Autism Spectrum Disorder is burdened by severely pathogenic variations within core domains of CHD8 and its CHD7-binding motif

Ashitha SNM[1], Suryanarayanan Balakrishnan[1], and Ramachandra Nallur[1]

[1]University of Mysore

August 6, 2020

Abstract

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder presented 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 gene that is a master regulator mediating the expression of genes controlling neuron functions. We collected 8,124 exonic SNPs in CHD8 from 4 databases representing the general and ASD populations; subjected them to multi-layered analyses on >20 computational tools. We observed that nsSNPs were common in the general population. Distinct hotspots for truncating and nsSNPs were identified within exons encoding the N and C terminals, respectively. Evolutionarily conserved regions involving CHD8 core domains: Helicase-C-terminal, Helicase-ATP-binding and SNF2_N domains, recorded the lowest density but severely pathogenic SNPs. Conversely, evolutionarily variable regions- CHD7-binding and BRK domains- hosted the highest SNPs, but were benign. Post-Translational-Modifications (PTMS) occurred on residues outside domains ($P<0.01$) i.e., non-conserved regions of CHD8 including the N and C terminals that were determined to be Intrinsically-Disordered-Protein-Regions (IDPRs) with 9 Molecular-Recognition-Features sites. Contrastingly, ASD population recorded significantly higher incidences of truncating SNPs than general population ($P<0.0001$). ASD-SNPs frequently occurring 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 PPIs effortlessly recapitulated the phenotypes presented by children with CHD8 mutations. 11/13 (84.6%) interacting molecules were IDPs. We identified 9 CHD8 nsSNPs that produced the strongest long-range disturbances, altering the modelled protein's global conformational dynamics.

Autism Spectrum Disorder is burdened by severely pathogenic variations within core domains of CHD8 and its CHD7-binding motif

Ashitha S Niranjana Murthy, Suryanarayanan Thangalazhi Balakrishnan and Nallur B Ramachandra*

Genetics and Genomics Laboratory, Department of Studies in Genetics and Genomics, University of Mysore, Manasagangotri-570006, Karnataka, India

Ms. ASHITHA S NIRANJANA MURTHY, M.Sc.

Ph.D. Student

Genetics and Genomics Laboratory,

Department of Studies in Genetics and Genomics,

University of Mysore,

Manasagangotri- 570006,

Karnataka, India

ashitha@zoology.uni-mysore.ac.in

Mr. SURYANARAYANAN THANGALAZHI BALAKRISHNAN, M.Sc.

M.Sc. Student

Department of Zoology

St Aloysius College, Elthuruth, Thrissur

Kerala - 680611

suryantb1995@gmail.com

*Corresponding Author: Dr. NALLUR BASAPPA RAMACHANDRA

Emeritus Professor and Principal Investigator Department of Studies in Genetics and Genomics, University of Mysore, Manasagangotri-570006, Karnataka, India nallurbr@gmail.com

ABSTRACT

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder presented 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 gene that is a master regulator mediating the expression of genes controlling neuron functions. We collected 8,124 exonic SNPs in*CHD8* from 4 databases representing the general and ASD populations; subjected them to multi-layered analyses on >20 computational tools. We observed that nsSNPs were common in the general population. Distinct hotspots for truncating and nsSNPs were identified within exons encoding the N and C terminals, respectively. Evolutionarily conserved regions involving CHD8 core domains: Helicase-C-terminal, Helicase-ATP-binding and SNF2_N domains, recorded the lowest density but severely pathogenic SNPs. Conversely, evolutionarily variable regions- CHD7-binding and BRK domains- hosted the highest SNPs, but were benign. Post-Translational-Modifications (PTMS) occurred on residues outside domains ($P<0.01$) i.e., non-conserved regions of CHD8 including the N and C terminals that were determined to be Intrinsically-Disordered-Protein-Regions (IDPRs) with 9 Molecular-Recognition-Features sites. Contrastingly, ASD population recorded significantly higher incidences of truncating SNPs than general population ($P<0.0001$). ASD-SNPs frequently occurring 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 PPIs effortlessly recapitulated the phenotypes presented by children with *CHD8* mutations. 11/13 (84.6%) interacting molecules were IDPs. We identified 9 *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*
CHD	CHD8 gene
aa	amino acid
ASD	Autism Spectrum Disorders
ATP	Adenosine triphosphate
BLAST	Basic Local Alignment Search Tool
BRK	Brahma and Kismet
CLS	Cytoplasmic Localization Sequence

CHD	*Chromodomain Helicase DNA-binding protein*
CONDEL	CONsensus DELeteriousness
D	Deleterious and/or Destabilizing
dbSNP	database of SNPs
DEG	Differentially Expressed genes
DEPICTER	DisorderEd PredictIon CenTER
DNA	Dioxy-ribonucleic 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 Database 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 Interactions
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

CHD	*Chromodomain Helicase DNA-binding protein*
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 & Gene Ontology
TF	transcription factor
TFBS	transcription factor binding sites
UTRs	3', 5' Untranslated Regions

