, an average cutoff scores of ≥ 0.7 and ≥ 0.9, respectively, were employed. MoRFs of length 5–25 residues were predicted with consensus across three tools including IUPred2A, MoRFchibi SYSTEM(Malhis et al. 2016) and Molecular Recognition Feature predictor (MoRFPred)(Disfani et al. 2012). Stringent cutoffs were set to emulate the best combined predictions.

2.6 Mutation cluster analysis

Two clustering tools were used to identify plausible clustering of pathogenic nsSNPs within CHD8 using Mutation3D and by manual segregation method. The tool Mutation3D auto-selected suitable PDB source for the input uniport protein to perform 3D clustering on input amino acid substitutions. It was based on complete-linkage clustering that used the coordinates of α-carbons in the protein backbones from models and crystal structures to compute the statistical significance (*P* value) of the discovered clusters (Meyer et al. 2016). The second method involved counting of damaging nsSNPs across all six CHD8 signature regions and nine MoRFs identified, comparing them with variations located outside these signature motifs and identifying a probable increased aggregation of pathogenic variants within signature regions. Subsequently, the statistical significance of these occurrences was tested using Fisher's two-tailed exact test (*P* value).

2.7 Analysis of physiochemical changes due to amino acid substitutions

Project Have Our Protein Explained (Project HOPE) is an automatic mutant analysis server that provides an insight into the physiochemical structural features of the native and variant amino acid. When input with protein sequence and mutant variants, Project HOPE server predicted structural variation between mutant and wild-type residues (Venselaar et al. 2010).

2.8 Protein–protein interaction (PPI) network construction

Ingenuity Pathway Analysis (IPA) software [IPA®, QIAGEN Redwood City]: The interacting partners of protein CHD8 were identified using IPA which enabled the construction of pathways around a single molecule in the context of its PPIs, protein–DNA, protein–RNA, RNA–RNA and RNA–DNA interactions within the organism, tissue and cell lines of interest. Only direct, experimentally observed, high-confidence and predicted molecular interactions involving all upstream and downstream genes measured in neuronal tissues were consulted for network building. Prominently, only specific developmental, neurological, psychological, hereditary, metabolic, connective tissue, skeletal and muscular disorders in ASD subjects were chosen for PPI network construction as in Ashitha and Ramachandra (2020). Additionally, molecular functions common to protein CHD8 interacting partners were identified through IPA and the gene-set enrichment analysis (GSEA) tool—EnrichR (Kuleshov et al. 2016).

2.9 Protein 3D modelling

SWISS-MODEL was utilized for protein homology modelling. For an input sequence, it performed a template search through BLAST and HHblits methods, ranked available templates based on global model quality estimate (GMQE) and quaternary structure quality estimate (QSQE) scores and generated a 3D model using ProMod3 modelling engine, which resolved unfavourable interactions or clashes introduced during the modelling process by energy minimization. SWISS-MODEL returned multiple predicted models whose quality was estimated using the GMQE score, i.e. ranging between 0 and 1 (higher value indicated higher reliability) and by qualitative model energy analysis (QMEAN) Z-scores, which was an estimate of the "degree of nativeness" of the modelled structure. QMEAN Z-scores around 0 indicated good agreement between the model structure and experimental structures of similar size (Waterhouse et al. 2018).

2.10 Protein dynamics analysis

Dynamut was employed to evaluate the conformational fluctuations caused by pathogenic nsSNPs and their effects on protein's dynamic motions. For stringency, only normal mode analysis (NMA)-based ENCoM scores DDG < − 0.5 were considered and delta vibrational entropy (DDS) scores > 0.5 were assigned as molecular flexibility increasing variants, whereas DDS < − 0.5 was predicted to increase molecular rigidity due to its decreased flexibility.

3 Results

3.1 N- and C-terminal exons hosted the highest truncating SNPs and nsSNPs, respectively

A total of 84,073 SNPs were collected from four databases (Fig. 1A, Supplementary Table S1). Within the general population, 1097 SNPs were retained after removing duplicates (Table 1). The general population accumulated more nsSNPs, whereas the ASD population reported a higher occurrence of truncating variations (52%) (Fig. 1B). Only 23 nsSNPs and 12 stop-gain SNPs were common between both general and ASD population, and 76.6% of ASD variations were unique (Table 1, Fig. 1C) including 3 truncating SNPs that were recurrently mutated in ≥ 2 unrelated ASD subjects (Supplementary Table S2).

To measure the relative abundance of SNPs across CHD8 exons and domains, they were mapped to their respective regions. The C-terminal region of protein CHD8 recorded higher frequency of variations (primarily nsSNPs), especially within the CHD7-binding/FAM124B-interacting region and BRK domain corresponding to exons 29 to 37. Exon 30 hosted the highest density of variations (73.24%). Truncating SNPs were common in N-terminal signature regions, including chromo, helicase ATP-binding and SNF2_N domain. Exons 17–20 encoding the helicase C-terminal region showed the lowest density of variations, followed by helicase ATP-binding and SNF2_N domains

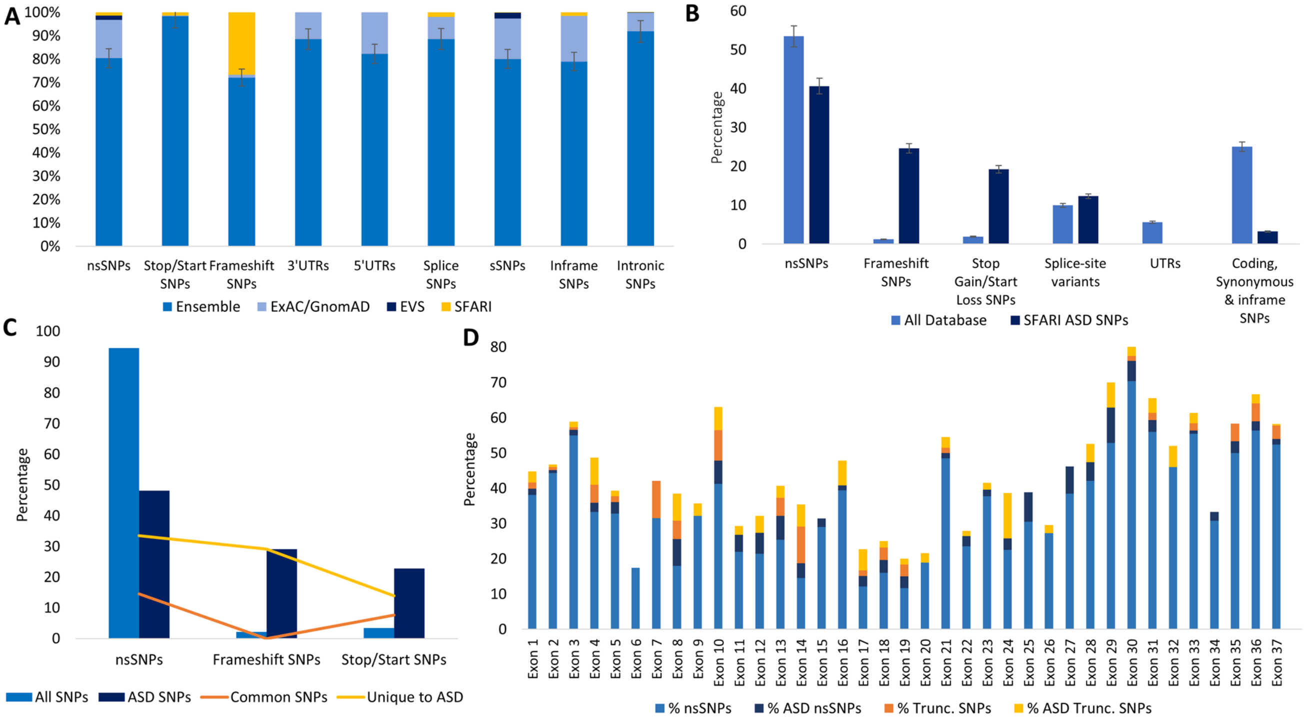

Fig. 1 **A** Comparison of SNPs in *CHD8* collected across databases such as Ensemble, ExAC/GnomAD, EVS and SFARI. Ensemble provided the highest variations, followed by GnomAD, whereas EVS had the least count. SFARI database had the highest percentage of truncating variations. **B** Frequency of different SNPs in the general population versus the ASD population. 53.45% of all variations identified in *CHD8* in the general population were nsSNP—the most common. However, truncating SNPs were the highest recorded variants within the ASD population. **C** Comparison of coding SNPs in the general vs ASD population. 94.5% of all variants collected in the general population were nsSNPs and truncating SNPs formed just 5.4%, whereas the ASD population had 51.89% truncating SNPs. 23 (27%) and 12 (35.7%) of ASD nsSNPs and stop-gain variants were common to both populations, whereas all frameshift variations identified in the ASD population were unique. **D** Exon-wise SNP density. Exon 30 recorded the highest SNP density, exon 6 had the lowest count of only nsSNPs from the general population, and exon 14 had the highest truncating SNPs. Exon 10 displayed the highest SNP density within the N-terminal region, and C-terminal exons 29–37 recorded higher SNPs except exon 34

