General Psychiatry

Genetic insights into male autism spectrum disorder in a small cohort of Indian simplex families: findings from whole exome sequencing

Srividhya Durbagula,[1,2] Snijesh Valiya Parambath,[3,4]
Ashitha Siddappa Niranjana Murthy,[2,3] Meghana Rameshraju K,[5]
Chetan Ghati Kasturirangan,[6] Gautham Arunachal Udupi,[6] Nallur B Ramachandra,[5]
Aparna Huligerepura Sosalegowda,[1] Vijaya Raman,[2,3] Aruna Korlimarla,[3,7]
Naveen Kumar Chandappa Gowda [ID] [3,4]

For numbered affiliations see end of article.

Correspondence to
Dr Naveen Kumar Chandappa Gowda;
naveen.gowda@sjri.res.in

Dr Aruna Korlimarla;
aruna.careadd@stjohns.in

To the editor:

Autism spectrum disorder (ASD) is believed to have a multifactorial aetiology involving both genetics and environmental factors. Evidence also emphasises that ASD is programmed during the in utero period, with multiple prenatal and postnatal factors influencing the epigenome and contributing to the onset of ASD.[1] Disruption in the neuronal network across various developmental stages leads to neurodevelopmental disorders,[2] which are characterised by dysregulated neuronal communications leading to a vast array of clinical features.[3] ASD predominantly impacts an individual's communication, behaviour and social interaction skills. The global burden of ASD is estimated to be 1 in 132 individuals.[4] However, in India, where comprehensive nationwide studies are lacking, studies conducted in smaller communities reported an ASD incidence of approximately 1 in 450 individuals aged 1.5–10 years.[5]

The role of genetics in the aetiology of ASD has been well documented for over two decades.[6] The underlying complex genetics of ASD involve a diverse array of pathological mechanisms that remain to be fully elucidated. While numerous population studies have undertaken extensive exome analyses, identifying hundreds of causal genes,[7] there is limited research specifically examining Indian subjects. To address this gap, this study examines a small cohort of boys with autism from simplex families to elucidate the genetic underpinnings of ASD within the Indian population. Whole exome sequencing was conducted on 23 trio familial samples (n=69), focusing on protein-damaging inherited and de novo variants using a computational method.

PARTICIPANT RECRUITMENT

A total of 23 trio simplex ASD families were recruited at the Centre for Advanced Research and Excellence in Autism and Developmental Disorders, St John's Research Institute, Bangalore, Karnataka, India. The diagnosis of ASD was conducted using the INCLEN Diagnostic Tool for ASD, developed based on Diagnostic and Statistical Manual of Mental Disorders, Text Revision, Fourth Edition.[8] Furthermore, clinical details, demographic information and other relevant details were collected. The study followed a detailed sampling protocol, including specific inclusion and exclusion criteria as described in online supplemental figure 1.

WHOLE EXOME SEQUENCING ANALYSIS

DNA extracted from peripheral blood samples was subjected to exome sequencing in two batches. The initial batch comprised 13 trio families, and their genetic material was sequenced using SureSelectXT Human All Exon (V6) on the Illumina HiSeq series. In the second batch, which included 10 trio families, the Twist Human Customised Core Exome Kit was used and sequenced on Illumina HiSeqX/NovaSeq. A total of 69 samples from 23 trio families were sequenced, generating 2×150 bp sequence reads at 80×–100× coverage.

Each trio set underwent a relatedness check, and the FASTQ files were subjected to

quality control using FASTQC. Low-quality reads (Phred score ≤20) and adapter contaminants were removed using TrimGalore and Cutadapt. High-quality paired-end reads were aligned to the GRCh38 reference genome with Burrows-Wheeler Aligner, Base Quality Score Recalibration using Genome Analysis Toolkit. Variant calling involved two phases: identifying inherited germline variants (single nucleotide variants and indels) with the Genome Analysis Toolkit best practice pipeline and calling de novo variants with the Genome Analysis Toolkit and VarScan. Finally, variants were annotated using the variant effect predictor by applying parameters such as depth ≥20, genotype quality ≥20, variant allele frequency ≥0.5 and allelic depth at the mutated base >10. Pathogenicity prediction tools such as SIFT, FATHMM, MutPred, PolyPhen-2, CADD, Mutational Taster, Bayesdel, MetaSVM and CLinPhred were used. The variants are classified as 'damaging (D)' if over 60% of the tools concurred on their deleterious nature. Furthermore, a global minor allele frequency threshold of <0.01 was applied to the filter, using data from the 1000 Genomes, gnomAD, ExAC and the Genome Asia databases. Subsequently, genes were prioritised based on two distinct criteria: (i) 'reported' genes in Simons Foundation Autism Research Initiative (SFARI) database, among these SFARI genes with an Evaluation of Autism Gene Link Evidence (EAGLE) score >7 were considered 'high-risk' genes associated with the ASD phenotype and (ii) 'unreported' genes were referred to as novel genes. The pictorial representation of workflow is depicted in online supplemental figure 2.