INTRODUCTION

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterised by social and communication deficits with repetitive, restricted behaviours and interests. Genetic aetiology of ASD is significantly influenced by rare *de novo* and common inherited variants (Michaelson et al., 2012; Krumm et al., 2014). Several studies 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 among the highest *de novo*loss-of-function (LoF) mutation rates in ASD encodes the*Chromodomain Helicase DNA-binding protein 8* (*CHD8*) protein which regulates gene expression through chromatin remodelling (Guo et al., 2018; Wade et al., 2019; Satterstrom et al., 2020). Mutations in*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 remodeling enzymes are crucial for the accurate organization of genomic DNA within chromatin. There are two classes of enzymes: ones that mediated post-translational histone modifications and others that utilize the energy derived from ATP hydrolysis to alter the histone-DNA contacts within the nucleosome (Marfella & Imbalzano, 2007). The family of ATP-dependent chromatin remodellers is characterised 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). CHD8 belongs to subfamily III (CHD6-CHD9) with additional functional motifs- BRK (Brahma and Kismet) domains, a SANT-like (switching-defective protein 3, adaptor 2, nuclear receptor co-repressor, transcription factor IIIB) domain, Helicase-C-terminal and a CHD7-binding motif (Marfella & Imbalzano, 2007). The DNA-binding SANT and SLIDE domain functions as a histone-binding module, confers nonspecific DNA binding, particularly to the linker DNA between nucleosomes (Micucci et al., 2015).

Expression studies revealed that *CHD8* mutations indirectly down-regulated gene expression in pathways involving neurodevelopment (Sugathan et al., 2014). Mouse knockdown models of *CHD8* resulted in defective Neuronal Progenitor Cells (NPC) proliferation and differentiation, causing abnormal neuronal morphology and behaviours in adult mice. *CHD8* disrupted expression of key transducers in Wnt signaling pathway-crucial for the correct balance between NPC proliferation and differentiation (Durak et al., 2016). *CHD8* was highly expressed in neurons, but at low levels in glial cells of humans and mice, playing an essential role in dendritic and axon development and migration of cortical neurons (Xu et al., 2018). Reduced *CHD8*expression led to profound alterations to both excitatory and inhibitory synaptic transmission resulting in a reduced excitatory:inhibitory balance (Ellingford et al., 2020). Thus, these multi-layered pieces of evidence have rightly prompted the categorisation of *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 mutational landscape of *CHD8* with respect to its domains across three different populations- ASD, cancer and general population. However, they relied on just one parameter for variant prioritisation, i.e., effect prediction score. Considering the immense genetic burden appended by *CHD8* on ASD manifestation, we performed a more comprehensive mutational burden analysis with emphasis on deciphering the specific roles of ASD-associated *CHD8* variations.

2. MATERIAL AND METHODS

1. **Single Nucleotide Polymorphisms (SNPs) and protein data collection**All SNPs within *CHD8* gene from the general population were retrieved from the National Center for Biotechnology Information (NCBI)- 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 & Packer, 2010). Regulatory SNPs (Splice-site, 3' and 5' UTR SNPs), intronic and inframe SNPs in non-canonical transcripts were excluded, the remaining SNPs in the coding region like missense and truncating SNPs were retained. These nonsynonymous SNPs (nsSNPs)/missense variations were subjected to pathogenicity prediction analysis to identify the most deleterious nsSNPs; while truncating SNPs like Frameshift deletion/insertion and stop gain or loss variants were considered as Loss of Function (LOF) SNPs. The mRNA transcripts of *CHD8* , its corresponding protein IDs were identified using NCBI, Ensemble and UniProt database and were subjected to protein domain prediction on tool InterPro.
2. **SNP effect prediction analysis:**

The effect of nsSNPs on protein functioning were analysed on 10 different prediction tools built on varying principals to obtain a holistic evaluation. The combined effect of these predictions determined pathogenicity of a nsSNP. It was declared as 'Deleterious/Damaging' '(D)' only if it was predicted so by $> 90\%$ of tools that provided results as described below: -

1. **Sorting Intolerant from Tolerant (SIFT):** predicted deleterious and tolerated SNPs to characterize amino acid (aa) substitutions causing phenotypic and functional changes in the protein. For each nsSNP, SIFT provided a tolerance index score, and score [?] 0.05 was considered a deleterious variant (Sim et al., 2012).
2. **Polymorphism Phenotyping v2 (PolyPhen2):** we input the details of aa substitutions and UniProtKB accession number or FASTA sequence. If the probabilistic score was > 0.85 the mutation was 'probably damaging' if score was > 0.15, then they were assigned as 'possibly damaging' and the remaining were 'benign' according to its specificity and sensitivity values (Adzhubei et al., 2010). Both HumDiv and HumVar analysis were performed.
3. **Protein Variation Effect Analyzer (PROVEAN):** measured the sequence similarity of a query sequence to a protein sequence homolog before and after introducing an aa variation to the query sequence. A protein variant was said to be "deleterious" if the final score was below the default threshold of -2.5 or was predicted to be "neutral" if the score was above the threshold (Choi et al., 2012).
4. **Functional Analysis Through Hidden Markov Models (FATHMM):**Uses HMMs to align homologous sequences and conserved domains and predict the effects of nsSNPs. A score > 0 is Tolerated substitution and < 0 is considered Damaging (Shihab et al., 2013).
5. **Single Nucleotide Polymorphism and Gene Ontology (SNPs &GO)**was a support vector machine (SVM) to predict disease-related mutations from protein sequences. If mutation measured a probability score higher than 0.5, then the protein was considered to have disease-related effect (Calabrese et al., 2009; Thomas and Kejariwal, 2004 2003).
6. **Screening for Non-Acceptable Polymorphisms 2 (SNAP2):** was developed based on a neural network classification method that predicted the effect of nsSNPs on protein function (Hecht et al., 2013). It provided scores for each substitution and predicted neutral or non-neutral effects when provided the input with FASTA sequence and a list of mutants.