compared to the N- and C-terminal regions that contained higher counts of SNPs (Fig. 1D and Fig. 2, Supplementary Table S3). The ASD population displayed higher density of SNPs within the core domains of protein CHD8.

3.2 Most deleterious nsSNPs were localized within CHD8 core domains: terminal regions contained benign nsSNPs

Only 135 out of 1037 nsSNPs (13%) from the general population were predicted to be deleterious. The highest density of such deleterious nsSNPs (> 34%) was found within helicase ATP-binding, SNF2_N (exons 11–15), followed by helicase C-terminal (exons 17, 18) and exons 19, 20. Two secondary peaks were observed in exons 24 and 30 that encoded a portion of SANT- and CHD7-binding region (Table 3, Fig. 3). Interestingly, nsSNPs within the N-terminal region, CHD7-binding site, BRK domain and C-terminal region recorded the highest count of nsSNPs, but were mostly benign. Among the 76 nsSNPs in the ASD population, 27 nsSNPs were predicted to be highly deleterious. Supporting the mutational patterns observed within the general population, ASD nsSNPs in the helicase C-terminal (exons 16–20), helicase ATP-binding and SNF2_N domains (exons 11, 13 and 14) and additionally exons 24 and 29 in SANT and CHD7-binding region were predicted to be deleterious than those nsSNPs located in non-domain regions (Fig. 3, Supplementary Tables S4–S8).

3.3 Helicase C terminal (exons 17–20) comprised the most destabilizing nsSNPs

All deleterious nsSNPs were further tested for their ability to cause protein stability change. A total of 51 moderate and 37 severely destabilizing nsSNPs were identified in the general population, of which only R912C and E1264K were common to both general and ASD populations (Table 2). Among the 27 deleterious ASD nsSNPs, 11 and 12 nsSNPs were severely and moderately destabilizing, respectively. The most deleterious and destabilizing nsSNPs were localized within helicase C-terminal, encoded by exons 17 to 20, followed by helicase ATP binding, SNF2_N. This trend was mirrored by nsSNPs found only in the ASD population. Combined, this study identified 48 severely damaging nsSNPs passing all thresholds of stringency (Table 2, Fig. 3, Supplementary Tables S4–S8). This pattern remained the same when a lower cutoff of DDG < – 0.5 was applied (Table 3).

Table 1 Summary of coding SNPs within the three longest transcripts of gene *CHD8**; comparison of the general and ASD populations

	All database		ASD SNPs		Common ASD SNPs			Unique ASD SNPs		
SNP types	# SNPs	%	# SNPs	%	# SNPs	% of ASD SNPs	% of all SNPs	# SNPs	% of ASD SNPs	% of all SNPs
nsSNPs	1037	94.53	76	48.11	23	14.55	3.2%	53	33.54	11%
Frameshift SNPs	23	2.10	46	29.12	0	0		46	29.11	
Stop/start gain/loss SNPs	37	3.37	36	22.79	12	7.6		22	14	
Total	1097		158		35			121		

*Transcript IDs: ENST00000399982.2/NM_001170629; ENST00000430710.3/NM_020920; ENST00000557364.1

Protein IDs: NP_001164100; NP_065971.2; Q9HCK8

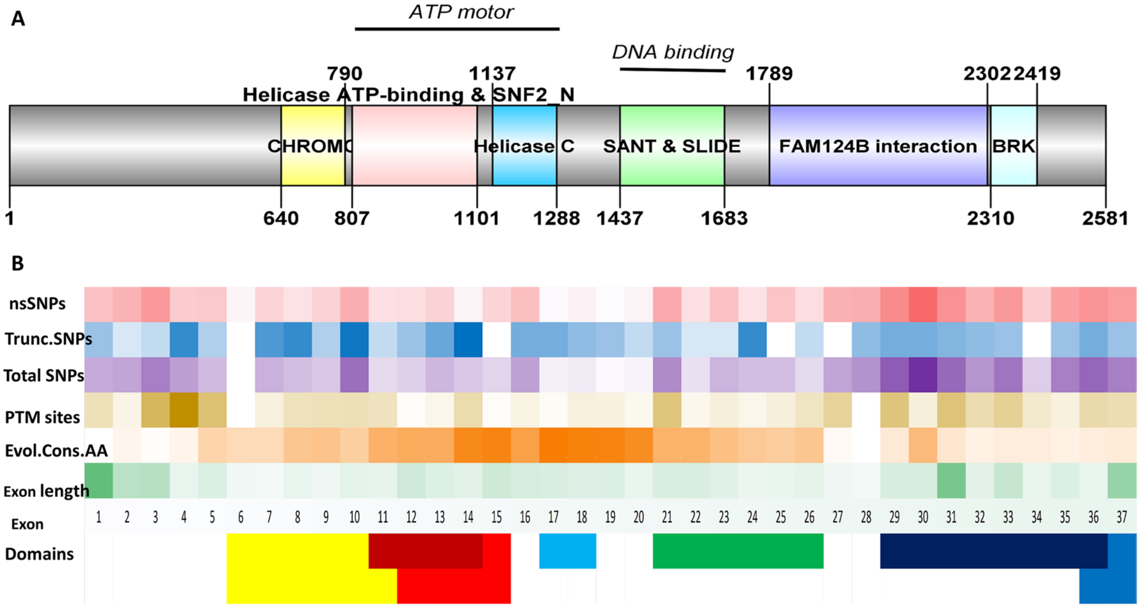

Fig. 2 **A** The longest protein sequence of CHD8 was identified to be 2581 amino acid in length, coded by mRNA transcript NM_001170629/ENST00000399982.2 composed of 37 exons, encoding protein ID NP_001164100/Q9HCK8. The protein CHD8 contains six important domains—chromo domain (640-790amino acid) is represented in yellow, helicase ATP-binding (807-1009amino acid in maroon/pink), SNF2_N (825–1101 amino acid in red/pink), helicase C-terminal (1137–1288 amino acid in light blue) and BRK domain (2310-2419amino acid in sky blue), DNA-binding site SANT and SLIDE (1437-1683amino acid in green) and a region between 1789 and 2302amino acid that binds to CHD7 and interacts with FAM124B (CHD7_BD, interaction with FAM124B) indicated in navy blue. **B** Heatmap representing exon-wise comparison of SNP density. nsSNPs were clustered within the C-terminal exons, including exons 2, 3, 10 and 21. Truncating SNPs often localized within the N-terminal exons, specifically exons 8, 10 and 14. The lowest SNP density was observed in exons 17–20 corresponding to the most conserved region of CHD8. Residues within the N-terminal exons 1–4 and C-terminal exons 31–37 were evolutionarily the most variable. Exons 3–5 contained the highest accumulation of PTMs, followed by exons 31, 29 and 21

3.4 Exons 14–20 encoding core CHD8 domains were the most evolutionarily conserved domains

The tool ConSurf provided a normalized evolutionary conservation score for each CHD8 protein residue, indicating their evolutionary status. This facilitated the identification of both evolutionarily conserved and variable residues, in addition to their relative positions on the protein structure. The helicase C-terminal was determined to be the most conserved region of protein CHD8, followed by SNF2_N and helicase ATP-binding domain corresponding to exons 14–20. Conversely, residues of exons 1–5 encoding the N-terminal region and the C-terminal exons encoding CHD-binding and BRK domain were highly variable in nature (Table 3, Figs. 2B, 3B, Supplementary Table S8).