FUNCTIONAL ANALYSIS

Novel genes were assessed for functional similarity with SFARI gene scores 1 and 2 by semantic similarity analysis. Wang's similarity metric was employed to compare the biological process hierarchy of these genes. The R tool GoSemSim was used to quantify semantic similarity between gene pairs, employing the Best-Match Average method to integrate the semantic relationship scores of multiple Gene Ontology (GO) terms. Further temporal dynamics of identified genes were explored by using gene expression profiles (https://www.brainspan.org/) from distinct brain cortical regions within two distinct groups, namely the prenatal group and the postnatal group. The period between 8 and 38 postconception weeks was considered prenatal, and after birth until 40 years was considered postnatal. Both groups were compared to identify differentially expressed genes (DEGs). Subsequently, the protein-protein interaction (PPI) network was constructed, incorporating all functionally similar 'novel genes' and 'reported genes' using the STRING database (https://string-db.org/). A medium confidence interaction score (0.400) was selected, and networks were visualised by Cytoscape (V.3.10.1). By directing attention to nodes with a degree >10, major protein hubs in the network were scrutinised. Subsequently, DEGs in the PPI network were focused and considered as candidate

novel genes that were functionally associated with ASD pathology as observed by ToppGene enrichment analysis (https://toppgene.cchmc.org/).

CLINICAL DETAILS

The study was conducted on 23 trio samples (n=69), comprising children with ASD and their biological parents of Indian origin. All participating children with ASD were born of non-consanguineous marriages and were reported as full-term births with an average birth weight of 3.0 kg. The Vineland Adaptive Behaviour Scale-2 results indicated that ASD probands showed deficits in the social and communication domains compared with the motor skill domain (online supplemental figure 3). Complete demographic details are shown in online supplemental table 1.

TOTAL GENETIC BURDEN OF INHERITED AND DE NOVO VARIANTS IN CHILDREN WITH ASD

The computational pipeline yielded a comprehensive prediction of 650 inherited missense variants and 339 loss-of-function (LoF) variants, involving 869 genes (figure 1A; online supplemental table 2). Notably, each proband was found to harbour over 30 inherited variants (figure 1B). Further analysis identified 38 de novo missense variants and 35 de novo LoF variants spanning 63 genes (figure 1C; online supplemental table 3). Remarkably, 73% of probands were found to harbour de novo variants, while 27% of probands reported no such variants (figure 1D). This highlights a significantly higher load of inherited variants than de novo variants within the study cohort. While focusing on genes with recurring mutations, it was found that 15 genes displayed recurrent mutations in at least 3 individuals, whose functional importance was determined later (figure 1E). Furthermore, the predictions suggested a higher likelihood of heterozygous variants (95%), rather than homozygous (5%) (figure 1F).

In total, 932 genes were identified with damaging variants, emphasising the heterogeneity within the cohort. Among them, 72 genes were previously known and reported in the SFARI database (figure 1G). Remarkably, 96% of cases harboured variants in either high confidence or strong candidate genes for ASD (SFARI categories 1 and 2) (figure 1H). Intriguingly, the study revealed rare variants in genes that were strongly associated with ASD phenotypes, as evidenced by an EAGLE score >7. We consider these genes as potential 'high-risk' candidate genes that account for 30% of children with ASD in our cohort (table 1).