7. **Predictor of human Deleterious Single Nucleotide Polymorphisms (PhD SNP):** PhD-SNP was SVM that predicted disease-associated and neutral aa substitution using reliability index scores between 0 and 9 (Capriotti et al., 2006).
8. **CONsensus DELeteriousness (CONDEL):** evaluated the deleterious missense variants and computed the complementary cumulative scores of deleterious and neutral mutations (González-Pérez & López-Bigas, 2011).
9. **Protein ANalysis THrough Evolutionary Relationships (PANTHER):** PANTHER estimated nsSNPs which impaired proper functioning of the protein. It also calculated the preservation time of a protein; longer the preservation time greater the functional impact (Tang & Thomas, 2016).
10. **Mutation Assessor (MA):** MA predicted the functional impact of aa substitutions in proteins. It gave the functional impact score (FIS), which was derived from multiple sequence alignments (MSA) of sequence homologs. Higher scores indicated a functional impact of a mutation (Reva et al., 2011).

1. **Protein stability prediction:**Only those nsSNPs determined to be deleterious were further subjected to protein stability change analysis to identify the most destabilising variants (D) predicted with DDG value < -1.0 across the three different tools used which are as described below:-
2. **I-Mutant (version 3.0)** is based on the SVM algorithm to predict the protein's stability due to a single aa variation, using protein sequence or structure information. It predicted DDG values as a regression estimator and the sign of the stability change. I-Mutant 3.0 classified mutations into three categories: neutral mutation (-0.5 [?] DDG[?]0.5), large decrease ([?]-0.5), and large increase (>0.5) (Capriotti et al., 2005).
3. **Integrated predictor for protein stability change upon single mutation (i-Stable):** uses SVM for the prediction of protein stability changes due to single aa variation. $\Delta\Delta G > 0$ was defined as a stabilizing change and a $\Delta\Delta G$ value < 0 was defined as a destabilizing variant (Chen et al., 2013).
4. **Prediction of Protein Stability Changes for Single-Site Mutations from Sequences (MUpro):** is a set of machine learning programs to predict the effects of single-site aa mutation on protein stability. 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).
5. **Evolutionary conservation analysis**

ConSurf: a web-based tool built using empirical Bayesian inference which automatically analysed the evolutionary conservation of aa substitutions in a protein. The FASTA sequence was provided as input data. The results were interpreted in the form of normalised conservation score, which ranged from highly conserved to the least conserved aa at a particular position (Glaser et al., 2003). ConSurf also provided information on the residue's location within the protein as either expose (e) or buried (b). The total number of conserved and non-conserved or variable residues within each domain and the regions surrounding it were counted to identify the most conserved region of CHD8. Two-tailed Fisher's exact test of independence was used to determine the dependencies between the set of conserved and variable residues.

Post- Translational Modification (PTM) prediction:

Modification Prediction (ModPred) predictor of PTM sites in proteins: was used to predict 23 different PTMs on a unified platform. (Pejaver et al., 2014). The total number of PTMs within and outside domains were calculated and tested for its statistical significance as described above.

Protein surface and solvent accessibility analysis

NetSurfP server ver. 1.1 : predicts the surface accessibility and also the secondary structure of aa (Petersen et al., 2009). It provided 3 subclasses for solvent accessibility of aa: high accessibility (exposed), moderate accessibility (partially buried) and low accessibility (buried).

1. **Predicting conformation switches in proteins: FlexPred server** : predicted residues acting as conformational switches. Proteins are dynamic and flexible macromolecules. When the environment changes, the protein backbone can undergo significant conformational changes and switch from one

folded conformation to another (Kuznetsov, 2008).

2. **Intrinsically Disordered Protein Regions (IDPRs) prediction:**
3. **DisorderEd PredictIon CenTER(DEPICTER):** was an intrinsic disorder and disorder function prediction server. It included the prediction of disorder protein-binding, RNA-binding, DNA-binding, linkers and multifunctional sites (Barik et al., 2020).
4. **Intrinsically unstructured/disordered proteins prediction tool (IUPred2A): is** a combined web interface that predicts IDPRs by generating energy-estimation based predictions for ordered and disordered residues by IUPred2 and for disordered binding regions by ANCHOR2. 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 cut-off score of [?] 0.7 and [?] 0.9 respectively were employed.
5. **Molecular Recognition Features (MoRFs) prediction:**MoRFs of length 5-25 residues, were predicted with consensus across three tools described below including IUPred2A. Stringent cut-offs were set to emulate the best combined predictions:
6. **MoRFchibi SYSTEM:**predicts MoRFs by generating six propensity scores for each residue as described in (Malhis et al., 2016) and our cut-off value was set to [?] 0.7.
7. **Molecular Recognition Feature predictor (MoRFPred)** : was used to annotate MoRF residues and prediction scores, which was set to a score [?] 0.4 (Disfani et al., 2012).
8. **Mutation cluster analysis:**
9. **Mutation3D** : To identify plausible clustering of pathogenic nsSNPs within CHD8, we performed two separate analysis 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 aa 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 discovered clusters (Meyer et al., 2016).
10. **Manual clustering:** The second method involved counting of deleterious and damaging nsSNPs across all 6 CHD8 signature regions and 9 MoRfs identified, compare them with variations located outside these signature motifs and identify 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).
11. **Analysis of physiochemical changes due to aa substitutions:Project Have Our Protein Explained (Project HOPE)** : is an automatic mutant analysis server that provides an insight on the physiochemical structural features of native and the variant aa. When input with protein sequence and mutant variants, Project HOPE server predicted structural variation between mutant and wild type residues (Venselaar et al., 2010).
12. **Protein-Protein Interaction (PPI) network construction:Ingenuity Pathway Analysis (IPA) software** [IPA®, QIAGEN Redwood City]: The interacting partners of 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, 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 only were consulted for network building. Prominently, only specific developmental, neurological, psychological, hereditary, metabolic, connective tissue, skeletal and muscular disorders disrupted in ASD subjects were chosen for PPI network construction similar to Ashitha & Ramachandra, 2020. Additionally, molecular functions common to CHD8 interacting partners were identified through IPA and the Gene-Set Enrichment Analysis (GSEA) tool- EnrichR (Kuleshov et al., 2016).
13. **Protein 3D modelling:**

SWISS-MODEL : was utilised 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 minimisation. SWISS-MODEL returned multiple predicted models whose quality was estimated using GMQE score, i.e., ranged 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).