3.5 CHD7-binding region had the highest PTM sites

A total of 311 PTM residues were identified within protein CHD8 (Q9HCK8), of which 86 and 79 phosphorylation and carboxylation sites were recognized, respectively, followed by 28, 20 and 17 acetylation, methylation and ubiquitination sites, respectively. Though PTMs were found throughout the protein, a higher aggregate was observed in regions outside the domain consisting of evolutionarily variable residues. CHD7-binding domain had the highest accumulation of PTM sites—especially exon 31 and subsequently exons 29 and 22, followed by the region between SANT and CHD7 binding (exons 27, 29) and the C-terminal tail (exons 34–47) (Table 3, Figs. 2B, 3B, Supplementary Table 9).

3.6 CHD8 is a highly disordered protein laden with nine high-confidence MoRFs

Our analysis identified that CHD8 is an intrinsically disordered protein (IDP). For reliable identification, we set propensity score cutoffs of ≈ ≥ 0.7 for tools MoRFCHiBi and IUPred2A, but selected a probability score of ≈ ≥ 0.4 for tool MoRFPred relative to the first two tools. Two distinct IDRs were detected at the N-terminal (1–600 amino acid) and C-terminal regions (2500–2570 amino acid) of CHD8 separated by exceptionally ordered, evolutionarily conserved domain region (Fig. 4). A total of nine high-confidence MoRF sites and seven disordered

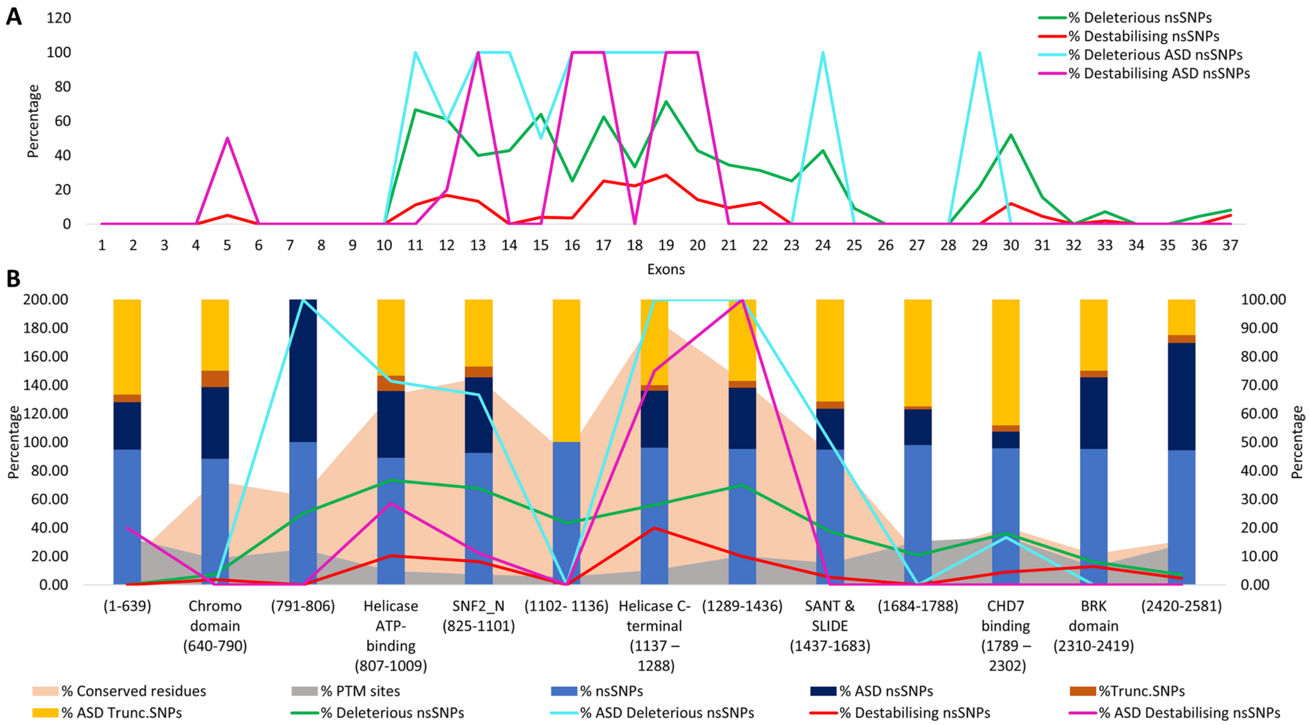

Fig. 3 Exon- and domain-wise distribution of SNPs across the general and ASD population represented in shades of blue (nsSNPs) and yellow (truncating SNPs) against the backdrop of evolutionary status of CHD8 residues (light pink area) and PTM sites (grey area) across exons in (**A**) and domain in (**B**)

binding sites were predicted within these two IDRs with consensus across tools (Table 4, Fig. 4).

Compositional bias between disordered and ordered residues was analysed. While no significant differences were observed among nonpolar residues, polar amino acids proline and serine were the most common residues within the disordered regions. An overall significant depletion in aromatic and positively charged amino acids and enrichment of polar uncharged amino acids were seen within the disordered regions (Supplementary Fig. S1). PTMs and IDRs commonly coincided—36% residues within IDRs also had PTM sites. The tool DISPHOS detected 84 PTM residues within these terminal IDRs, only 34 PTM sites contained nsSNPs and just 1 nsSNP (S1759G) was predicted to be damaging. Additionally, these IDRs were found to be prominent sites for DNA and protein binding (Fig. 5A).

3.7 Cluster analysis reveals several key characteristics of CHD SNPs

Mutation cluster analysis of the final 48 severely damaging nsSNPs identified 4 statistically significant mutations clusters. It was visualized within the PDB model 3mwy, selected by Mutant3D to evaluate the spatial arrangements of these variants (Supplementary Table S10A and Figure S2). The first significant mutation cluster included residue numbers 1051, 1264, 1325 and 1333, located around SNF2_N and helicase C-terminal domains. Two additional clusters were identified within the helicase ATP-binding and SNF2_N domains involving residues 834, 865, 952, 991 and 861, 920, 943 (Fig. 5B, and Supplementary Fig. S2), indicating that these three domains were central to the efficient functioning of protein CHD8.

Additionally, we looked for patterns of association between severely deleterious and destabilizing variations, evolutionarily conserved and variable residues and PTM sites based on their locations within or outside protein domains. Our analysis revealed a significant difference in the occurrence of truncating SNPs between the general and ASD population (*P* value 0.0001). The general population was enriched with nsSNPs (Supplementary Table S10B). Residues within domains hosted severely deleterious amino acid substitutions than residues outside (*P* value 0.0001). However, nsSNPs localization within domains was mostly stabilizing in nature (Supplementary Table S10D). Evolutionarily

Table 2 SNP effect analysis and protein stability change predictions identified 48 severely damaging nsSNPs in *CHD8*