PRIORITISATION OF 'NOVEL' GENES BY FUNCTIONAL ANALYSIS

Gene expression profiles in brain cortical regions unveiled distinct temporal patterns, identifying two gene groups: one upregulated prenatally and the other
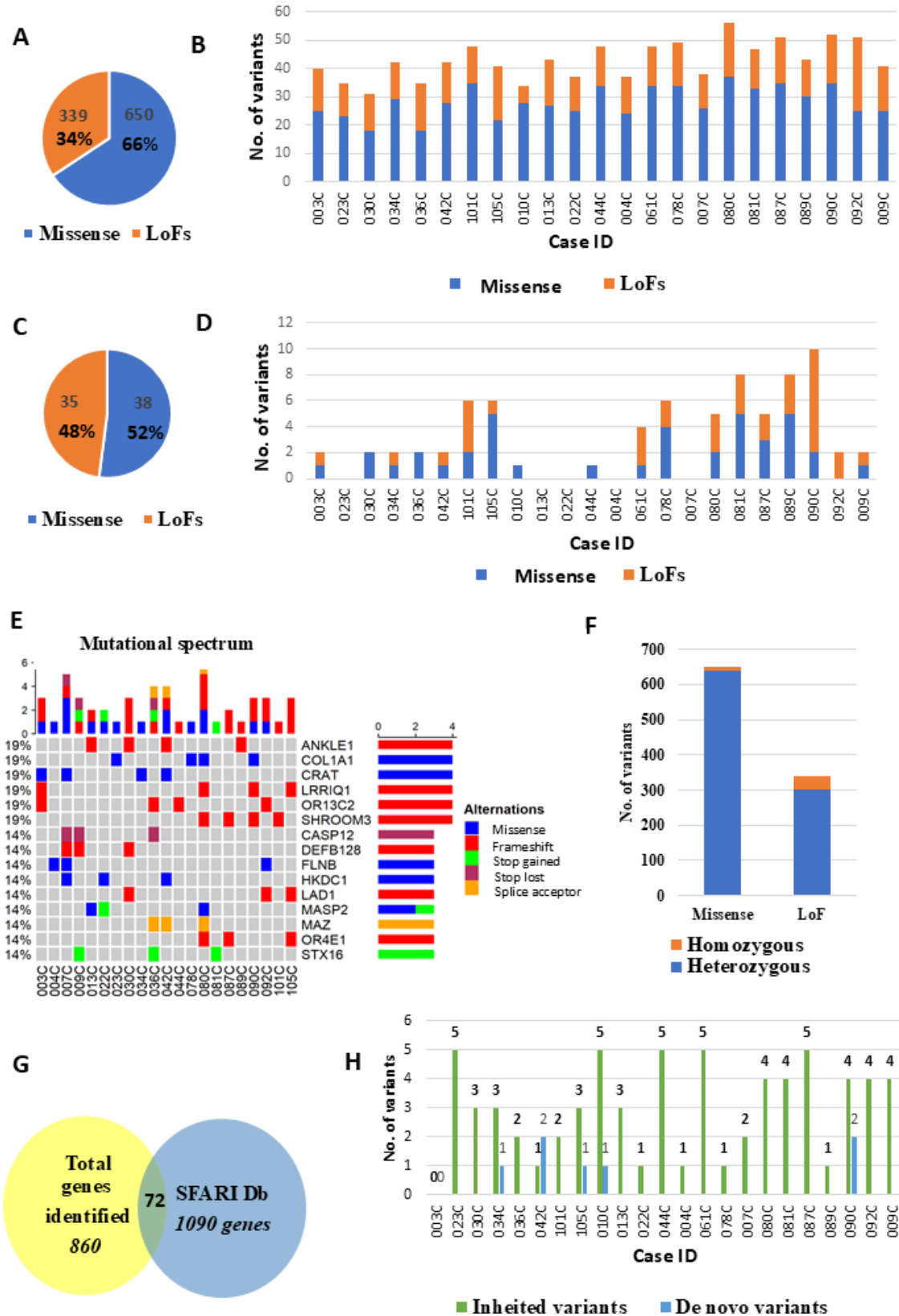
Figure 1 Total genetic burden in the ASD probands. (A) Total number of missense and LoF inherited variants identified; (B) total number of inherited variants identified across the subjects with ASD; (C) total number of missense and LoF de novo variants; (D) total number of de novo variants identified across the subjects with ASD; (E) recurrently mutated genes (genes mutated in more than three subjects) within the cohort; (F) gene counts with heterozygous or homozygous variants; (G) number of genes carrying variants overlapping with the SFARI Db; (H) number of high confidence and strong candidate genes carrying damaging variant as reported by the SFARI Db identified across subjects with ASD. ASD, autism spectrum disorder; ID, identification number; LoF, loss of function; SFARI Db, Simons Foundation Autism Research Initiative database.

