UGC MAJOR RESEARCH PROJECT

MRP ID: MRP-MAJOR_GENE_2013-19809

SUBMISSION OF FINAL REPORT FOR ASSESSMENT BY EXTERNAL EXPERTS

Title of the project

Elucidation and validation of the burden of DNA variations in Autism Spectrum disorders to assess the impact on the genetic pathways

Principal Investigator

Dr. N. B. Ramachandra
Professor and Principal Investigator
Department of Studies in Genetics and Genomics
University of Mysore
Manasagangotri
Mysuru - 570 006

Co-Investigators

Dr. K. C. Shyamala
Professor and HOD
Department of Speech Language Pathology
Chairperson, Autism Spectrum Disorders Unit
All India Institute of Speech and Hearing
Manasagangotri, Mysuru - 570 006

Dr. Prakash Padakannaya
Professor and Coordinator for Innovative program on LD
Department of Studies in Psychology
University of Mysore
Manasagangotri, Mysuru - 570 006

INTRODUCTION

Autism is a genre of serious neurodevelopment disorders, including difficulties in reciprocal social interactions and abnormalities in verbal and nonverbal communications along with strong repetitive behaviours and stereotyped interests (Cook et al., 2001). The most prevalent and exclusive comorbidities of autism are hypersensitivity, impulsivity, agitation, mood swings, mild to severe levels of impairments in cognitive functions ranging from above average to intellectual disability, accompanied by seizures and language impairment (Cook, 1998). Prevalence of these deficits in the functional domains and defects in one or more of these areas before the age of three years results in autism onset (Cook et al., 2001). Copy number variation (CNV) studies on autism have discovered large rare and common variants conferring varying effects on autism risk in the general population (Girirajan et al., 2012; Marshall et al., 2012). Genome and exome Sequencing in Autistic families have identified several *de novo,* novel, deleterious variations in greater than 1000 genes in studies thus far. According to Simons Foundation Autism Research Initiative (SFARI) curative database for Autism, there are 1036 scored autistic gene, 2273 CNV loci, and about 2190 animal models to study autism many of these gene variants affects the biochemical pathways and are known to cause autism.

CNV are known to increase risk for Autism in 5-10% of the cases (Vorstman et al. 2006; Pinto et al. 2010,2014). Its occurrence can be limited to a single gene or a contiguous set of genes and result in dosage sensitive nature which can contribute to human phenotypic variability, complex behavioural traits and disease susceptibility. These can cause functional loss either by disrupting regulatory elements, generating fusion proteins or through position effect variegation. About 600 genes are reported till date which causes autism in varied ways, out of which 354 genes have convincing genetic evidences to be ascertained as causal with replication in literature.

In 2010, the Autism Genome Project Consortium reported that in a sample of nearly 1,000 people with autism, compared with a group of matched controls, there was a significantly higher burden of rare CNVs involving functional genes. The researchers conclude that CNVs are likely to play a role in causing autism (Gross, 2011). Girirajan et al. (2012) proposed a "two-hit" or "second site" model that is based on the observation that affected persons with a micro deletion on chromosome 16p12.1 are more likely to have additional large copy-number variants than are controls. This second set included copy-

number variants known to be associated with a much more variable set of outcomes, ranging from neuropsychiatric disease to severe intellectual disability such as 16p12.1 (*CDR2*) deletion, 3q29 (*DLG1*) duplication and rare copy-number variants that are potentially pathogenic (Girirajan et al., 2012).

Analysis of structural variations such as CNVs in normal cohorts has enabled its utility in development of biomarkers in diagnosis and prognosis of neurodevelopment phenotypes, which until now was limited to leukemia and lymphoma. Increased probability of accumulating CNVs or exposure to triggering elements could enhance the development of autism. Similar such studies on diseases like leukemia and lymphoma have shown varied presences across healthy populations (Yasukawa et al.,2001). These findings highlight the role of such variations present in normal, healthy groups and would require contributions from other factors. It is hypothesized that the CNV may not be the initiating event in the pathogenesis, and additional preceding mutations may be required to induce autism (Girirajan et al.,2012).

Single-nucleotide polymorphisms (SNPs) are genetic markers that enable researchers to search for genes associated with complex diseases. There exist many case-control studies of SNPs associated with ASD. For example, Cheng et al. (2009) showed significant differences in allele frequencies between the ASD and control groups in an association study of four BDNF polymorphisms. Little is known about the degree to which genetic variations underlie symptom variability in ASD; this is especially true for children with the more common form of ASD (typical autism), at least in part because the phenotypic patterns of the three ASD core impairments is extremely variable (Parks et al., 2009). To shed greater light on this issue, researchers have sought to study overall symptom severity in young children with ASD, by comparing known or strongly suspected ASD-related genes or SNPs in such genes, with overall symptom severity. Lerer et al. (2008) found SNPs and haplotypes in the OXTR genes that are associated with IQ and total VABS (Vineland adaptive behavioural scales, (Sparrow & Cicchetti, 1985) scores (as well as the communication, daily living skills and socialization subdomains).

SNPs have been used to help predict whether an individual has a certain disease, or a subtype of a certain disease (Schwender, Ickstadt, & Rahnenfuhrer, 2008). SNP-based diagnostic models have been constructed for several diseases. For example, Huang et al., (2004) used Flextree to differentiate hypertension from hypotension and obtained sensitivity 0.65 and specificity 0.54. To our knowledge, SNP-based diagnostic models have not been

reported in previous ASD symptom-severity studies. Thus, this study focused on finding and studying genetic variations prevailing in Autistic subjects from India, for the first, to elucidate and validate the burden of DNA variations and assess their impacts on genetic pathways. In order to achieve these goals, the following objectives were proposed:

Objectives

A. To analyze for the presence of DNA variations (SNPs, InDels and CNVs) in Autism spectrum disorders in the human whole genome sequences from the 1000 genome project and from 1746 human genome CNP and CNV data

B. To perform DNA variations genotyping on the ASD subjects and controls.

C. To construct and establish various genetic pathways and molecular interaction network affected by these DNA variants in ASDs.