Chr position	Ref/ Alt	rsID	aa cords	Domain	SNP effect > 90%	Protein Stability DDG < – 1.0 prediction (Value)			Sources
					%	I-Mutant	Mupro DDG	iStable (Confidence Score)	
21,869,215	C/G	rs1293008333	D1397H		90.9	D(– 1.31)	D(– 1.21)	D(0.66)	All DB
21,862,621	T/C	rs773818606	D1805G	CHD7_binding, Interaction with FAM124B	90.9	D(– 1.32)	D(– 1.40)	D(0.87)	All DB
21,870,587	C/T	rs1389713229	E1264K	Helicase C-terminal	100	D(– 1.17)	D(– 1.03)	D(0.91)	Both
21,862,550	A/C	rs767254646	F1829V	CHD7_binding, Interaction with FAM124B	100	D(– 1.99)	D(– 1.39)	D(0.81)	All DB
21,875,163	A/G	rs1248698098	F920S	Helicase ATP-binding, SNF2_N	100	D(– 2.22)	D(– 1.88)	D(0.83)	All DB
21,870,518	C/A	rs1200201759	G1287C	Helicase C-terminal	100	D(– 1.52)	D(– 1.05)	D(0.81)	All DB
21,870,180	A/T	rs948525922	I1333N		100	D (– 1.68)	D (– 2.13)	D (0.82)	All DB
21,862,285	A/G	rs1272412242	I1890T	CHD7_binding, Interaction with FAM124B	90.9	D(– 1.85)	D(– 3.16)	D(0.86)	All DB
21,876,607	A/G	rs1467509220	I865T	Helicase ATP-binding, SNF2_N	90.9	D(– 2.49)	D(– 1.59)	D(0.81)	All DB
21,871,331	G/A	rs540325439	L1187F	Helicase C-terminal	90.9	D(– 1.20)	D(– 1.19)	D(0.79)	All DB
21,862,295	G/C	rs369825360	L1887V	CHD7_binding, Interaction with FAM124B	90.9	D(– 1.21)	D(– 1.27)	D(0.85)	All DB
21,862,219	A/G	rs768411068	L1912S	CHD7_binding, Interaction with FAM124B	90.9	D(– 1.94)	D(– 1.72)	D(0.77)	All DB
21,854,301	A/G	rs546916768	L2406S		90.9	D(– 2.57)	D(– 1.42)	D(0.88)	all DB
21,871,297	G/C	rs778266688	P1198R	Helicase C-terminal	90.9	D (– 1.11)	D (– 1.08)	D (0.8)	All DB
21,869,208	G/C	rs771856418	P1399R		100	D(– 1.22)	D(– 1.41)	D(0.84)	All DB
21,862,325	G/A	rs755813740	P1877S	CHD7_binding, Interaction with FAM124B	90.9	D(– 1.71)	D(– 1.11)	D(0.78)	All DB
21,854,292	G/T	rs375361952	P2409H		100	D(– 2.02)	D(– 1.14)	D(0.76)	all DB
21,854,293	G/A	rs1027979929	P2409S		90.9	D(– 2.10)	D(– 1.06)	D(0.75)	All DB
21,876,620	G/T	rs1392213269	P861T	Helicase ATP-binding, SNF2_N	100	D(– 1.77)	D(– 1.49)	D(0.77)	All DB
21,869,187	C/T	rs770193381	R1406H		90.9	D(– 1.43)	D(– 1.49)	D(0.89)	all DB
21,868,725	G/A	rs1307220437	R1473C		90.9	D(– 1.20)	D(– 1.08)	D(0.81)	All DB
21,868,658	C/T	rs376523446	R1495H		90.9	D(– 1.38)	D(– 1.31)	D(0.83)	All DB
21,862,552	C/T	rs199908540	R1828H	CHD7_binding, Interaction with FAM124B	90.9	D(– 1.39)	D(– 1.40)	D(0.76)	All DB
21,862,139	G/C		R1939G	CHD7_binding, Interaction with FAM124B	90.9	D(– 1.31)	D(– 1.75)	D(0.87)	All DB
21,862,138	C/T	rs751815253	R1939H	CHD7_binding, Interaction with FAM124B	90.9	D(– 1.19)	D(– 1.29)	D(0.84)	All DB

Table 2 (continued)

Chr position	Ref/Alt	rsID	aa cords	Domain	SNP effect > 90%	Protein Stability DDG < – 1.0 prediction (Value)			Sources
					%	I-Mutant	Mupro DDG	iStable (Confidence Score)	
21,860,965	G/C	rs371915075	R2158G	CHD7_binding, Interaction with FAM124B	100	D (– 1.45)	D (– 1.19)	D (0.85)	All DB
21,854,266	G/C	rs372825432	R2418G		90.9	D (– 1.49)	D (– 1.14)	D (0.85)	All DB
21,854,260	G/C	rs371294659	R2420G		90.9	D(– 1.64)	D(– 1.60)	D(0.83)	All DB
21,854,259	C/T	rs771590165	R2420H		90.9	D (– 1.24)	D (– 1.17)	D (0.81)	All DB
	C/T	rs769243605	R790H	Chromo domain	90.9	D(– 1.38)	D(– 1.48)	D(0.74)	All DB
21,875,188	G/A	rs776713016	R912C	Helicase ATP-binding, SNF2_N	100	D (– 1.13)	D (– 1.26)	D (0.73)	Both
21,875,094	C/T	rs773585104	R943H	Helicase ATP-binding, SNF2_N	90.9	D(– 1.42)	D(– 1.08)	D(0.81)	All DB
21,862,511	T/C	rs1000377731	S1842G	CHD7_binding, Interaction with FAM124B	90.9	D(– 1.21)	D(– 1.77)	D(0.74)	all DB
21,873,523	A/C	rs794727141	V1051G	SNF2_N	90.9	D(– 1.63)	D(– 2.61)	D(0.85)	All DB
21,862,483	A/C	rs1462738501	V1851G	CHD7_binding, Interaction with FAM124B	90.9	D(– 3.27)	D(– 1.67)	D(0.85)	All DB
21,868,662	A/G	rs952240294	Y1494H		100	D (– 1.54)	D (– 1.7)	D (0.84)	All DB
21,862,615	T/G	rs992466357	Y1807S	CHD7_binding, Interaction with FAM124B	90.9	D(– 1.57)	D(– 2.04)	D(0.75)	All DB
21,870,587	C/T	rs1389713229	E1264K	Helicase C-terminal	100	D(– 1.17)	D(– 1.03)	D(0.91)	Both
21,870,204	A/G	c.3974 T > C	F1325S		100	D(– 1.42)	D(– 1.43)	D(0.84)	Only SFARI
21,876,700	A/G	c.2501 T > C	L834P	Helicase ATP-binding, SNF2_N	100	D(– 1.93)	D(– 1.89)	D(0.88)	Only SFARI
21,873,959	A/T	c.2972 T > A	L991H	Helicase ATP-binding, SNF2_N	100	D(– 2.1)	D(– 2.00)	D(0.85)	Only SFARI
21,871,358	G/A	c.3532C > T	R1178C	Helicase C-terminal	100	D(– 1.24)	D(– 1.12)	D(0.75)	Only SFARI
21,868,658	C/T	rs376523446	R1495H		90.9	D(– 1.38)	D(– 1.31)	D(0.83)	Both
21,884,050	C/T	c.1733G > A	R578H		90.9	D(– 1.5)	D(– 1.55)	D(0.76)	Only SFARI
21,875,188	G/A	rs776713016	R912C	Helicase ATP-binding, SNF2_N	100	D (– 1.13)	D (– 1.26)	D (0.73)	Both
21,875,068	G/C	c.2854C > G	R952G	Helicase ATP-binding, SNF2_N	100	D(– 1.41)	D(– 1.65)	D(0.89)	Only SFARI
21,869,662/ 21,867,772	A/C	c.4073 T > G	V1358G		100	D(– 2.92)	D(– 1.89)	D(0.82)	Only SFARI
21,871,628	A/T	c.3502 T > A	Y1168N	Helicase C-terminal	90.9	D(– 1.3)	D(– 1.53)	D(0.66)	Only SFARI

Rows highlighted in red include nsSNPs unique to ASD population

D refers to Destabilizing; DDG values are written within brackets ()

Gen. Pop. general population, *ASD pop.* ASD population and it means that nsSNP was identified uniquely in ASD population only, 'Both' refers to general as well as ASD population

Table 3 Summary of SNPs identified across signature regions of CHD8, their evolutionarily conservation and PTM status