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 were predicted to increase molecular rigidity due to its decreased flexibility.

3. RESULTS:

1. **N and C terminal exons hosted the highest truncating SNPs and nsSNPs, respectively:** A total of 84,026 *CHD8* SNPs were collected from 4 databases. 90.73% of all SNPs were gathered from Ensemble, followed by GnomAD, and SFARI database for ASD variants with 143 SNPs (0.17%) (Figure 1A, Supplementary Table S1); removing duplicates retained 1037 nsSNPs (94.53%) and 23 Frameshift insertion/deletion SNPs (2.09%) and 35 Stop gain SNPs and 2 Start loss SNPs (3.37%) in the general population. SFARI database reported a higher occurrence of truncating variations (55%) among ASD population (Figure 1B) and within the general population, missense SNPs were the most common variation type. Only 14 nsSNPs and 10 stop-gain SNPs (i.e., 2.19% of all SNPs) were common between both general and ASD population and 79.3% of variations (i.e., 8.38% of all SNPs) identified in ASD population were unique (Table 1, Figure 1C); including 9 SNPs that were found recurrently mutated in [?]2 unrelated ASD subjects and 3 SNPs overlapping both populations (Supplementary Table S2). To measure the relative abundance of different types of variations across exons and domains, they were mapped onto their respective regions. A higher proportion of variations occurred within the C terminal end of protein CHD8, especially within CHD7-binding/ FAM124B-interacting region and BRK domain- corresponding to exons 29 to 37 (>50% variations) and were primarily composed of nsSNPs. Within this region, exon 30 hosted the highest density of variations (73.24%) and exon 34 the least. Overall, truncating SNPs were more common in N terminal signature regions, including Chromo, Helicase ATP-binding and SNF2_N domain involving exons 14,10,8 and 7 that recorded the highest aggregate. Exons 17-20 encoding the Helicase-C-terminal region showed the lowest density of variations, followed by Helicase-ATP-binding and SNF2_N domains, whereas the N and C terminal region displayed a higher occurrence of SNPs (Figure 1D and 2, Supplementary Table S3). ASD population displayed higher density of SNPs within core domains of CHD8.
2. **Most deleterious nsSNPs were localised within CHD8 core domains; terminal regions contained benign nsSNPs:**Just 135 out of 1037 nsSNPs (13%) were predicted to be damaging by > 90% of tools; 3 nsSNPs overlapped between general and ASD populations. 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 (Figure 3). Notably, 47-68% nsSNPs within N terminal region, CHD7-binding site, BRK domain including C-terminal regions recorded the highest count of nsSNPs, but were benign and tolerated. Among the 52 nsSNPs in the ASD population, 4 nsSNPs were not processed by any of the tools, and 19 nsSNPs were predicted to be highly deleterious. Supporting the pathogenicity patterns observed in 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) in addition to exons 24 and 29 in SANT and CHD7 binding region were most aggregated with deleterious nsSNPs than regions outside (Table 3, Figure 3, Supplementary Tables S4, S5 and S8).
3. **Helicase-C-terminal (exons 17-20) comprised the most destabilizing nsSNPs:**The 135 dele-

terious nsSNPs were further tested on a different category of tools for their ability to cause protein stability changes, which identified 101 moderately and 42 severely destabilising and highly pathogenic nsSNPs and just one variant common to both populations (Table 2). The most deleterious and destabilising nsSNPs were localised within Helicase C-terminal, the region encoded by exons 17 to 20, followed by Helicase ATP-binding, SNF2_N. This trend was mirrored by the 'ASD-only' nsSNPs of ASD population, where 3 out of 4 nsSNPs displayed a robust destabilising effect on the protein. Combined, this analysis identified 42 severely pathogenic variants passing all thresholds of stringency (Tables 2 and 3, Figure 3, Supplementary Tables S6 to S8). This pattern remained the same when a lower cut-off of DDG < -0.5 was applied and thereby reconfirmed our findings.

4. **Exons 14-20 encoding core CHD8 domains were the most evolutionary conservation domain:** ConSurf provided a comprehensive evaluation of the evolutionary status of CHD8 protein residues and normalised conservation score, which facilitated the identification of a wide range of evolutionarily conserved and variable residues along with the information of their relative positions on the protein structure. The Helicase-C-terminal was identified as the most conserved region of CHD8, followed by SNF2 _N and Helicase ATP-binding domain corresponding to exons 14 to 20. Conversely, residues of exons 1-5 encoding the N terminal region and the last 10 C terminal exons encoding CHD-binding and BRK domain were highly evolutionarily variable by nature (Figure 2B and 3B, Supplementary Table S9).
5. **CHD7-binding site had the highest PTM sites:**A total of 311 PTM sites were identified in protein CHD8 (Q9HCK8), of which 86 and 79 phosphorylation and carboxylation sites were identified, respectively, followed by 28, 20 and 17 acetylations, methylations and ubiquitination sites respectively. Though PTMs were found throughout the protein, a higher aggregate was observed in regions outside domain consisting of evolutionarily variable residues. CHD7-binding site 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) (Figure 2B and 3B, Supplementary Table S10).
6. **CHD8 is a highly disordered protein laden with 9 high-confidence MoRFs:**