D. Validation and annotation of the newly identified probable candidate genes obtained from the genetic pathways and molecular interaction network studies in ASD subjects

General Criteria used in all studies for selection of Autism subjects: -

Inclusion Criteria:

- Individuals must have a confirmed diagnosis of Autism/ASD/PDD-NOS/PDD/ Asperger's syndrome with or without any comorbidity.
- Cases described as having autistic traits/autistic features were included
- PubMed cited studies on CNV using discovery methods like SNP array, array Comparative Genomic Hybridization (aCGH), Next Generation Sequencing (NGS).
- Analysis performed on blood or brain tissue samples

Exclusion Criteria:

- Syndromic cases of ASD (eg: Rett syndrome, Tourett syndrome etc) and non ASD cases (eg: ADHD, OCD, BP, SCZ etc)
- If the original study did not identify individuals with an ASD diagnosis and if there is an ambiguity whether the individual is autistic or not.
- Studies using less specific CNV discovery method like karyotyping, FISH, qPCR or multiplex ligation-dependent probe amplification (MLPA)
- Controls, unaffected siblings/parents and individuals with unknown diagnosis

RESULTS

Objective A

To analyze for the presence of DNA variations (SNPs, InDels and CNVs) in Autism spectrum disorders in the human whole genome sequences from the 1000 genome project and from 1746 human genome CNP and CNV data

A. i) To analyze for the presence of CNVs in Autism in the human whole genome sequences from 1746 human genome CNP and CNV data. (Manuscript communicated)

Sample Selection: Human whole genome genotyping data performed using Affymetrix Genome-Wide SNP array 6.0 of 12 populations of 1746 (Including 43 Indian) present in the lab has been used to identify hot spots in the regions of the genome bearing ASD genes.

Result: About 34.8% of the subjects were found to have CNVs in the autism sub-genome across the study population. These CNVs were found in individuals across all populations ranging from 8.88 - 49.05% with the highest frequency identified in Australia (49%) and the least in HapMap YRI (8%). A total of 44109 CNVs containing 126190 genes from 1715 individuals. Autism genes were found more in deletion regions (73%) compared to duplication regions (27%), distributed in varying frequencies across populations. Further analysis revealed 73 singleton autism genes out of 354 genes found to be encompassing 838 CNVs in 598 individuals across 12 populations ranging from 0.2 - 55%. Frequency of genes *ARHGAP11B* was found the highest followed by *DUSP22* and *CHRNA7*. Analysis of highly enriched-genes using IPA unravelled the complex interaction network of proteins, and revealed the possible role of novel genes and miRNAs in autism.

Discussion:

Autism-CNVs burden

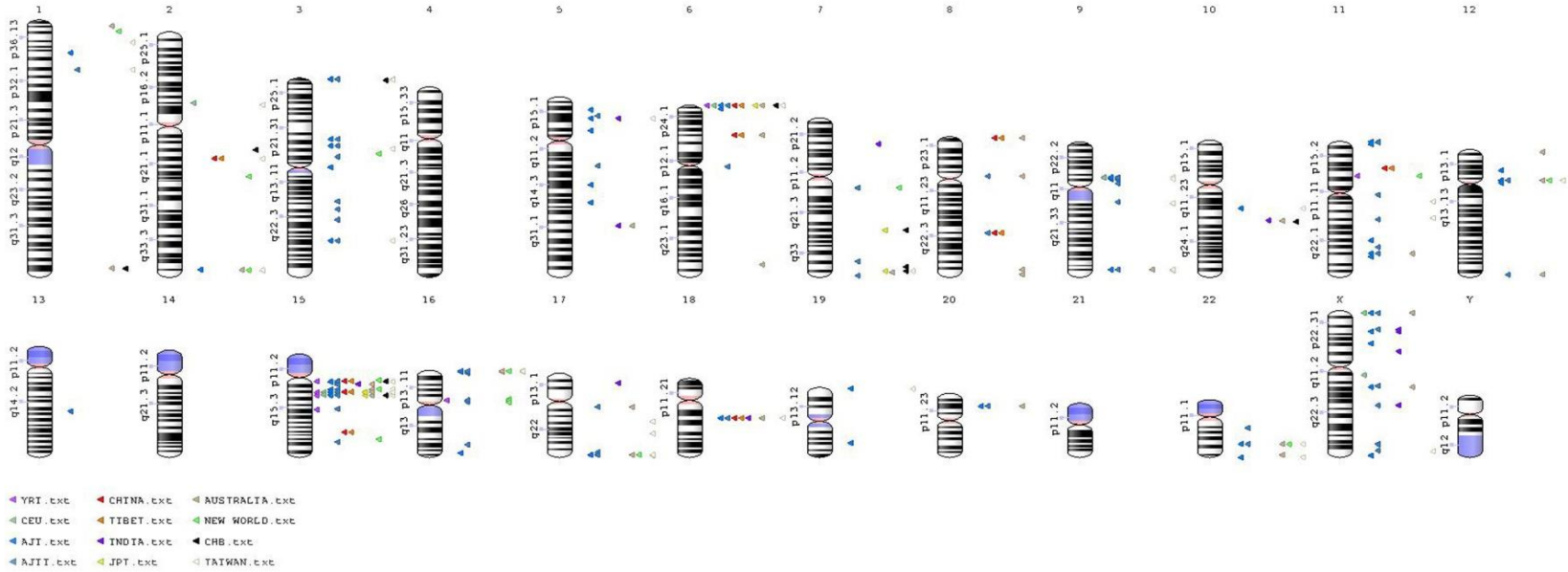
We identified a total of 838 autism gene-CNVs and 3039 miRNA gene-CNVs, out of which 73 autism genes and 243 miRNA genes were singletons across all the 12 populations. Varied distribution of CNVs was observed across both autism and miRNA sub-genome in all the populations. Autism gene-CNV duplications outnumbered the deletion regions evident in AJI and AJII, Australia and Taiwan, while HapMap YRI and HapMap CEU contained equal amounts of duplication and deletions. Duplications and deletions in the sub genome manifests into varied complex diseases owing to the gain and loss of function (Haraksingh et al., 2013).

Duplication or deletion CNVs in exon regions of autism genes impacts the protein structure by either increasing or decreasing the length of the protein (Shishido et al.,2014) while CNVs in the regulatory regions will alter the transcriptional activity of that gene. CNVs in autism genes ranged from 100 kb - 5796 kb and was found distributed in almost all the chromosomes across all 12 populations with a mean size of 261.35 kb. Among all the chromosomes, Chromosome 15 was found containing 3 autism causal genes *CYFIP1*, *NIPA1* and *TUBGCP5* in 192 individuals which were under recurrent CNV events.

Table 1: Number of individuals identified with autism gene CNVs and gender wise distribution of autism CNV duplication and deletions across population.