Table 1 'High-risk' genes associated with ASD phenotype carrying damaging variants

Case ID	Symbol	Inherited/ De novo	MAF (Genome Asia)	Consequences	Variant type	Phenotype associated (animal models)	VABS score
030C	*CACNA1D*	Inherited	0.000289	c.1697A>G	Missense variant	Restricted, repetitive and perseverative behaviour	Moderately low
105C	*RELN*	Inherited	–	c.5305G>A	Missense variant	Decreased social interaction	Low
081C	*NRXN2*	Inherited	–	c.3461G>A	Missense variant	Vocalisation pattern/ social behaviour pattern	Low
092C	*SHANK2*	Inherited	–	c.400G>T	Missense variant	Vocalisation pattern/ social behaviour pattern	Moderately low
092C	*ZNF462*	Inherited	–	c.5043C>G	Missense variant	Syndromic ASD form with ADHD	Moderately low
090C	*WDFY3*	De novo	–	c.7753–7754 CCA>C	LoF	Macrocephaly, deficit in motor coordination, associative learning	Low
034C	*BRSK2*	De novo	–	c.978G>C	LoF	Social impairment	Moderately low
010C	*DEAF1*	De novo	–	c.670C>T	Missense variant	Memory deficits and increased anxiety-like behaviour	Low

ADHD, Attention deficit hyperactivity disorder; ASD, autism spectrum disorder; ID, identification number; LOF, loss-of-function variants; MAF, minor allele frequency; VABS, Vineland Adaptive Behaviour Scale.

postnatally (online supplemental figure 4A). Among the novel genes that functionally align with SFARI genes, 63 exhibited significant differential expression patterns between prenatal and postnatal periods, with a logFC >|1| and an adjusted p value <0.05 (online supplemental figure 4B). Of these genes, 63% were upregulated prenatally and 27% postnatally. Enrichment analysis indicated that prenatally upregulated genes were enriched for distinct biological processes such as transcriptional and translational mechanisms, while postnatally upregulated genes were enriched for cell junction, cell assembly and cell matrix adhesion. Common biological processes between the groups included extracellular matrix (ECM) organisation, GTPase regulator activity and calcium channel activity (online supplemental figure 4C).

PPI NETWORK

The PPI network comprised 302 nodes and 644 edges, with an average node degree of 2.96 (online supplemental figure 5A). The significant PPI enrichment p value of 2.89×10^{-9} emphasised a notable association among the proteins within this network. Enrichment analysis indicated that these proteins were involved in various biological processes (online supplemental figure 5B). Notably, 30 proteins in the network exhibited a degree >10, with 76% of these proteins falling within the same subnetwork. This subnetwork was constructed with FN1 protein that

had a higher degree of 37 (online supplemental figure 5C). GO predicted that these proteins in the subnetwork were associated with cell-cell communication (online supplemental figure 2). Additionally, SMARCA4 with 16 edges and SCN5A with 10 edge counts formed subnetworks that were enriched for positive regulation of transcription by RNA polymerase II (GO:0045944) and high voltage-gated calcium channel activity (GO:0008331), respectively (online supplemental figure 5E,F).

Furthermore, focusing on DEGs in PPI subnetworks, we found 14 novel genes as candidate genes in this study with functional significance by the adapted integrated analysis (online supplemental table 4). Among these, 10 DEGs were present in the major network: 6 postnatally upregulated genes (*TLN1, THBS2, ITGA7, CD63, ALCAM, TNS1*) that were enriched for Kyoto Encyclopedia of Genes and Genomes ECM receptor interaction pathway (M7098) with a false discovery rate-adjusted p value <0.001, and 4 prenatally upregulated genes (*HDAC2, SOS1, HMGA2* and *FBN3*) that were enriched for ECM organisation and epithelial development, interconnected with known SFARI genes (online supplemental figure 5C). Additionally, we observed five DEGs in a small subnetwork associated with positive regulation of transcription, all of which were prenatally upregulated (*FOXM1, HDAC2, MED17, VTA1, SS18L1*). The study also highlighted DEGs such as *FOXM1, TNS1, FBN3* in these

subnetworks, which were recurrently mutated in our cohort.