Our analysis identified that CHD8 is an Intrinsically Disordered Protein (IDP). For reliable identification, we set a probability/propensity score cut-offs 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 (around aa 1-600) and C terminal regions (2500-2570 aa) of CHD8 separated by exceptionally ordered, evolutionarily conserved domain region (Figure 4). Although IDRs predicted by different tools were in broad agreement, we observed wide contradictions while detecting specific MoRF sites. However, 9 high-confidence MoRF sites and 7 disordered binding sites were predicted (with consensus across tools) within these two large terminal IDRs (Table 4, Figure 4).

Compositional bias between disordered and ordered residues was analysed. While no significant differences were observed among nonpolar residues, polar aa Proline and Serine were the most common residues within disordered regions. An overall significant depletion in aromatic and positively charged aa and enrichment of polar uncharged aa were seen within disordered regions (Supplementary Figure S1). Since PTMs and IDRs commonly coincided, 36% residues within IDRs accumulated PTM sites; with 50, 33, 4 and 3 Phosphorylation, Carboxylation, Ubiquitination, Sulfation and Acetylation sites, respectively; but just two Methylation sites that are known as prominent histone modifiers. The tool DISPHOS detected 84 PTM residues within these terminal IDRs, only 34 PTM sites contained nsSNPs and just one nsSNP (S1759G) was predicted to be pathogenic SNP effect analysis. Additionally, these IDRs were found to be prominent sites for DNA and protein binding (Figure 5A).

Cluster analysis reveals several key characteristics of CHD SNPs:

Mutation cluster analysis of the prioritised 42 severely pathogenic nsSNPs identified two statistically significant clusters of aa substitutions above the clustering MPQS threshold of [?] 0.5 (Supplementary Table S11A and Figure S2). Tool Mutant3D auto-selected PDB model 3mwy to evaluate the spatial arrangements

of these variants. The first significant cluster was identified within the Helicase ATP-binding and SNF2_N domains involving residues 861, 920, 943 (Figure 5B, and Supplementary Figure S2- shown in yellow); whereas the second cluster included residues 1051, 1264, 1325 and 1333- located around SNF2_N and Helicase C-terminal domain (shown in red) indicating that these three domains are central to the precise functioning of the protein CHD8.

Additionally, we looked for statistically significant patterns of association between the occurrence of deleterious and destabilising variations, evolutionarily conserved and variable residues, as well as PTM sites on residues located within or outside domains of the protein. Our analysis revealed that there is a significant difference in the occurrence of truncating SNPs between general and ASD population at *P-value 0.0001* (Supplementary Table S11B); that domain residues hosted severely deleterious aa substitutions than residues outside (*P-value 0.0001*); however, nsSNPs localisation within domains was moderately destabilising or stabilising (Supplementary Table S11C). Evolutionarily 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*).

Importantly, 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 destabilising in nature. ASD population did not contain any SNPs within MoRF sites.

1. **CHD8 PPI network recapitulates common phenotypes associated with *CHD8* mutations:**Remarkably, 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 majority of these protein interactors had a neurodevelopmental role. The 13 prominent networking partners of CHD8, namely: AGR2, CREB1, CTNNB1, CASR, CHD7, ESR2, EZH2, NR2C2, KMT2A, SMARCA1, SOX2, TNIK and TP53 were involved in transcription of DNA/RNA (12 genes), formation of the brain (4 genes), gastrointestinal tract (6 genes), body axis and long-term memory (2 genes each) and elicited important ASD associated phenotypes such as macrocephaly, anxiety (5 genes) and impaired social behaviour (2 genes) (Figure 6 and Supplementary Figure S3A). The most critical CHD8 interactors identified- CTNNB1 and CREB1- were found to produce 5 and 4 ASD associated phenotypes, respectively. Remarkably, 11 out of these 13 CHD8 protein interactors (84.6%) were disordered proteins. Proteins like 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 localisation sequence binding. Eukaryotic transcription initiation, androgen receptor, miRNA regulation, Wnt and TGFB signaling pathways were the other prominent pathways (Supplementary Figure S3B). This PPI network included 24 Zinc Finger domain-containing molecules, followed by CHD core domains containing molecules.
2. **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- passed the necessary quality threshold. Although both these models computed the same GMQE score of 0.1, the structure with a higher QMEAN Z-score (-1.97) was finalised as the best estimated CHD8 model for residues between aa cords 800-1340. (Supplementary Figure 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 and thereby, limiting our downstream analysis to these modelled residues of the core CH8 domains-Helicase ATP-binding, SNF2_N and Helicase C-terminal regions.
3. **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 800-1340 aa, were analysed on DynaMut to assess the impact of these mutations within the Helicase ATP-binding, SNF2_N and Helicase C-terminal domains on protein dynamics and stability. 56 nsSNPs were predicted to be destabilising (DDG <-0.1), of

which only 27 nsSNPs crossed the DDG threshold <-0.5; 54 nsSNPs were found to increase molecular flexibility, but only 11 nsSNPs were above the DDS vibrational entropy cut-off >0.5. Similarly, 33 nsSNPs increased molecular rigidity (DDS < -0.1) and only 9 nsSNPs were above the cut off DDS < -0.5 (Supplementary Table S12). Overall, 28.38% of nsSNPs within SNF2_N domain were destabilising in nature, which was the highest. Helicase C-terminal region had more of flexibility increasing variations, whereas the Helicase ATP-binding domain had rigidity increasing SNPs (Table 5). DynaMut analysed 15 out of the 42 severely pathogenic 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 molecular conformation (Supplementary Table S12 and Figure S5).