Sl. No	Population	Total individuals analyzed	No. of individuals with CNVs	CNVs				Genes			
				Male (%)		Female (%)		Male (%)		Female (%)	
				Duplication	Deletion	Duplication	Deletion	Duplication	Deletion	Duplication	Deletion
1	HapMap YRI	90	9	42.86	57.14	100	0	50	50	100	0
2	CEU	90	9	25	75	57.14	42.86	33.33	66.67	50	50
3	Ashkenazi Jews I & II	944	415	20.83	79.17	225	74	68.18	31.82	76.47	23.53
4	China & CHB	199	29	43.75	56.25	50	50	45.45	54.55	77.78	22.22
5	Tibet	31	15	71.43	28.57	73.33	26.67	66.67	33.33	70	30
6	India	38	7	100	0	80	20	100	0	80	20
7	Japan	45	16	33.33	66.67	31.25	68.75	33.33	66.67	33.33	66.67
8	Australia	53	26	82.5	17.5	90.47	9.52	75	25	85.71	14.29
9	New World	41	17	65.21	34.78	-	-	57	43	-	-
10	Taiwan	184	57	92	8	82.76	17.24	86.67	13.33	87.5	12.5
	Total	1715	599	73.19	26.81	73.66	21.95	67.14	32.86	76.38	23.61

a.



b.

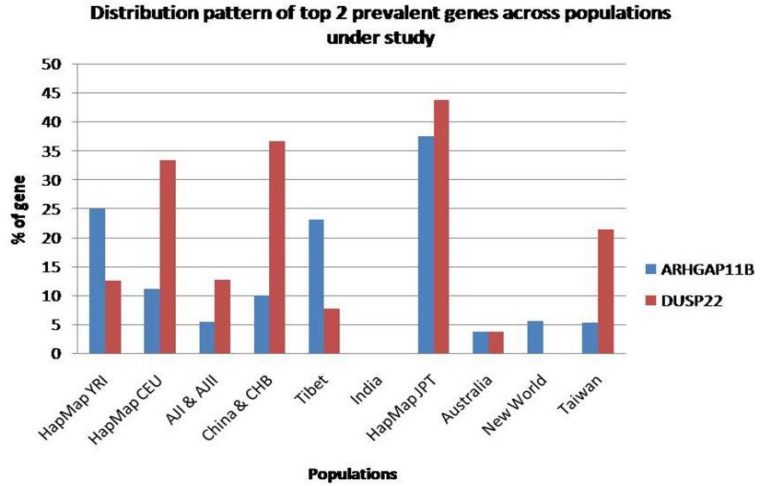


Figure 1 (a) Karyogram of autism genes across populations. CNV burden is prominent in chromosome 15, 16, 6 across all the populations and Chromosome 15 has multiple regions for the CNV prevalence catering to the CNV burden (4-5). There are chromosomes like 2q21.1, 19p13, 20p11, 13q14 which are specific to China and Tibet and Ashkenazi Jews respectively. In case of sex chromosome, the distribution of CNV is varied and HapMap YRI and Tibet is absent altogether. Chromosome 4, 14, 21 and Y have total absence of any CNVs. **(b)** Percentage of top two prevalent autism genes: DUSP22 and ARHGAP11B across populations. Across the 12 populations, HapMap JPT has the highest percentage of both the genes while India has the least.

ii) To analyze for the presence of DNA variations (SNPs, InDels and CNVs) in Autism spectrum disorders in the human whole genome sequences from the 1000 genome project data.

Sample Selection: Whole genome sequences data of all the populations from the 1000 genome project has been analyzed using Genome Browse V1.5, SVS Golden Helix 7.2 Version, RegulomeDB and other softwares. Mapping of high-quality reads of these sequences was performed against the Human reference genome (hg19) to perform various heuristic algorithm based alignments to identify gaps, base substitutions, and synonymous/nonsynonymous changes, Insertions and Deletions (InDels), and Copy Number Variations (CNVs) in ASD-associated genes and also in other regions of the genome interacting with ASD-associated genes.

Result: Variant frequency of Autism specific SNPs in 1000 Genome Project has been found along with RegulomeDB score and frequency of individuals carrying the variations in various ethnicities. Further, followed by expression Quantitative Trait Loci (eQTL) analysis of SNPs, interactive network has been built to show the gene-gene, gene- protein and protein-protein interactions.

Discussion:

Around 25 significant Autism specific SNPs have been identified with eQTL. Score of 1 and 2 indicating significant contribution to the manifestation of autism. The eQTL study has identified 16 genes which are known to cause autism. Pathway built using the 16 identified autism genes gives an insight into the different mechanisms by which manifestation of autism can be carried out. Various autism genes namely, *PITX1*, *BST1*, *PRKCB* and *DLX2* are found to be involved in multiple pathways causing severity in autism cases: 1) *PRKCB* is implicated in circadian rhythms, and learning and memory Serine/threonine-specific protein kinase that mediates B cell activation, T cell migration, antigen-presenting cell function and cytokine release. 2) *BST1* is a glycosylphosphatidylinositol-anchored molecule that promotes pre-B-cell growth, genetic variants have been identified for autism for the same. 3) PITX1 is the key regulator of hormones within the pituitary-hypothalamic axis such as ACTH,cortisol, and beta-endorphin; histone family member H2AFY, which is involved in X-inactivation in females and therefore could be a positional candidate that could explain the 4:1 male:female gender distortion present in autism. 4) DLX genes encode homeodomain transcription factors that

control craniofacial patterning and differentiation and survival of forebrain inhibitory neurons.

Table 2: Variant frequency of Autism specific SNPs in 1000 Genome Project along with regulomeDB score and number of individuals carrying the variations.