Robust prevalence studies, which are currently lacking, are crucial to understanding the true burden of ASD in the Indian population. Moreover, there are a limited number of genetic studies on Indian families affected by ASD. To identify potential disease-causing genes in the Indian population and to understand the intricate molecular mechanisms underlying the pathology of ASD, we conducted whole exome sequencing on 23 Indian trio familial samples. Given the higher prevalence of ASD in males, as well as factors such as the female protective effect and sex hormones that make females less susceptible to ASD symptoms,[9] this study focused exclusively on males.

Our investigation identified inherited and de novo variants, revealing a higher burden of inherited variants despite 72% of families reporting no history of psychiatric or neurological disorders. These findings are noteworthy given the prevailing belief in the higher heritability rate in ASD, as supported by the literature.[10] In contrast, a recent study on the Indian population observed a higher prevalence of de novo mutations. However, a limitation of this study is that exome sequencing was conducted only on probands. Additionally, the recurrent mutations in the *MECP2* gene observed may be attributed to sample bias.[11]

The study identified 96% of subjects with ASD in our cohort as harbouring deleterious variants, either in high confidence or strong candidate genes for ASD, suggesting their potential relevance to disease manifestation. Interestingly, the study highlights eight 'high-risk' genes carrying rare protein-damaging variants, potentially contributing to the ASD phenotype in 30% of the subjects with low to moderately low adaptive function. Among the eight high-risk genes, *CACNA1D, RELN, NRXN2, SHANK2, ZNF462* harboured inherited rare missense variants and were identified as candidate genes in four subjects with ASD (030C, 105C, 081C, 092C) (table 1). These high-risk genes are associated with a diverse array of symptoms, including intellectual disability, developmental disorder, neurological disorder, decreased social interaction and restrictive repetitive behaviour.[7 12]

Additionally, in three probands (090C, 034C and 010C) with no reported family history of ASD, we infer that the ASD phenotype is driven by high-risk genes (*WDFY3, BRSK2, DEAF1*) carrying de novo variants. This is consistent with the literature suggesting the implication of de novo variants in sporadic ASD.[13] Literature underscores that autism-associated genes are more intolerant to LoF variants; even carriers of these variants exhibit defective cognitive function, highlighting their significant impact.[14] Hence, we consider *BRSK2* and *WDFY3* with high-impact variants as candidate genes in subjects 010C and 034C with ASD, respectively. Another de novo missense variant in *DEAF1* could be a candidate gene in subject 010C. This inference is supported by studies reporting multiple missense variations in the *DEAF1* gene, predominantly

residing in the Sp100, AIRE-1, NucP41/75, DEAF-1 (SAND) domain, similar to our current findings.[7]

Considering genetic heterogeneity, the study investigated novel genes that could contribute to the ASD phenotype, potentially arising from ethnic differences. The integrated functional analysis identified 14 novel DEGs that are functionally important during the critical stage of brain development. The prenatally upregulated genes, including *FOXM1, HDAC2, MED17, VTA1, SS18L1* and *HMGA2,* are recognised as transcriptional modulators. Given that the prenatal period is characterised by intense synaptogenesis, it can be inferred that the mutation burden in this phase affecting the dynamic regulation of transcription and translation processes may impact epigenetic regulation, contributing to dysregulated neuronal development.

Furthermore, our functional analysis prioritised *TLN1, THBS2, ITGA7, CD36, ALCAM, TNS1, SOS1* and *FBN3,* which were enriched for various biological processes such as ECM and cell junction organisation. These processes converge into similar molecular mechanisms, contributing to fundamental aspects of cellular communication within the brain. Remarkably, the 'high-risk' genes identified in this study, such as *SHANK2, NRXN2* and *RELN*, were also enriched for cell-cell communication pathways. This finding aligns with a broader body of research that supports the pathology of ASD mediated by cell adhesion molecules (the Neurexin gene family) and neural ECM molecules.[15] Interestingly, one of the large exome studies revealed that out of 102 identified genes, 29 were associated with neuronal communications, including *NRXN2* and *SHANK2*, leading to defective synaptogenesis and synaptic plasticity.[7] Collectively, we posit that there is a dysregulation of intercellular communication that might have caused defective neurogenesis and synaptogenesis in ASD.