DISCUSSION: We evaluated the gene *CHD8'* s intrinsic mutability in the ASD population in the background of its mutational propensity within the general population to report the first detailed *in silico*mutational burden analysis to date. Cumulatively, nsSNPs were the most common type of variations identified frequently within exons encoding C terminal region of *CHD8* ; whereas, truncating SNPs usually occurred in the N terminal side (highest in exon 14, 10 and 8). We observed that exons 14-20 encoded the most conserved regions of CHD8 and thereby, displayed the lowest SNP density, but the highest sensitivity to variations- especially Helicase-C-terminal with the least SNPs among all CHD8 domains. Overall, nsSNPs identified within the core CHD8 domains- Helicase-ATP-binding, SNF2_N and Helicase-C-terminal regions were extremely pathogenic, reflecting their crucial functional roles as evolutionary essential regions of CHD8 (Figure 3). An auxiliary peak was observed within the CHD7-binding region, especially due to pathogenic variations within exon 30. ASD population recorded a significantly higher frequency of truncating SNPs ($P < 0.0001$) (Wilkinson et al., 2015) (An et al., 2020). Although ASD variants were not localised to any specific regions of CHD8, >30% of ASD SNPs, remarkably 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 the highest accumulation of truncating SNPs and severely damaging nsSNPs than the general population (An et al., 2020). Gene *CHD8* recorded 12 different independently occurring recurrent ASD SNPs, 8 of these SNPs (66.7%) were located within signature regions, 3 and 2 out of 7 such truncating SNPs occurred within CHD7-binding motif and SANT and SLIDE DNA binding domain which 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 that encode the 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 9 MoRF sites of < 12 aa 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 & Uversky, 2018). Phosphorylation mediates specific but weak interactions with partners, modulates the binding affinity of transcription factors to their coactivators and DNA and thereby alter gene expression affecting cell growth and differentiation (Darling & Uversky, 2018). These disordered regions of CHD8 were observed to have larger incidences of ASD associated truncating SNPs. An et al., utilised the Chd1 crystal structure (PDB code 5O9G) in their study and remapped *CHD8* mutations onto it. However, 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 800-1340 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 Figure S4)- supported by our finding that CHD8 is a highly disordered protein. 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.

CHD8 mutations have been consistently associated with phenotypes like ASD, macrocephaly, ID and GI complications that were recapitulated in animal models by silencing *CHD8* 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. CHD8 is known to regulate gene expression through protein interactions.

A study utilised both transcriptome and ChIP sequencing in human neural progenitor cells (hNPCs) and identified 1,756 Differentially Expressed Genes (DEGs) and demonstrated widespread binding to chromatin (Sugathan et al., 2014). Another study exploring transcriptional changes due to CHD8 knockdown in hNSCs identified 1,715 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 like social behaviour, anxiety and long-term memory. We suspect that aberrant CHD8 dosage leads to altered regulation of gene expression, cause cumulative changes to these molecular interactions and produce ASD and comorbidities associated with CHD8 mutation and needs further investigation.

Thereby *CHD8* is indeed a master regulator of neuronal and GI functions and hence a potent contributor of ASD. Our in-dept 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.

ACKNOWLEDGEMENT

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

CONFLICT OF INTEREST STATEMENT

The authors declare no competing financial interests.

REFERENCES

An, Y., Zhang, L., Liu, W., Jiang, Y., Chen, X., Lan, X., Li, G., Hang, Q., Wang, J., Gusella, J. F., Du, Y., & Shen, Y. (2020). De novo variants in the Helicase-C domain of CHD8 are associated with severe phenotypes including autism, language disability and overgrowth. Human Genetics. https://doi.org/10.1007/s00439-020-02115-9 Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7:248–249. Ashitha, S. N. M., & Ramachandra, N. B. (2020). Integrated Functional Analysis Implicates Syndromic and Rare Copy Number Variation Genes as Prominent Molecular Players in Pathogenesis of Autism Spectrum Disorders. Neuroscience. 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. In DMM Disease Models and Mechanisms. https://doi.org/10.1242/dmm.005439 Barik, A., Katuwawala, A., Hanson, J., Paliwal, K., Zhou, Y., & Kurgan, L. (2020). DEPICTER: Intrinsic Disorder and Disorder Function Prediction Server. Journal of Molecular Biology. https://doi.org/10.1016/j.jmb.2019.12.030 Barnard, R. A., Pomaville, M. B., & O'Roak, B. J. (2015). Mutations and modeling of the chromatin remodeler CHD8 define an emerging autism etiology. In Frontiers in Neuroscience. https://doi.org/10.3389/fnins.2015.00477 Bernier, R., Golzio, C., Xiong, B., Stessman, H. A., Coe, B. P., Penn, O., Witherspoon, K., Gerdts, J., Baker, C., Vulto-Van Silfhout, A. T., Schuurs-Hoeijmakers, J. H., Fichera, M., Bosco, P., Buono, S., Alberti, A., Failla, P., Peeters, H., Steyaert, J., Vissers, L. E. L. M., ... Eichler, E. E. (2014). Disruptive CHD8 mutations define a subtype of autism early in development. Cell. https://doi.org/10.1016/j.cell.2014.06.017 Calabrese, R., et al., 2009. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum. Mutat. 30, 1237–1244. Capriotti, E., Calabrese, R., Casadio, R. (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics, 22:2729-2734. Capriotti, E., Fariselli, P., Casadio, R., 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 33, 306–310. http://dx.doi.org/10.1093/nar/gki375. Chen, C., Lin, J. & Chu, Y. iStable: off-the-shelf predictor integration for predicting protein stability changes. BMC Bioinformatics 14, S5 (2013).