SL. No	SNP	SNP Position	RDB Score	Gene	Ref/Alt	All Indiv Freq	European Allele Freq (EUR_AF)	African American Allele Freq (AFR_AF)	American Allele Freq (AMR_AF)	South Asian Allele Freq (SAS_AF)	EastAsian Allele Freq (EAS_AF)
1	rs2470207	chr17:31237993-31237993	2c	-	G/T	0.459065	0.3598	0.5212	0.5245	0.3354	0.5516
2	rs3796863	chr4:15849986-15849986	2c	CD38	G/T	0.435104	0.327	0.6876	0.3444	0.3691	0.3383
3	rs788172	chr2:172953438-172953438	2c	DLX1	G/A	0.402756	0.335	0.7088	0.3127	0.3405	0.1915
4	rs10512479	chr17:35249517-35249517	2b	-	A/C	0.0335463	0.0765	0.0083	0.0375	0.0552	0
5	rs1322784	chr1:231928935-231928935	2b	DISC1/ TSNAX-DISC1	G/A	0.692492	0.7624	0.621	0.7867	0.7495	0.5962
6	rs148637931	chr16:24231344-24231344	2b	PRKCB	C/T	0.0003993	0	0.0015	0	0	0
7	rs1858830	chr7:116312439-116312439	2b	MET	C/G	0.457468	0.4493	0.2209	0.5432	0.5112	0.6647
8	rs1861973	chr7:155254145-155254145	2b	EN2	T/C	0.758986	0.7058	0.7148	0.768	0.6728	0.9474
9	rs198198	chr16:24128397-24128397	2b	PRKCB	T/A	0.638578	0.5815	0.916	0.464	0.4775	0.6081
10	rs226216	chr17:37033996-37033996	2b	LASP1	G/T	0.65595	0.67	0.5651	0.621	0.5562	0.8819
11	rs3732383	chr3:39226127-39226127	2b	XIRP1	C/T	0.130791	0.2207	0.0121	0.1354	0.184	0.1419
12	rs6596188	chr5:134368103-134368103	2b	PITX1	A/T	0.179513	0.1054	0.4561	0.2061	0.047	0.001
13	rs6596189	chr5:134368169-134368169	2b	PITX1	C/T	0.179513	0.1054	0.4561	0.2061	0.047	0.001
14	rs743605	chr2:172967302-172967302	2b	DLX2	C/T	0.316893	0.4533	0.2337	0.4337	0.3436	0.1835
15	rs758158	chr12:2025204-2025204	2b	CACNA SD4	A/G	0.945687	0.8757	0.9962	0.9063	0.9223	0.999
16	rs8076066	chr17:26295337-26295337	2b	-	A/G	0.723243	0.7455	0.8684	0.7752	0.729	0.4692
17	rs11671930	chr19:8117309-8117309	2a	CCL25	T/C	0.0642971	0.16	0.0189	0.1138	0.0573	0.001
18	rs150447075	chr7:145813047-145813047	2a	-	T/G	0.0161741	0.0249	0.0204	0.0187	0.0164	0
19	rs34712024	chr7:145813093-145813093	2a	-	A/G	0.0199681	0.0258	0.0401	0.0274	0.002	0
20	rs10001565	chr4:15722574-15722574	1f	BST1	G/A	0.0001996	0	0	0	0.001	0
21	rs237889	chr3:8802483-8802483	1f	OXTR	T/C	0.70028	0.5994	0.9077	0.8026	0.5665	0.5883
22	rs3861787	chr11:4647770-4647770	1f	-	T/G	0.855232	0.9433	0.5605	0.9135	0.9724	1
23	rs344781	chr19:44174788-44174788	1d	-	C/T	0.736422	0.7505	0.9561	0.719	0.6605	0.5198
24	rs9900089	chr17:30772577-30772577	1d	PSMD11	C/T	0.520966	0.6312	0.7572	0.5504	0.4949	0.1062

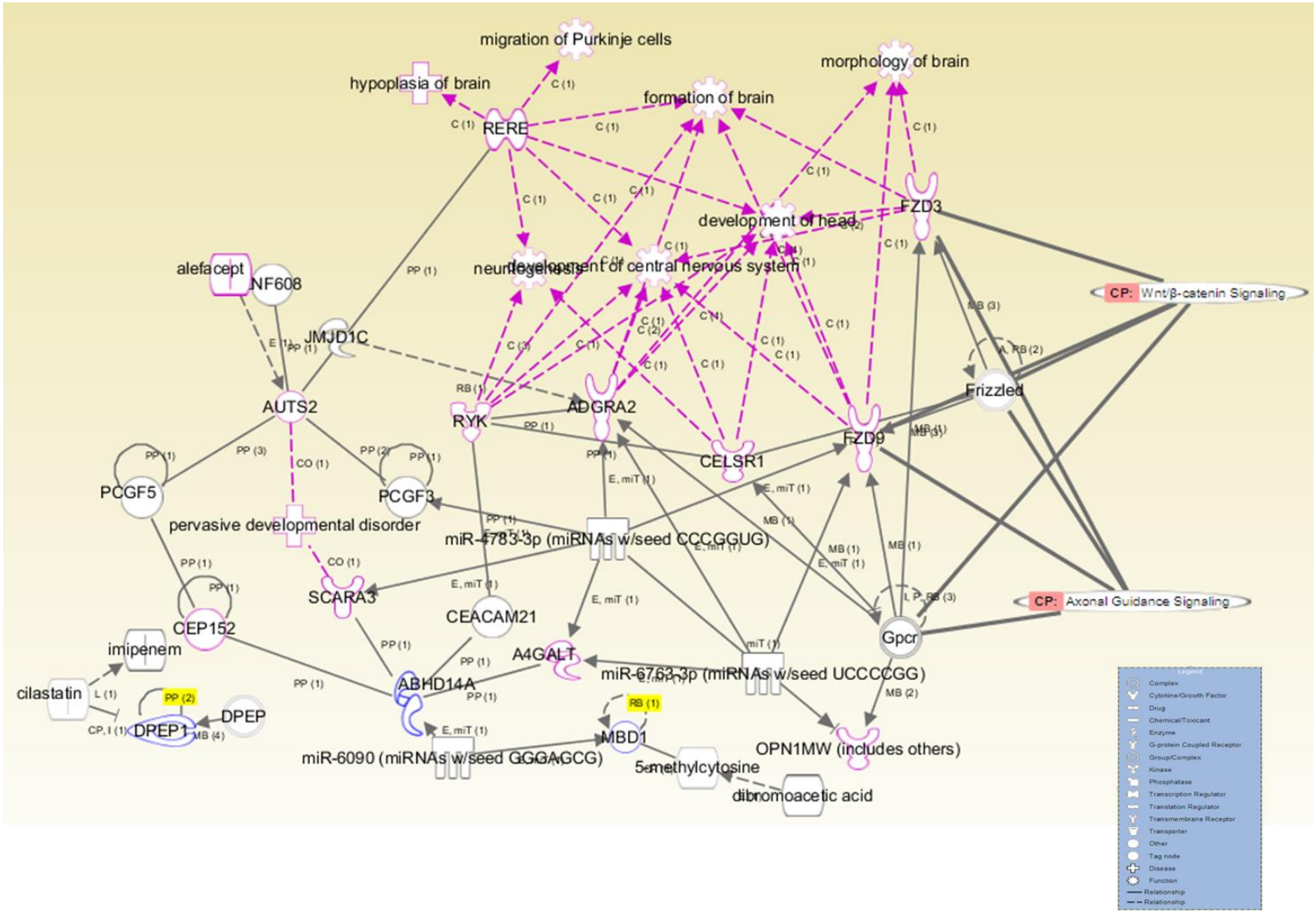


Figure 2: Pathway predicted for genes having Single Nucleotide Polymorphism (SNP) for autism from 1000 genome project

OBJECTIVE B

To perform DNA variations genotyping on the ASD subjects and controls.