Domains (aa cords)	Length	Total # SNPs	# ns SNPs	%	# Trunc SNPs	%	SNP Effect (D > 90%)		SNP Effect (D < 50%)		Prot.stab. (DDG < – 1.0)		Evol. Cons residues		Evol. Variable residues			PTM residues
							# ns SNPs	%	# ns SNPs	%	# ns SNPs	%	# aa	%	# a a	%	# aa	%
(1–639)	639	286	271	94.76	15	5.24	0	0	178	**65.68**	0	0	47	7.36	592	**92.64**	106	**16.59**
Chromo domain (640–790)	151	61	54	88.52	7	11.48	2	3.70	21	38.89	1	1.85	55	36.42	96	**63.58**	14	9.27
(791–806)	16	4	4	100	0	0	1	25.00	1	25	0	0	5	31.25	11	**68.75**	2	12.50
Helicase ATP-binding (807–1009)	203	55	49	89.09	6	10.91	18	**36.73**	12	24.49	5	**10.20**	135	**66.50**	68	33.50	10	4.93
SNF2_N (825–1101)	277	80	74	92.50	6	7.50	25	**33.78**	18	24.32	6	**8.11**	201	**72.56**	76	27.44	10	3.61
(1102–1136)	35	23	23	100	0	0	5	21.74	7	30.43	0	0	16	45.71	19	54.29	1	2.86
Helicase C-terminal (1137 – 1288)	152	26	25	96.15	1	3.85	7	**28.00**	2	8	5	**20**	142	**93.42**	10	6.58	8	5.26
(1289–1436)	148	42	40	95.24	2	4.76	14	**35.00**	16	40	4	**10**	106	**71.62**	42	28.38	15	10.14
SANT and Slide (1437–1683)	247	79	75	94.94	4	5.06	14	18.67	35	46.67	2	2.67	114	46.15	133	53.85	19	7.69
(1684–1788)	105	49	48	97.96	1	2.04	5	10.42	31	**64.58**	0	0	11	10.48	94	**89.52**	16	**15.24**
CHD7_binding, Interaction with FAM124B (1789 – 2302)	514	302	289	95.70	13	4.30	52	17.99	168	**58.13**	13	4.50	104	20.23	410	**79.77**	87	**16.93**
BRK domain (2310–2419)	110	65	62	95.38	3	4.62	5	8.06	29	46.77	4	6.45	12	10.91	98	**89.09**	7	6.36
(2420–2581)	162	90	85	94.44	5	5.56	3	3.53	58	**68.24**	2	2.35	25	15.43	137	**84.57**	23	**14.20**
Total	2581		1037				135				34		851		1730		311	

nsSNPs nonsynonympusSNPs, *Trunc.SNPs* truncating SNPs, *Prot.stab.* protein stability, *Evol.cons.* evolutionarily conserved

Prominent relatively higher values are highlighted in bold

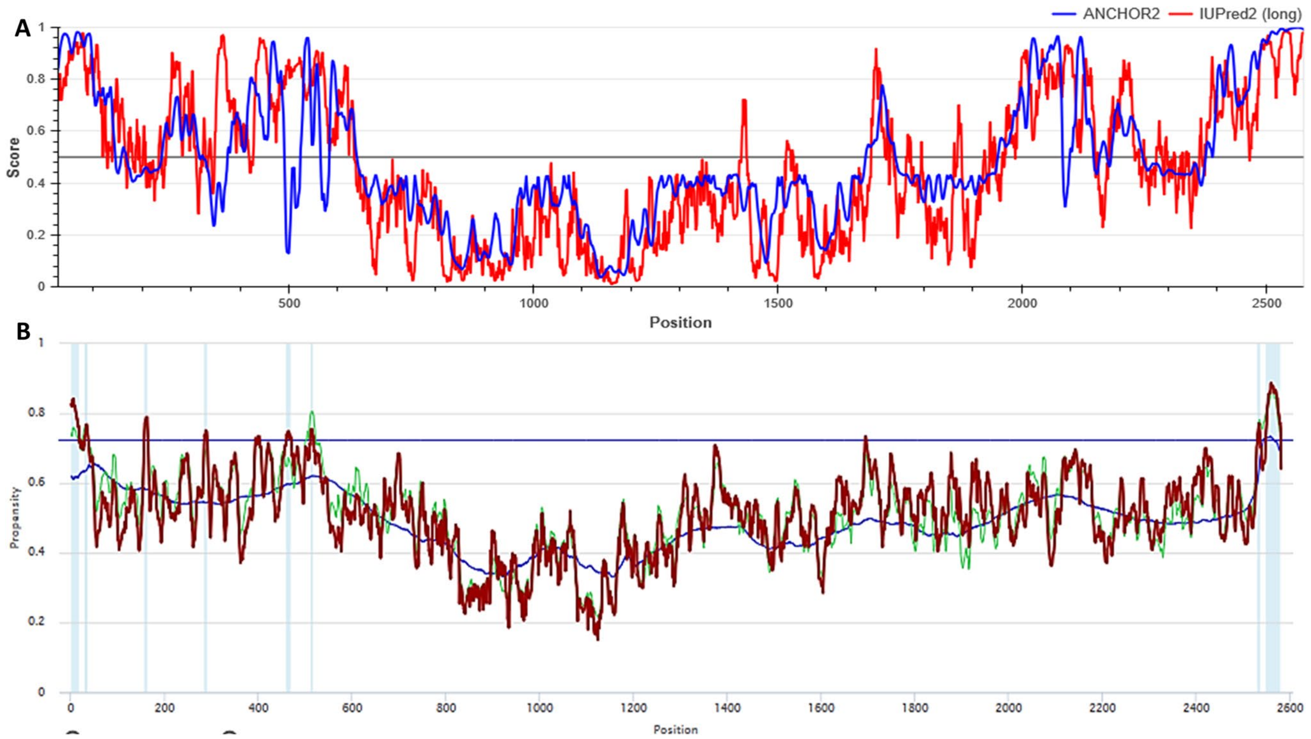

Fig. 4 Comparison of CHD8 protein disorder prediction by tools IUPred2A in (**A**) and MoRFchibi SYSTEM in (**B**). In both, each residue is plotted against its disorder probability score in the Y axis. Within (**B**), the MoRF predictions were displayed as Toggle MoRF bands in light blue colour

conserved residues were prominently segregated within signature regions (*P* value 0.0001) and remarkably PTMs were most often located outside domains (*P* value 0.0108) (Supplementary Table S10E, F).

A detailed inspection of the 9 MoRFs identified that they did not host any truncating SNPs, but contained 21 nsSNPs (2% of the general population), which were not predicted to be deleterious, but were destabilizing in nature. The ASD population did not contain any SNPs within MoRF sites.

3.8 CHD8 PPI network recapitulates common phenotypes associated with *CHD8* mutations

CHD8 was found to interact with 137 different proteins involving several cell cycle proteins and significantly enriched with DNA/RNA transcription regulation proteins, which were pooled out. An investigation for additional common molecular functions identified that the majority of these protein interactors had a neurodevelopmental role. The 13 prominent networking protein partners of CHD8, namely AGR2, CREB1, CTNNB1, CASR, CHD7, ESR2, EZH2, NR2C2, KMT2A, SMARCA1, SOX2, TNIK and TP53, were involved in the transcription of DNA/RNA (12 proteins), formation of the brain (4 proteins), gastrointestinal tract (6 proteins), body axis and long-term memory (2 proteins each) and elicited important ASD-associated phenotypes such as macrocephaly, anxiety (5 genes) and impaired social behaviour (2 proteins) (Fig. 6 and Supplementary Fig. S3A). The most critical CHD8 protein interactors identified were CTNNB1 and CREB1 found to produce five and four ASD-associated phenotypes, respectively. Eleven out of these 13 CHD8 protein interactors (84.6%) were disordered proteins. Proteins such as CASR, CHD7, KMT2A, SOX2, TNIK and TP53 were strongly disordered proteins, except ESR2 and NR2C2.

In addition, GSEA revealed that DNA transcription regulation was the single most enriched function involving 42 out of 137 (30%) interacting molecules, followed by histone methyltransferase activity and nuclear localization sequence binding. Eukaryotic transcription initiation, androgen receptor, miRNA regulation, Wnt and TGFB signalling pathways were the other prominent pathways (Supplementary Fig. S3B). This PPI network included 24 zinc finger domain-containing molecules, followed by CHD core domains containing molecules.