https://doi.org/10.1186/1471-2105-14-S2-S5 Cheng J, Randall A, Baldi P. 2006. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 62:1125–1132. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., Chan, A.P., 2012. Predicting the functional effect of amino acid substitutions and indels. PLoS One 7 (10), e46688. http://dx.doi. org/10.1371/journal.pone.0046688. Darling, A. L., & Uversky, V. N. (2018). Intrinsic disorder and posttranslational modifications: The darker side of the biological dark matter. In Frontiers in Genetics. https://doi.org/10.3389/fgene.2018.00158 Disfani, F. M., Hsu, W. L., Mizianty, M. J., Oldfield, C. J., Xue, B., Keith Dunker, A., Uversky, V. N., & Kurgan, L. (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, Y. J., Rueda, R., Martorell, A. J., Nott, A., Liu, C. Y., Watson, L. A., & Tsai, L. H. (2016). Chd8 mediates cortical neurogenesis via transcriptional regulation of cell cycle and Wnt signaling. Nature Neuroscience. https://doi.org/10.1038/nn.4400 Ellingford, R. A., Meritens, E. R. de, Shaunak, R., Naybour, L., Basson, M. A., & Andreae, L. C. (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, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N. 2003. ConSurf: Identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 19:163–164. Gonzalez-Perez, Abel, and Nuria Lopez-Bigas. "Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel." The American Journal of Human Genetics 88.4 (2011): 440-449. Guo, H., Wang, T., Wu, H., Long, M., Coe, B. P., Li, H., Xun, G., Ou, J., Chen, B., Duan, G., Bai, T., Zhao, N., Shen, Y., Li, Y., Wang, Y., Zhang, Y., Baker, C., Liu, Y., Pang, N., ... Xia, K. (2018). Inherited and multiple de novo mutations in autism/developmental delay risk genes suggest a multifactorial model. Molecular Autism, 9(1), 64. https://doi.org/10.1186/s13229-018-0247-z Hecht, M., Bromberg, Y., Rost, B., 2013. News from the protein mutability landscape. J. Mol. Biol. 425 (21), 3937–3948. http://dx.doi.org/10.1016/j.jmb.2013.07.028. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., ... MacArthur, D. G. (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, B. J., Shendure, J., & Eichler, E. E. (2014). A de novo convergence of autism genetics and molecular neuroscience. In Trends in Neurosciences. https://doi.org/10.1016/j.tins.2013.11.005 Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., & Ma'ayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Research. https://doi.org/10.1093/nar/gkw377 Kuznetsov, Igor B. "Ordered conformational change in the protein backbone: Prediction of conformationally variable positions from sequence and low-resolution structural data." Proteins: Structure, Function, and Bioinformatics 72.1 (2008): 74-87. Malhis, N., Jacobson, M., & Gsponer, J. (2016). MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. Nucleic Acids Research. https://doi.org/10.1093/nar/gkw409 Marfella, C. G. A., & Imbalzano, A. N. (2007). The Chd family of chromatin remodelers. Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis. https://doi.org/10.1016/j.mrfmmm.2006.07.012 Meszaros, B., Erdos, G., & Dosztanyi, Z. (2018). IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Research. https://doi.org/10.1093/nar/gky384 Meyer, M. J., Lapcevic, R., Romero, A. E., Yoon, M., Das, J., Beltran, J. F., Mort, M., Stenson, P. D., Cooper, D. N., Paccanaro, A., & Yu, H. (2016). mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome. Human Mutation. https://doi.org/10.1002/humu.22963 Michaelson, J. J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., Jian, M., Liu, G., Greer, D., Bhandari, A., Wu, W., Corominas, R., Peoples, A., Koren, A., Gore, A., Kang, S., Lin, G. N., Estabillo, J., Gadomski, T., ... Sebat, J. (2012). Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell. https://doi.org/10.1016/j.cell.2012.11.019 Micucci, J. A., Sperry, E. D., & Martin, D. M. (2015). Chromodomain helicase DNA-binding proteins in stem cells and human developmental diseases. In Stem Cells