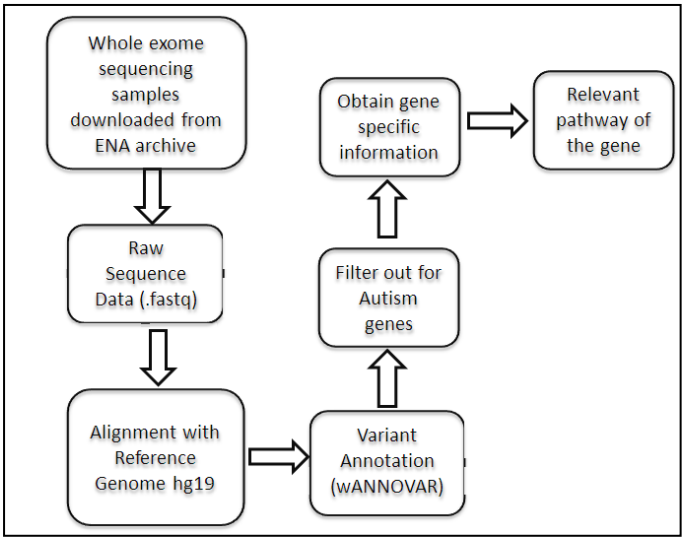
B. i) Deciphering of high-risk genes in global whole exome sequences of autism subjects

Source of Samples: We utilized the sequenced datasets through various research papers and at the European Nucleotide Archive in .fastq format. Sequencing was carried out using Illumina Genome Analyzer IIx paired end sequencing platform, with library preparation according to Illumina protocols. In this study, we opted for 180 global whole exome datasets and 15 Indian samples.

Sequence Reads Alignment: The raw sequence reads were aligned against hg19 build of the human reference genome using Strand and Partek software. Strand NGS is a software platform for next generation sequencing data analysis which can import raw read sequences from sequencing platforms. Post alignment, quality check (QC) was performed. Further, variants were called using the variant calling programme and the files with variants were exported in .vcf format.

Variant Annotation: Variant annotating programme web ANNOVAR (wANNOVAR) and Variant effector Predictor (VeP) were used to annotate the .vcf files that were exported in the previous step. The .vcf files were annotated based on position, gene, and amino-acid change, zygosity and mutation effects. Variant calls which had higher read depth (>20) were included in the study. The .CSV files were exported and priority based classification for the known candidate genes for the two phenotypes were conducted. The 63 candidate genes were filtered for deleterious and damaging mutations (stop gain, stop loss, missense and non-synonymous) and the frequency of these mutations were limited to less than 0.5 based on pathogenicity and haploinsufficiency.

Workflow



Results:

Whole exome sequence (WES) was done at coverage of >100X and obtained total number of reads (R1+ R2) of more than 47 million. The mean read quality (Phred score) was found to be 33.32 for R1 and 33.10 for R2 in both read orientation. The Q score for 30-32 and 33-35 were 27.65% and 66.47% respectively. The mean read length was 100bp and a total of 10.19GB raw data was obtained from sequencing reaction. Annotating the variants from WES revealed a total of 13,097 Single base Variants (SBVs) and 13,102 Multi Base Variants (MBVs).

Mutations for autism manifestation:

Mutations were found across regulatory, untranslated region (UTRs), exons, and introns, downstream and intergenic regions of gene. Annotation of the variants present in the coding region of the DNA revealed several deleterious mutations in genes *KMT2C*, *CNTNAP*, *CACNA1H*, *ANK2*, *CNTN4* and *KATNAL2*. Out of these genes, *KMT2C*, *CNTNAP* and *CACNA1H* were present in 170, 164 and 140 individuals respectively. These mutations were nonsynonymous, stopgain and frameshift deletion of heterozygous nature with haplo insufficiency score of 3 and above and at read depth (RD) of >70. These mutations were predicted to be Damaging/Deleterious by both SIFT/PolyPhen.

KMT2C is a histone methyltransferase that methylates 'Lys-4' of histone H3 - represents a specific tag for epigenetic transcriptional activation present on chromosome 7q36.1. It has shown to have five *de novo* mutations loss of function, c.5216del (p.Pro1739Leufs_2), c.7550C>G (p.Ser2517_), c.1690A>T (p.Lys564_), c.10812_10815del (p.Lys3605fs) an intragenic 203kb *de novo* deletion (Chr7:151858920±152062163) (Koemans et al., 2017).

CNTNAP2 encodes a cell adhesion protein that regulates signaling between neurons at the synapse positioned at 7q35-q36.1. Previously reported deleterious and damaging mutations have been seen in the coding region, promoter, exonic regions (Penagarikano et al., 2016).

CACNA1H encodes a T-type member of the alpha-1 subunit family, a protein in the voltage-dependent calcium channel complex positioned at 16p13.3. There are no relevant previously reported mutations in this gene. However, mutations have been seen in the Calcium signaling family of genes for *CACNA1C* and *CACNA1D* (Li et al., 2015).

Studying the effects of mutations on protein stability and function is important in understanding its role in disease. Site Directed Mutator analyses for *KMT2C*, *CNTNAP* and

CACNA1H was performed and the OSP values were obtained. It was observed that the stability has been reduced considerably in the mutated genes.

Discussion

Whole exome sequence datasets represents autism subjects characterized by repetitive behaviours and compromised socio communication interactions. These cases showed features of intellectual disability, delayed speech language and global developmental delay relevant to autism. WES was employed as a single genetic test along with high throughput computational program to identify causal mutations and disease pathways of autism. Variations across regulatory regions, exons, introns and downstream regions were identified in protein coding as well as non-protein regions. Focusing on Deleterious/Damaging mutations resulted in identification of 1849 mutations that disrupted the normal gene function. Furthermore, utilizing custom developed pipeline with stringent filters revealed 24 significant disease-causing variants. Thus, screening these variants led to the identification of rare mutations in *KMT2C, CNTNAP* and *CACNA1H.*

Table 3: Distribution of autism high risk gene mutations in 180 global whole exome