3.9 Protein 3D model of CHD8 core domains

Two 3D models built by SWISS-MODEL (using the template 5jxr.1.A) with 44.36% and 47.61% sequence identity

Table 4 Details of MoRFs and disordered binding site prediction in protein CHD8 with consensus across all three tools represented with their average probability/propensity scores

MoRF Predictions						Disorder Binding Prediction on ANCHOR2		
#	aa cords	MoRFCHiBi SYSTEM*		IUPred2A	MoRFPred			
		MCW	MCL			#	aa cords	Score
1	8–15	0.798	0.742	0.626	0.535	1	1–99	0.9215
2	28–45	0.713	0.731	0.781	0.315			
3	285–292	0.734	0.687	0.734	0.414			
4	394–401	0.722	0.718	0.621	0.451			
5	449–469	0.703	0.668	0.791	0.316	2	462–476	0.911
6	482–489	0.674	0.651	0.779	0.474			
7	505–517	0.712	0.776	0.840	0.400			
						3	531–543	0.9229
						4	2017–2080	0.9019
						5	2113–2129	0.9089
						6	2421–2435	0.9007
8	2528–2539	0.724	0.750	0.956	0.528	7	2468–2581	0.9612
9	2548–2570	0.846	0.841	0.858	0.412			

MoRF Propensities scores descriptions:

MoRFCHiBi_Web (MCW): an overall MoRF prediction propensity score generated by incorporating (MC) and (MDC) scores

MoRFCHiBi_Light (MCL): MoRF prediction propensity score generated by incorporating (MC) MoRF prediction and (IDP) protein disorder prediction scores. This score mainly targeted longer MoRFs

MoRFCHiBi (MC): MoRF prediction solely based on the local physiochemical properties of the amino acid sequence

MoRFDC (MDC): MoRF prediction based on the protein disorder prediction (IDP) and conservation information (ICS)

Disordered Propensity (IDP): IDP provided long trends protein disordered prediction

Conservation Propensity (ICS): ICS provided a general conservation propensity score assembled by aligning the query sequence

The predicted probability sores from IUPred2A, MoRFPred with an average cut off ≥ 0.7 and ANCHOR2 with average cut off ≥ 0.9 were used

passed the necessary quality threshold. The structure with a higher QMEAN *Z*-score (− 1.97) was finalized as the best estimated CHD8 model for residues between amino acid cordinates 800–1340 (Supplementary Fig. S4A–F). Appropriate structure templates with > 25% sequence identity were not available for the rest of the protein, likely because of their high disorder propensity, thereby limiting our downstream analysis to these modelled residues of the core CHD8 domains— helicase ATP-binding, SNF2_N and helicase C-terminal regions.

3.10 SNF2_N domain nsSNPs caused severe alterations to protein dynamic motions

A chromatin remodeller like CHD8 functions by binding DNA/proteins; hence, it is a highly dynamic protein constantly undergoing conformational changes to facilitate these interactions. A total of 131 nsSNPs, found within the modelled region of CHD8 between 800 and 1340 amino acids, were analysed on DynaMut to assess the impact of these mutations within the helicase ATPh-binding, SNF2_N and helicase C-terminal domains on protein dynamics and stability.

56 nsSNPs were predicted to be destabilizing, of which only 27 nsSNPs crossed the DDG threshold; 54 nsSNPs were found to increase molecular flexibility, but only 11 nsSNPs were above the DDS vibrational entropy cutoff > 0.5. Similarly, 33 nsSNPs increased molecular rigidity (DDS < – 0.1) and only 9 nsSNPs were above the cutoff DDS < – 0.5 (Supplementary Table S11). Overall, 28.38% of nsSNPs within the SNF2_N domain were destabilizing in nature, which was the highest. Helicase C-terminal region had more flexibility increasing variations, whereas the helicase ATP-binding domain had rigidity-increasing SNPs (Table 5). DynaMut analysed 15 out of the severely damaging nsSNPs within the modelled CHD8 structure and identified that 8 nsSNPs (including 3 ASD nsSNPs) within helicase ATP-binding, SNF2_N and helicase C-terminal domains produced strong dynamic fluctuations that altered the molecular conformation (Supplementary Table S11 and Fig. S5).

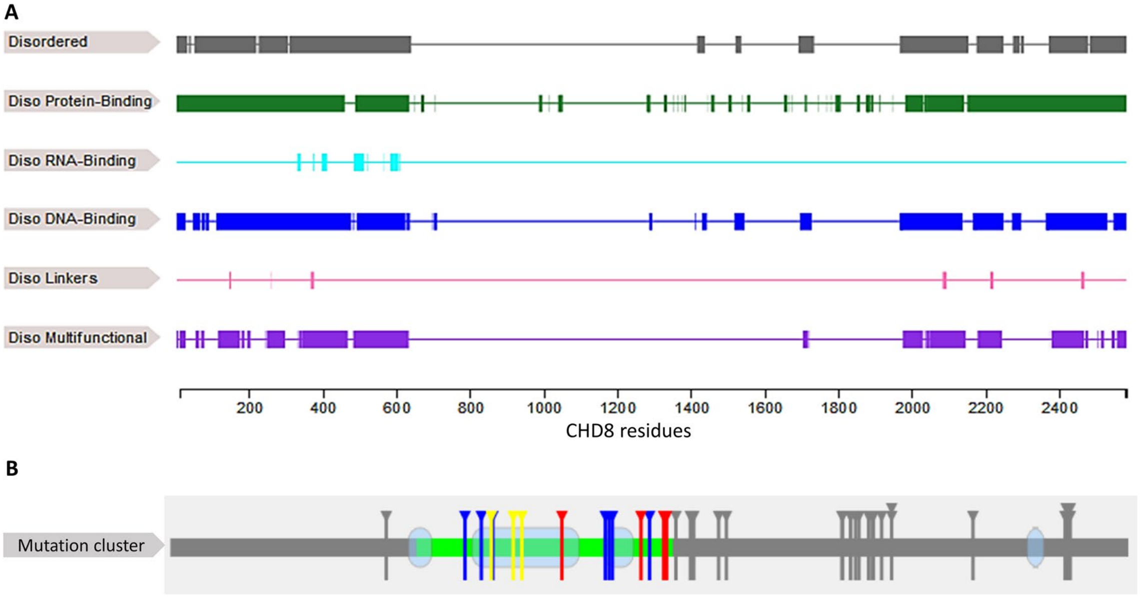

Fig. 5 **A** DEPICTER predictions of disordered regions across the protein CHD8 and its corresponding protein-binding, RNA-binding, DNA-binding, linkers and multifunctional disordered sites. **B** Mutation cluster predictions by tool Mutant 3D. The core domain regions are highlighted in fluorescent green and nsSNPs are represented as vertical pins along the CHD8 protein 2D structure. Mutations belonging to significant mutation clusters are represented in yellow and red colour code separately. Further details are available in Supplementary Fig. S2

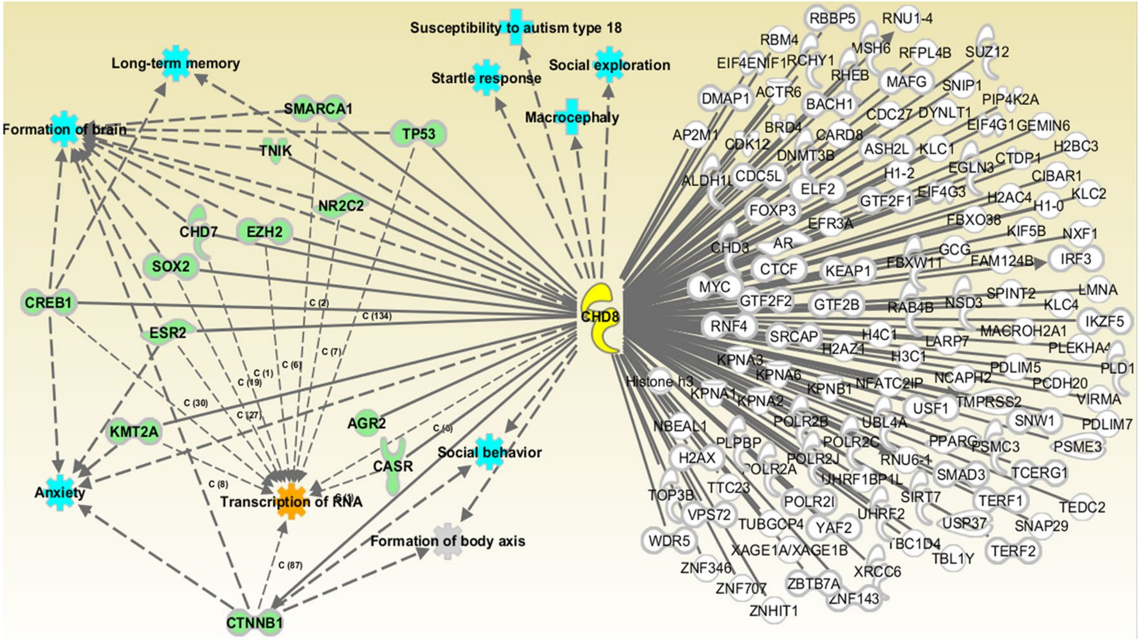

Fig. 6 Protein–protein interaction network constructed for the enzyme CHD8 (in yellow). Stringent network building rules were applied to obtain 13 direct interactions with protein partners that are represented in green. Molecular functions directly associated with ASD are presented in turquois, regulatory function in orange and others in grey