It is crucial to acknowledge the limitations of our study, particularly the limited sample size. Nevertheless, the study successfully identified candidate genes associated with 30% of subjects with ASD in our cohort. The identified novel genes carrying protein-damaging variants may add to the existing literature. Further research and evidence are warranted to substantiate these findings.

Author affiliations
¹Department of Studies in Biotechnology, University of Mysore, Mysore, Karnataka, India
²St John's National Academy of Health Sciences, Bangalore, Karnataka, India
³CAREADD, St John's Research Institute, Bangalore, Karnataka, India
⁴Department of Molecular Medicine, St John's National Academy of Health Sciences, Bangalore, Karnataka, India
⁵Department of Studies in Genetics and Genomics, University of Mysore, Mysore, Karnataka, India
⁶Department of Human Genetics, National Institute of Mental Health and Neurosciences (NIMHANS), Bangalore, Karnataka, India
⁷Department of Research, Sri Shankara Cancer Hospital and Research Center, Bangalore, Karnataka, India

ORCID iD
Naveen Kumar Chandappa Gowda http://orcid.org/0000-0003-2944-6965

REFERENCES

1 Durbagula S, Korlimarla A, Ravikumar G, *et al*. Prenatal epigenetic factors are predisposing for neurodevelopmental disorders-considering placenta as a model. *Birth Defects Res* 2022;114:1324–42.
2 Penn AA, Shatz CJ. Brain waves and brain wiring: the role of endogenous and sensory-driven neural activity in development. *Pediatr Res* 1999;45:447–58.
3 Heck DH, Lu L. The social life of neurons: synaptic communication deficits as a common denominator of autism, schizophrenia, and other cognitive disorders. *Biol Psychiatry* 2012;72:173–4.
4 Lord C, Brugha TS, Charman T, *et al*. Autism spectrum disorder. *Nat Rev Dis Primers* 2020;6:5.
5 Arun P, Chavan BS. Survey of autism spectrum disorder in Chandigarh, India. *Indian J Med Res* 2021;154:476–82.
6 Folstein S, Rutter M. Genetic influences and infantile autism. *Nat New Biol* 1977;265:726–8.
7 Satterstrom FK, Kosmicki JA, Wang J, *et al*. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* 2020;180:568–84.
8 American Psychiatric Association. *Diagnostic and statistical manual of mental disorders, fourth edition, text revision (Dsm-Iv-Tr)*. Washington,DC: American Psychiatric Association, 2000.
9 Ferri SL, Abel T, Brodkin ES. Sex differences in autism spectrum disorder: a review. *Curr Psychiatry Rep* 2018;20:1–17.
10 Sandin S, Lichtenstein P, Kuja-Halkola R, *et al*. The heritability of autism spectrum disorder. *JAMA* 2017;318:1182–4.
11 Sheth F, Shah J, Jain D, *et al*. Comparative yield of molecular diagnostic algorithms for autism spectrum disorder diagnosis in India: evidence supporting whole exome sequencing as first tier test. *BMC Neurol* 2023;23:292.
12 Weiss K, Wigby K, Fannemel M, *et al*. Haploinsufficiency of ZNF462 is associated with craniofacial anomalies, corpus callosum dysgenesis, ptosis, and developmental delay. *Eur J Hum Genet* 2017;25:946–51.
13 O'Roak BJ, Deriziotis P, Lee C, *et al*. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* 2011;43:585–9.
14 Rolland T, Cliquet F, Anney RJL, *et al*. Phenotypic effects of genetic variants associated with autism. *N Med* 2023;29:1671–80.
15 Guang S, Pang N, Deng X, *et al*. Synaptopathology involved in autism spectrum disorder. *Front Cell Neurosci* 2018;12:415409.

Durbagula Srividhya is a doctoral candidate at the University of Mysore, Department of Studies in Biotechnology, Mysore, India. She obtained her Master's degree and M. Phil degree in Biotechnology from Madurai Kamaraj University, Tamil Nadu, India, in 2011. She began her career in autism genetic research at the All-India Institute of Speech and Hearing, Mysore. In 2015, she joined the Centre for Advanced Research & Excellence in Autism and Developmental Disorders (CARE-ADD) at St. John's Research Institute, Bangalore, as a Senior Research Fellow. She registered for her Ph.D. in 2021. Her main research interests include understanding the genetic underpinnings in children with autism spectrum disorder and exploring the impact of mutations on molecular mechanisms using model systems.