Genes	Gene Variants Identified in the current Study				Previously Reported Gene Variants	
	Chromosome Position	Reported Cases %	Zygosity	Haplo insufficiency Score	Mutation	Reference
KMT2C	7q36.1	170 (94.44)	Het	53.58	c.5216del (p.Pro1739Leufs_2),c.7550C>G (p.Ser2517_), c.1690A>T (p.Lys564_),c.10812_10815del (p.Lys3605fs)	Koemans et al. 13.10 (2017):
CNTNAP2	7q35-q36.1	164 (91.11)	Het	4.94	3709delG - Coding region, Exon 22; CNV (deletion) - 7q33–q35, Several genes, Exons 1–3rs7794745 - Major allele T,Intron 2 CNV (deletion) – Promoter, H275Ab - Coding region, Exon 6	Peñagarikano et al., (2012):
CACNA1H	16p13.3	140 (77.78)	Hom, Het	71.69	None	
ANK2	4q25-q26	122 (67.78)	Het	24.10	p.D2894Afs*20, Frameshift AD, de novo,	Rossi, Mari, et al. (2017)
SHANK3	22q13.33	118 (65.56)	Het, Hom	60.09	p.A1243GfsX69, Frameshift AD, de novo	
CNTN4	3p26.3-p26.2	104 (57.78)	Het, Hom	6.89	CNV Location and type - 3p26.3, deletion- duplication	Sener, Elif Funda. 2014
KATNAL2	18q21.1	103 (57.22)	Hom, Het	41.98	None	
KDM5B	1q32.1	99 (55.00)	Het	27.99	None	
SYNGAP1	6p21.32	83 (46.11)	Het	19.64	None	
CHD8	14q11.2	79 (43.89)	Het, Hom	11.24	p.C944YfsX3 Frameshift AD, likely de novo, Heterozygous	Rossi, Mari, et al. (2017)
WDFY3	4q21.23	78 (43.33)	Het	21.08	c. 978 T>A nonsense	Iossifov, Ivan, et al (2012):
ANKRD11	16q24.3	74 (41.11)	Het, Hom	75.55	c.7570-1G>C (p.Glu2524_Lys2525del); c.2305delT (p.Ser769GlnfsX8); c.7189C>T (p.Gln2397X); c.5953_5954delCA(p.Gln1985GlufsX46); c.6071_6084delCGTACGCTCTGCCC	Sirmaci, Asli, et al (2011)
IRF2BPL	14q24.3	70 (38.89)	Het, Hom	48.87	None	
SHANK2	11q13.3-q13.4	70 (38.89)	Het, Hom	45.57	15q11.2 deletion, 16p12.1 deletion	Chung et al (2014
ASH1L	1q22	53 (29.44)	Het	23.94	None	
MBOAT7	19q13.42	53 (29.44)	Het, Hom	58.85	None	

ii. **Meta-analysis of Copy Number Variations (CNVs) in individuals with Autism Spectrum Disorder (ASDs) using Global ASD case and control datasets**

Sample Source: The study used Copy Number Variation (CNV) data from Simons Foundation Autism Research Initiative (SFARI) Gene which curates CNV findings from 515 annotated reports on autism and other related neurological disorders comprising of more than 10,000 individuals. CNV data was extracted from SFARI CNV module that provides exhaustive detail on all CNVs reported in individuals with autism for further study.

Database of Genomic Variants (DGV) (MacDonald et al., 2013) curates a catalogue and comprehensive summary of structural variation in human genome of healthy individuals. In general, it serves as a reference control for all CNV studies. For this study we used data from Copy Number Variation Map created by Zarrei, Mehdi, et al. (2015) and submitted to DGV which provides a stringent list of CNV calls in healthy individuals (controls).

Data filtering: The CNV data obtained from SFARI was cleaned from all associated neurological disorders. Inclusion and exclusion criteria were considered as discussed before. CNV data included in the study were further screened for case duplicates. Since different studies use different technology yielding different coverage and resolution, all duplicates were retained with separate identifiers to avoid loss of CNV information. However, CNV data with accurate breakpoint resolution were used for results interpretation. Control data within SFARI CNV module were separated out and used as our internal control dataset.

SFARI Case data processing: The number of individuals under each of the 1382 chromosome loci within SFARI case cohort was identified. Percentage population frequency-defined as the number of individuals under each chromosome locus to the total number of individuals in SFARI case cohort-were calculated. SFARI case chromosome loci were then classified into 4 different groups.

Group ID	Chromosome loci with population frequency
GP 1	≥ 2%
GP 2	1-2%
GP 3	1-0.5%
GP 4	≤ 0.5%

The 3 groups were analyzed separately with the same workflow (as given in CNV Burden Analysis in Cases). Similarly, SFARI control dataset were processed to identify percentage frequency of occurrence of each chromosome loci.

Control Analysis: Chromosome loci of SFARI cases were manually screened for presence of CNVs in two different control datasets: DGV and SFARI controls. Genes falling within CNV regions present in DGV controls were excluded from further analysis and only those genes outside control CNV regions were included in our study.

CNV Burden Analysis in Cases:

Genes outside the burden of CNVs found in control datasets but present within CNVs in SFARI case datasets were selected and scored for its percentage population frequency. The corresponding CNV breakpoints encompassing the genes were identified. The total numbers of studies reporting CNV in these genes and its CNV status (deletions/duplications), average CNV size, gender (number of males/females showing CNV in the region), mode of inheritance (maternal/paternal) were tabulated. Literature was reviewed to identify roles of these genes in autism. When genes were identified as associated with autism, they were further classified based on the Gene Scores (GS) categories provided in SFARI Gene module. SFARI Gene also comprised of Human Gene module that lists 881 genes implicated in autism, with annotations and links to published papers. Most of these genes have been score based on a set criteria formulated by SFARI and categorized into 6 different categories (described below) called gene score (GS) categories and used throughout our study to prioritize gene findings in all chromosome loci.

GENE SCORE (GS) CATEGORY	**DESCRIPTION**
Syndromic (S)	Syndromic
Categories 1 (C1)	High confidence
Categories 2 (C2)	Strong candidate
Categories 3 (C3)	Suggestive evidence
Categories 4 (C4)	Minimal evidence
Categories 5 (C5)	Hypothesized but untested
Categories 6 (C6)	Evidence does not support a role

Scientific reports used in SFARI database are categorized as Major and Minor reports. A research publication is classified as a major report if independent secondary methodology was used to confirm CNV results and as a minor report if no subsequent validation or

confirmation step was performed following the initial discovery. In addition to literature review, this classification was used to further validate the results of our meta-analysis.

Results

SFARI ASD case cohort: The CNV module of SFARI Gene database, as of March 2017, contained 515 annotated reports on autism and related neurological disorders comprising of 2173 chromosome loci. Stringent criteria adopted to include only autism cases into the study resulted in a separate SFARI ASD cohort comprising of 1,382 chromosome loci in 9,635 individuals from 422 annotated reports.

	Cases	**Controls**	
Databases	SFARI	DGV	SFARI controls
# of studies	422	26	52
# of chromosome loci	1,382	11,732 (CNVs)	946
# of individuals	9,635	2,647	~6,700

SFARI ASD CNVs

CNV data processing performed based on chromosome locus yielded in four groups- GP1 to GP4 that were categorized on percentage population frequency. There were 15 chromosome loci with percentage population frequency greater than 2%, the highest being 6.63% and thereby placed in GP1, 68 chromosome loci in GP2 with population frequency between 1-2 % and GP3 contained 135 and GP4 had 404 chromosome loci. The findings of the detailed CNV burden analysis performed on GP1 chromosome loci present in greater than 2% of the SFARI ASD cohort is tabulated in table below.