Table 5 DynaMut prediction of molecular dynamic changes caused by 131 pathogenic nsSNPs located within the core CHD8 domains

Domains (aa cords)	# Total nsSNPs	DDG Dynamut < – 0.5		Increased FLEXIBILITY Delta_vibrational entropy (DDS) > 0.5		Increased RIGIDITY, Delta_vibrational entropy (DDS) < – 0.5		Delta stability Encom DDG < – 0.5	
		#nsSNPs	%	#nsSNPs	%	#nsSNPs	%	#nsSNPs	%
Helicase ATP-binding (807–1009)	49	13	26.53	4	8.16	5	10.20	2	4.08
SNF2_N (825–1101)	74	21	28.38	5	6.76	7	9.46	3	4.05
(1102–1136)	23	0	0.00	3	13.04	1	4.35	3	13.04
Helicase C-terminal (1137 – 1288)	25	5	20.00	3	12.00	1	4.00	3	12.00
(1289–1436)	40	1	0.60	0	0.00	0	0.00	0	0.00

4 Discussion

This is the first comprehensive *in silico CHD8* gene mutational burden analysis to date. We evaluated the intrinsic mutability of gene *CHD8* in the ASD population against the backdrop of its mutational landscape within the general population. Cumulatively, nsSNPs were the most common type of variations identified frequently within exons encoding the C-terminal region of protein CHD8, whereas truncating SNPs usually occurred in the N-terminal side (highest in exon 14, 10, 8 and 24). We observed that exons 14–20 encoded the most conserved regions of CHD8 and thereby displayed the lowest SNP density, but highest sensitivity to SNPs, especially the helicase C-terminal region. Overall, nsSNPs identified within the core CHD8 domains are helicase ATP-binding, SNF2_N and helicase C-terminal regions that are severely damaging, reflecting their crucial functional roles as evolutionarily essential regions of CHD8 (Fig. 3). An auxiliary peak was observed within the CHD7-binding region, especially due to damaging variations within exon 30.

The ASD population recorded a significantly higher frequency of truncating SNPs (P < 0.0001) compared to the general population (Wilkinson et al. 2015; An et al. 2020). Although ASD variants were not localized to any specific regions of CHD8, > 30% of ASD SNPs occurred frequently within the highly conserved signature regions in contrast to the observations made in the general population. Notably, the helicase C-terminal region had frequent accumulation of truncating SNPs and severely damaging nsSNPs than the general population (An et al. 2020), followed by helicase ATP-binding and SNF2_N domains. Gene *CHD8* had recurrent ASD SNPs within the CHD7-binding motif, especially G1602Vfs*13 in SANT and SLIDE DNA-binding domain. This could lead to loss of PTM sites and alter CHD8's chromatin remodelling functions, respectively, known to disrupt protein function.

Additionally, the N- and C-terminal regions of CHD8, involving exons 1–6 and exons 27–37 encoding CHD7-binding and BRK domains respectively, contained the highest nsSNPs that were mostly benign (> 65%). Apart from being highly tolerant to variations, these regions were identified as intrinsically disordered with nine MoRF sites of < 12 amino acid length. These IDRs were evolutionarily variable, prone to higher accumulation of tolerant SNPs, especially the C-terminal end. PTMs are known to be strongly associated with IDRs. 58% of phosphorylation sites in CHD8 were within IDRs, the most common type of PTM found within IDRs (Darling and Uversky 2018). Phosphorylation mediates specific, but weak interactions with partners, and modulates the binding affinity of transcription factors to their coactivators and DNA, thereby altering the gene expression affecting cell growth and differentiation (Darling and Uversky 2018). These disordered regions of CHD8 were observed to have larger incidences of ASD-associated truncating SNPs.

An et al. (2020) utilized the Chd1 crystal structure (PDB code 5O9G) in their study and remapped gene *CHD8* mutations onto it. To study the conformational disturbances caused by nsSNPs to the dynamic motions in CHD8, we performed protein homology modelling. Only the core domains of CHD8 between 800 and 1340 residues were successfully modelled due to the unavailability of reliable 3D templates for the rest of the protein with a minimum 30% sequence similarity (Supplementary Fig. S4). Interestingly, missense variations at the core of CHD8 produced long-range fluctuations altering the global dynamic motions of this complex, not observed in residues outside these domains.

Mutations in gene *CHD8* have been consistently associated with phenotypes such as ASD, macrocephaly, ID and GI complications that were recapitulated in animal models by silencing the *CHD8* gene expression (Bernier et al. 2014; Xu et al. 2018). However, to date, limited explanations have been provided on the molecular mechanisms responsible for such comorbidities. Protein CHD8 is known to regulate gene expression through protein interactions. A study utilized both transcriptome and ChIP sequencing in human neural progenitor cells (hNPCs) and identified 1756 differentially expressed genes (DEGs) and demonstrated widespread binding to chromatin (Sugathan et al. 2014). Another study exploring transcriptional changes due to *CHD8* gene knockdown in hNSCs identified 1715 DEGs (Wilkinson et al. 2015) and SFARI database's protein interaction analysis identified 3,583 CHD8 interactors with > 100 ASD-associated genes. However, our stringent PPI analysis identified 137 protein interactors of CHD8 participating in DNA/RNA transcription regulation, formation of brain, body axis and GI tract and additionally produced ASD traits such as social behaviour, anxiety and long-term memory. We suspect that aberrant CHD8 dosage leads to altered regulation of gene expression due to cumulative changes to these molecular interactions and consequently produce ASD and comorbidities associated with CHD8 mutation which needs further investigation.

Therefore, gene*CHD8* is indeed a master regulator of neuronal and GI functions and hence a potent contributor to ASD. Our in-depth *in silico* analysis provides a blueprint of the mutational landscape and pathogenicity patterns of *CHD8*. ASD is burdened by the variations occurring within core domains and frequently occurring truncating SNPs, especially within CHD7-binding site.