and Development. https://doi.org/10.1089/scd.2014.0544 Pejaver, V., Hsu, W. L., Xin, F., Dunker, A. K., Uversky, V. N., & Radivojac, P. (2014). The structural and functional signatures of proteins that undergo multiple events of post-translational modification. Protein Science. https://doi.org/10.1002/pro.2494 Petersen, B., et al., 2009. A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct. Biol. 9, 51 http://dx.doi.org/10.1186/ 1472-6807-9-51. Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: Application to cancer genomics. Nucleic Acids Research. https://doi.org/10.1093/nar/gkr407. Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J. Y., Peng, M., Collins, R., Grove, J., Klei, L., Stevens, C., Reichert, J., Mulhern, M. S., Artomov, M., Gerges, S., Sheppard, B., Xu, X., Bhaduri, A., Norman, U., ... Buxbaum, J. D. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell. https://doi.org/10.1016/j.cell.2019.12.036 Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., Day, I. N. M., & Gaunt, T. R. (2013). Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. Human Mutation. https://doi.org/10.1002/humu.22225 Sim, N. L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. (2012). SIFT web server: Predicting effects of amino acid substitutions on proteins. Nucleic Acids Research. https://doi.org/10.1093/nar/gks539 Sugathan, A., Biagioli, M., Golzio, C., Erdin, S., Blumenthal, I., Manavalan, P., Ragavendran, A., Brand, H., Lucente, D., Miles, J., Sheridan, S. D., Stortchevoi, A., Kellis, M., Haggarty, S. J., Katsanis, N., Gusella, J. F., & Talkowski, M. E. (2014). CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. Proceedings of the National Academy of Sciences, 111(42), E4468–E4477. https://doi.org/10.1073/pnas.1405266111 Tang, Haiming, and Paul D. Thomas. "PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation." Bioinformatics 32.14 (2016): 2230-2232. Thomas, P.D. and Kejariwal, A. (2004) Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. Proc Natl Acad Sci U S A. 101:15398-15403. Wade, A. A., Lim, K., Catta-Preta, R., & Nord, A. S. (2019). Common CHD8 genomic targets contrast with model-specific transcriptional impacts of CHD8 haploinsufficiency. Frontiers in Molecular Neuroscience. https://doi.org/10.3389/fnmol.2018.00481 Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., De Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. Nucleic Acids Research. https://doi.org/10.1093/nar/gky427 Wilkinson, B., Grepo, N., Thompson, B. L., Kim, J., Wang, K., Evgrafov, O. V., Lu, W., Knowles, J. A., & Campbell, D. B. (2015). The autism-associated gene chromodomain helicase DNA-binding protein 8 (CHD8) regulates noncoding RNAs and autism-related genes. Translational Psychiatry. https://doi.org/10.1038/tp.2015.62 Xu, Q., Liu, Y. Y., Wang, X., Tan, G. H., Li, H. P., Hulbert, S. W., Li, C. Y., Hu, C. C., Xiong, Z. Q., Xu, X., & Jiang, Y. H. (2018). Autism-associated CHD8 deficiency impairs axon development and migration of cortical neurons. Molecular Autism. https://doi.org/10.1186/s13229-018-0244-2

WEBLINKS

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 org/
http://mutationassessor.org/r3/
http://folding.uib.es/i-mutant/i-mutant3.0.html
http://predictor.nchu.edu.tw/iStable/

https://www.ebi.ac.uk/interpro/
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/
http://exac.
https://gnomad.
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/

LEGENDS:

Figure 1: (A) Comparison of SNPs in *CHD8* collected across databases like 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 general population versus ASD population. 53.45% of all variations identified in *CHD8* in general population were nsSNP- the most common. However truncating SNPs were the highest recorded variants within ASD population. **(C)** Comparison of coding SNPs in general vs ASD population. 94.5% of all variants collected in general population were nsSNPs and truncating SNPs formed just 5.4%. Whereas ASD population had 55.17% truncating SNPs. 14 (27%) and 10 (35.7%) of ASD variants were common to both population, whereas all frameshift variations identified in 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 general

population, exon 14 had the highest truncating SNPs. Exon 10 displayed the highest SNP density within the N terminal region, C terminal exons 29 to 37 recorded higher SNPs except exon 34.

Figure 2: (A) The longest protein sequence of CHD8 was identified to be 2,581 aa 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 (640-790aa) represent in yellow, Helicase ATP-binding (807-1009aa in maroon/pink), SNF2_N (825-1101aa in red/pink), Helicase C-terminal (1137 – 1288aa in light blue) and BRK domain (2310-2419aa in sky blue), DNA-binding site SANT and SLIDE (1437-1683aa in green) and a region between 1789 – 2302aa 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 C-terminal exons and including exons 2, 3, 10 and 21. Truncating SNPs often localised within the N terminal exons- specifically exons 8,10 and 14. Lowest SNP density was observed in exons 17-20 corresponding to the most conserved region of CHD8. Residues within N terminal exons 1-4 and C terminal exons 31-37 were evolutionarily the most variable. Exons 3 to 5 contained the highest accumulation of PTMs, followed by exons 31, 29 and 21.

Figure 3: Exon and Domain wise distribution of SNPs across 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 fig. **(A)** and domain in fig. **(B).**

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

Figure 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 Figure S2.

Figure 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 to ASD are presented in turquois, regulatory function in orange and others in grey.

Hosted file

`Table 1.docx` available at https://authorea.com/users/349188/articles/474271-autism-spectrum-disorder-is-burdened-by-severely-pathogenic-variations-within-core-domains-of-chd8-and-its-chd7-binding-motif

Hosted file

`Table 2.docx` available at https://authorea.com/users/349188/articles/474271-autism-spectrum-disorder-is-burdened-by-severely-pathogenic-variations-within-core-domains-of-chd8-and-its-chd7-binding-motif

Hosted file

`Table 3.docx` available at https://authorea.com/users/349188/articles/474271-autism-spectrum-disorder-is-burdened-by-severely-pathogenic-variations-within-core-domains-of-chd8-and-its-chd7-binding-motif

Hosted file

`Table 4.docx` available at https://authorea.com/users/349188/articles/474271-autism-spectrum-disorder-is-burdened-by-severely-pathogenic-variations-within-core-domains-of-chd8-and-

its-chd7-binding-motif

Hosted file

Table 5.docx available at https://authorea.com/users/349188/articles/474271-autism-spectrum-disorder-is-burdened-by-severely-pathogenic-variations-within-core-domains-of-chd8-and-its-chd7-binding-motif