A total of 15 loci were found to be present in more than 2% of the SFARI ASD case population comprising a total of 9,635 individuals. Locus 15q11.2 was present in 6.63% of the population i.e 639 individuals. Of this 492 had duplications and 147 showed deletions in 15q11.2 locus. Literature review identified 15q11.2 as causal to autism and contained 5 autism associated genes. However, screening against control datasets retained only 75 genes and 1 out of 5 autism associated genes i.e UBE3A, which was present in only 0.84% of the SFARI ASD population. UBE3A gene's intra-locus population frequency was 12.67%.

The second most common loci was 11q11 with percentage population frequency of 4.73% and present in 456 individuals. Most of the CNVs and genes within 11q11

chromosome locus were present in the DGV control dataset and also SFARI CNV control dataset. On manual screening of 11q11 locus we identified only 24 genes were outside the region of control CNVs and therefore considered for further analysis. All genes except one belonged to the olfactory receptor gene family. However, none of them were present in significant number of individuals.

Locus 16p11.2, the third most commonly occurring chromosome locus of SFARI ASD cohort was present in 4.34% of the population i.e. 419 individuals. However most of the CNVs within 16p11.2 locus were present in DGV control dataset and also our SFARI CNV control datasets. There were 240 deletions and 179 duplications with an average CNV size of 316,238bp. The locus hosts 4 autism associated gene. On manual screening of locus 16p11.2 we identified that only 70 genes were absent in controls and therefore were considered for further analysis. The number of individuals carrying each of these 70 genes was calculated. We identified several significant genes like SPN, PPP4C present in 1.55% of the population; DOC2A, YPEL3 and MAPK3 present in 1.54%; SEZ6L2; TBX6 present in 1.53%; C16orf92, ALDOA in 1.52%; PRRT2, MVP, CDIPT, ASPHD1, KCTD13, FAM57B in 1.51%; MAZ, TMEM219, TAOK2, HIRIP3, INO80E, GDPD3 in 1.5%; QPRT and C16orf54 in 1.49% of the SFARI ASD population.

The fourth common locus of SFARI ASD cohort was 8p22. It was present in 301 individuals and 3.12% of the population. Manual screening of 8p22 chromosome locus identified only 15 genes outside control CNVs. FGL1 gene is the top hit within the locus and present in 0.86% of the population frequency. Followed by MTUS1, MSR1, PSD3 and SGCZ genes in 0.47%, 0.42%, 0.311%, and 0.26% of the SFARI ASD population respectively. 131 and 170 individuals had deletions and duplications respectively and the average CNV size was 95,831bp. The locus contains 2 autism associated genes MSR1 and PSD3.

1q44 locus is the fifth common locus identified with population frequency of 3.08% and present in 297 individuals. We identified only 53 genes outside control CNVs. The number of individuals carrying each of these 53 genes was identified and percentage frequency was calculated. The locus contained 190 individuals with deletions and 107 individuals with duplications.

Table 4: Top 15 GP1 chromosome loci and their relevance to autism

Chromo-some Loci	# of individuals	%population Frequency	#of deletions	#of duplications	Avg CNV size(bp)	#of Autism genes	Gene score category	#of other genes (absent in controls)	Significant genes	Evidence from literature on association with Autism
15q11.2	639	6.63	147	492	318174	5	3S,2S	77	UBE3A	Syndromic ASD
11q11	456	4.73	166	290	77826.6	1	None	24	none	No association
16p11.2	419	4.34	240	179	316238	4	4, 2	70	PPP4C,MAPK1	Known
8p22	301	3.12	131	170	95831.7	2	4	15	FGL1	Indirectly linked- display autistic features
1q44	297	3.08	190	107	171549	2	4	49	OR2T29	Indirectly linked -display seizures and congenital anomalies
5p15.33	295	3.06	100	195	210559	4	3	22	ZDHHC11	Syndromic ASD & other Suggestive evidence
3q26.31	286	2.96	160	126	20479.8	2	None	4	NLGN1, NAALADL2	Known
19q13.2	274	2.84	80	194	26283.3	2	None	47	none	Suggestive evidence
9p23	226	2.34	138	88	95302.2	None	None	2	PTPRD	Associated-with strong evidence
13q12.11	206	2.13	90	116	48318.4	None	None	25	SKA3	Suggestive evidence
14q11.2	197	2.04	103	94	242379	1	1S	74	none	Known
4q13.2	193	2	53	140	95579.9	None	None	None	None	No association
22q11.21	192	1.90	122	70	757178	6	3,3S ,4	61	PI4KA, RTN4R	Known
16p13.3	190	1.97	78	112	209483	2	2	None	None	known
5q33.1	190	1.97	128	62	22938.5	2	None	None	None	Syndromic ASD-Tuberous Sclerosis

Significant gene of this locus was OR2T29 belonging to olfactory receptor class and was present in 1.36% of the population. Two other genes, KIF26B and SMYD3 were present in 0.529% and 0.2% of the population whereas the rest of the genes were not present in significant number of individuals.

Locus 5p15.33 was identified as the common locus with population frequency of 3.06% and was present in 295 individuals. Manual screening of 5p15.33 chromosome locus against control datasets identified only 22 genes outside the influence of control CNVs. We identified 100 deletions and 195 duplications incidences. The average CNV size of this locus was 210,559bp. The number of individuals carrying each of these 22 genes was identified and percentage population frequency was calculated. We identified ZDHHC11 gene to be present in 2.15% of the population.

3q26.31.33 is the seventh chromosome locus with population frequency of 2.96% and present in 286 individuals. 160 and 126 individuals contained deletions and duplications respectively. Average CNV size was 20,479bp. Most of CNVs within 19q13.2 chromosome locus were present in the DGV control dataset and also SFARI CNV control dataset. Therefore, manual screening of 3q26.31 locus identified only 4 genes were absent in controls and therefore were considered for further analysis. The number of individuals carrying each of these 4 genes was identified and percentage population frequency for the gene was calculated. The locus hosted 2 autism associated gene NLGN1 and NAALADL2 with significant population frequency of 1.71% and 1.19% respectively.

The population frequency of locus 9p23 was 2.34% and was in 226 individuals with deletions in 138 and duplications in 88 individuals. Average CNV size was 95302bp. However, almost all of CNVs within 9p23 chromosome were present in the DGV control dataset and also our SFARI CNVs control dataset. On manual screening of 9p23 chromosome locus we identified only 2 genes. These genes were absent in controls and therefore were considered for further analysis. PTPRD gene was present in 0.47% of the population frequency.