5 WEBLINKS accessed before 31st October 2020

https://www.ebi.ac.uk/interpro/
http://sift.jcvi.org/
http://genetics.bwh.harvard.edu/pph2/
http://provean.jcvi.org/index.php
http://fathmm.biocompute.org.uk/inherited.html
http://snps.biofold.org/snps-and-go/snps-and-go.html
https://rostlab.org/services/snap2web/
http://snps.biofold.org/snps-and-go/snps-and-go.htm
http://bbglab.irbbarcelona.org/fannsdb/home
http://www.pantherdb
http://mutationassessor.org/r3/org/
http://folding.uib.es/i-mutant/i-mutant3.0.html
http://predictor.nchu.edu.tw/iStable/
http://mupro.proteomics.ics.uci.edu/
http://consurf.tau.ac.il/
http://montana.informatics.indiana.edu/ModPred/
http://www.cbs.dtu.dk/services/NetSurfP/
http://flexpred.rit.albany.edu/
http://biomine.cs.vcu.edu/servers/DEPICTER/
https://iupred2a.elte.hu/
https://morf.msl.ubc.ca/index.xhtml
http://biomine.cs.vcu.edu/servers/MoRFpred/
http://mutation3d.org/advanced_form.shtml
https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/analysis-and-visualization/qiagen-ipa/
https://amp.pharm.mssm.edu/Enrichr/
https://swissmodel.expasy.org/
http://biosig.unimelb.edu.au/dynamut/
http://grch37.ensembl.org/index.html
https://evs.gs.washington.edu/EVS/
https://exac.broadinstitute.org/
https://gnomad.broadinstitute.org/
http://www.ncbi.nlm.nih.gov/gene
https://gene.sfari.org/database/human-gene/
https://www.ncbi.nlm.nih.gov/
https://www.uniprot.org/
https://www.rcsb.org/structure/1d5r
http://fathmm.biocompute.org.uk/index.html
http://mmb.irbbarcelona.org/PMut
http://snps.biofold.org/snps-and-go
https://bbglab.irbbarcelona.org/fannsdb/
http://snps.biofold.org/phd-snpg/
https://loschmidt.chemi.muni.cz/predictsnp2/
http://snps.biofold.org/snps-and-go/index.html
http://mutpred.mutdb.org/about.html
http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi
http://montana.informatics.indiana.edu/ModPred/index.html
https://consurf.tau.ac.il/index_proteins.php
http://www.pantherdb.org/tools/csnpScoreForm.jsp
https://www3.cmbi.umcn.nl/hope/

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13721-022-00377-z.

Acknowledgements We thank all databases and tools included in this study for making them readily available for research use. We acknowledge the support and thank members of the Genetics and Genomics Laboratory, Department of Studies in Genetics and Genomics for their help and encouragement.

Author contributions ASNM: conceptualization, planning, design, execution and manuscript preparation; STB: execution; NBR: guidance, editing and evaluation.

Funding We immensely thank ICMR for senior-research fellowship provided to ASNM (Award: F. No. 45/28/2018-HUM/BMS). This research did not receive any specific research grant from funding agencies in the public, commercial, or not-for-profit sector.

Declarations

Conflict of interest The authors declare no competing financial interests.

Ethics approval Not required.

Consent to participate and for publication Not required.

References

An Y, Zhang L, Liu W et al (2020) De novo variants in the Helicase-C domain of CHD8 are associated with severe phenotypes including autism, language disability and overgrowth. Hum Genet 139:499–512. https://doi.org/10.1007/s00439-020-02115-9

Ashitha SNM, Ramachandra NB (2020) Integrated functional analysis implicates syndromic and rare copy number variation genes as prominent molecular players in pathogenesis of autism spectrum disorders. Neuroscience 438:25–40. https://doi.org/10.1016/j.neuroscience.2020.04.051

Banerjee-Basu S, Packer A (2010) SFARI Gene: an evolving database for the autism research community. DMM Dis Model Mech 3:133–135

Barik A, Katuwawala A, Hanson J et al (2020) DEPICTER: intrinsic disorder and disorder function prediction server. J Mol Biol 432:3379–3387. https://doi.org/10.1016/j.jmb.2019.12.030

Barnard RA, Pomaville MB, O'Roak BJ (2015) Mutations and modeling of the chromatin remodeler CHD8 define an emerging autism etiology. Front Neurosci 9:477

Bernier R, Golzio C, Xiong B et al (2014) Disruptive CHD8 mutations define a subtype of autism early in development. Cell 158:263–276. https://doi.org/10.1016/j.cell.2014.06.017

Cheng J, Randall A, Baldi P (2006) Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 62:1125–1132

Darling AL, Uversky VN (2018) Intrinsic disorder and posttranslational modifications: the darker side of the biological dark matter. Front Genet 9:158

Disfani FM, Hsu WL, Mizianty MJ et al (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics. https://doi.org/10.1093/bioinformatics/bts209

Durak O, Gao F, Kaeser-Woo YJ et al (2016) Chd8 mediates cortical neurogenesis via transcriptional regulation of cell cycle and Wnt signaling. Nat Neurosci 19:1477–1488. https://doi.org/10.1038/nn.4400

Ellingford R, de Meritens ER, Shaunak R et al (2020) Cell-type-specific synaptic imbalance and disrupted homeostatic plasticity in cortical circuits of ASD-associated Chd8 haploinsufficient mice. bioRxiv. https://doi.org/10.1101/2020.05.14.093187

Glaser F, Pupko T, Paz I et al (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 19:163–164. https://doi.org/10.1093/bioinformatics/19.1.163

Guo H, Wang T, Wu H et al (2018) Inherited and multiple de novo mutations in autism/developmental delay risk genes suggest a multifactorial model. Mol Autism. https://doi.org/10.1186/s13229-018-0247-z

Karczewski KJ, Francioli LC, Tiao G et al (2019) Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv. https://doi.org/10.1101/531210

Krumm N, O'Roak BJ, Shendure J, Eichler EE (2014) A de novo convergence of autism genetics and molecular neuroscience. Trends Neurosci 37:95–105

Kuleshov MV, Jones MR, Rouillard AD et al (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44:W90–W97. https://doi.org/10.1093/nar/gkw377

Malhis N, Jacobson M, Gsponer J (2016) MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. Nucleic Acids Res 44:W488–W493. https://doi.org/10.1093/nar/gkw409

Marfella CGA, Imbalzano AN (2007) The Chd family of chromatin remodelers. Mutat Res Fundam Mol Mech Mutagen 618:30–40. https://doi.org/10.1016/j.mrfmmm.2006.07.012

Mészáros B, Erdös G, Dosztányi Z (2018) IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res 46:W329–W337. https://doi.org/10.1093/nar/gky384

Meyer MJ, Lapcevic R, Romero AE et al (2016) mutation3D: cancer gene prediction through atomic clustering of coding variants in the structural proteome. Hum Mutat 37:447–456. https://doi.org/10.1002/humu.22963

Michaelson JJ, Shi Y, Gujral M et al (2012) Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell 151:1431–1442. https://doi.org/10.1016/j.cell.2012.11.019

Micucci JA, Sperry ED, Martin DM (2015) Chromodomain helicase DNA-binding proteins in stem cells and human developmental diseases. Stem Cells Dev 24:917–926

O'Roak BJ, Vives L, Fu W et al (2012) Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science (80-) 338:1619–1622. https://doi.org/10.1126/science.1227764

O'Roak BJ, Stessman HA, Boyle EA et al (2014) Recurrent de novo mutations implicate novel genes underlying simplex autism risk. Nat Commun. https://doi.org/10.1038/ncomms6595

Pejaver V, Hsu WL, Xin F et al (2014) The structural and functional signatures of proteins that undergo multiple events of post-translational modification. Protein Sci 23:1077–1093. https://doi.org/10.1002/pro.2494

Satterstrom FK, Kosmicki JA, Wang J et al (2020) Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. Cell 180:568-584.e23. https://doi.org/10.1016/j.cell.2019.12.036

Sugathan A, Biagioli M, Golzio C et al (2014) CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. Proc Natl Acad Sci U S A 111:E4468–E4477. https://doi.org/10.1073/pnas.1405266111

Venselaar H, te Beek TAH, Kuipers RKP et al (2010) Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-11-548

Wade AA, Lim K, Catta-Preta R, Nord AS (2019) Common CHD8 genomic targets contrast with model-specific transcriptional impacts of CHD8 haploinsufficiency. Front Mol Neurosci. https://doi.org/10.3389/fnmol.2018.00481

Waterhouse A, Bertoni M, Bienert S et al (2018) SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 46:W296–W303. https://doi.org/10.1093/nar/gky427

Wilkinson B, Grepo N, Thompson BL et al (2015) The autism-associated gene chromodomain helicase DNA-binding protein 8 (CHD8) regulates noncoding RNAs and autism-related genes. Transl Psychiatry. https://doi.org/10.1038/tp.2015.62

Xu Q, Liu YY, Wang X et al (2018) Autism-associated CHD8 deficiency impairs axon development and migration of cortical neurons. Mol Autism. https://doi.org/10.1186/s13229-018-0244-2

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.