13q12.11 locus has a population frequency of 2.13% and this loci present in 206 individuals. However the most of CNVs within 13q12.11 chromosome were present in the DGV control dataset and also our SFARI CNVs control dataset. On manual screening of 13q12.11 chromosome locus we identified only 25 genes. These genes were absent in controls and therefore were considered for further analysis. The number of individuals

carrying each of these 25 genes was identified and population frequency was calculated. We identified SKA3 gene to be present in 1.42% of the population.

14q11.2 the eleventh locus in GP1, had a population frequency of 2.04% and this locus present in 197 individuals with 103 and 94 individuals with deletion and duplication. Average CNV size was 242,370bp The locus has one autism associated gene CHD8 with a population frequency of 0.05% i.e it was present in 5 individuals only. On manual screening of 14q11.2 locus we identified only 74 genes in outside control CNVs and therefore were considered for further analysis. We did not get any significant genes in the chromosome locus 14q11.2.

Chromosome locus 4q13.2was identified in 2% of the SFARI ASD cohort i.e in 193 individuals with 53 individuals with deletions and 140 individuals with duplications. Average CNV size was found to be 95,580bp. Manual screening of 4q13.2 locus did not identify any gene significantly present in SFARI ASD case cohort nor did it contain any autism associated gene.

Locus 22q11.21 with population frequency of 1.90% was present in 192 individuals. There were 122 and 70 individuals with deletion and duplications respectively. Manual screening of 22q11.21 locus against control datasets retained only 61 genes for further analysis. The number of individuals carrying each of these 61 genes was identified and population frequency was calculated. We identified two significant genes PI4KA and RTN4R to be present in 0.93% and 0.72% of the population. The locus contained 6 autism associated genes and was previously implicated in playing an active role In causing ASD.

16p13.3 locus was present in 190 individuals i.e. 1.97% of the SFARI ASD cohort. The locus contained 78 and 122 individuals with deletions and duplications respectively. Average CNV size was 209,483bp. Manual screening of 16p13.3 locus identify any genes that was present in a significant number of individuals of the SFARI ASD cohort but the locus contained 2 autism associated genes RNPS1 and CACNA1H.

Locus 5q33.1 is also present in 190 individuals with percentage population frequency of 1.97%. 128 individuals had deletion event in this locus and 62 individuals showed duplications. There were two autism associated genes in this locus. However, manual screening of locus we did not identify any genes significantly present in the SFARI ASD cohort.

Discussion

The study hypothesis that investigating CNV influences on Autism Spectrum Disorders (ASD) in a large disease cohort would identify active participants of disease etiology. SFARI database provided us with CNV findings on ASD and other neurological disorders in more than 16,000 individuals. The stringent inclusion-exclusion criteria used ensured that only ASD cases with or without comorbidities were included in our study cohort named SFARI ASD cases. This resulted in a study cohort with 9,635 individuals with ASD. The most common CNVs present in more than 2% of these individuals were identified under group1 (GP1) category. These CNV loci were expected to host active genetic players leading to autism. Additionally, using DGV and SFARI control datasets as our internal control has resulted in a stringent list of genes probably involved in brain development and removal of genes under the influence of CNVs in the general population.

Locus 15q11.2 was identified in highest number of individuals in SFARI ASD cohort i.e 6.63%. However, majority of the region within this locus was found to be present in DGV control dataset. Off the 77 genes outside the influence of control CNVs, only UBE3A was autism associated gene. The remaining genes viz. MAGEL2, NDN, SNURF, MKRN3, NPAP1, IPW, critical genes for PWS likeSNORD64, SNORD108, SNORD109A, SNORD109B, SNORD116 and SNORD115 clusters (Dykens et al., 2011), along with GOLGA6L1 and GOLGA6L2 were present in less than 0.8% of the population. Genes from this region are implicated in PraderWilli Syndrome (PWS) and Angelman Syndrome (AS) which are both caused by 15q11-q13 deletion but differ by their parent of origin. Individuals with PWS and AS are known to exhibit autistic features. Patients with MAGEL2, MKRN3, NDN deletions often showed intellectual disability, obesity and an ASD diagnosis (Schaaf et al., 2013). CNV involving UBE3A gene was found in 0.84% of the population. Maternal deletions of UBE3A is known to cause AS, however maternal duplications are associated with autism (Glessner et al.,2009).

Table 5: Genes within CNVs identified in locus 15q11.2 and their relevance to Autism:

Genes in 15q11.2	% population frequency	Associated conditions	Relevance to autism
UBE3A	0.84	Angelman Syndrome (Deletion of maternal origin)	Maternal duplications associated with autism and cognitive deficits
MAGEL2, NDN	0.75	PraderWilli (PW) Syndrome (Deletion of paternal origin)	19% of individuals with PWS met diagnostic criteria for ASDs (Descheemaeker, Mie-Jef, et al. 2006)
SNURF, MKRN3	0.69		
NPAP1	0.45		
IPW SNORD64, SNORD108, SNORD109A, SNORD109B SNORD116 & SNORD115 clusters,	0.32		
GOLGA6L1	0.30		
GOLGA6L2	0.26		
MIR4508	0.20		

CNVs in locus 11q11, although present in 456 individuals i.e. 4.73% of the SFARI ASD cohort, contained several groups of Olfactory Receptor (OR) genes. Only 24 OR genes were outside CNV regions of control but they were present in less than 4 individuals of the SFARI ASD cohort. Therefore locus 11q11 did not contain any significant gene known to be involved in autism. However, role of OR genes and non-coding regions of 11q11 locus could be investigated further.

16p11.2 deletion and duplications were found to produce different effect on individuals and also affect different pathways. For example, the deletion alters the expression of 291 genes and the duplication affects 665, with very little overlap — only 6 genes — between the two (Luo, Rui, et al.2012). Vast majority of genes in the 16p region affect brain development (Blaker et al. 2012). A mouse model lacking one copy of the 16p11.2 region exhibited a tendency to run in circles around their cages, also reminiscent of repetitive behaviors associated with autism.

Locus 1q44 contained 53 genes outside the influence of control CNVs. Although OR2T29 gene was found in 1.36% of the population, the role of OR genes in autism is yet to be understood. Rare variants in the OR1C1 gene have been identified with autism (Bucan et al., 2009). Two other genes KIF26B and SMYD3 were found in 0.53% and 0.2% of the population and microdeletions involving these two genes are associated to microcephaly (Raun et al., 2017).

Of the 30 genes outside control CNVs, locus 5p15.33 contained only one significant gene ZDHHC11 gene with a significance presence of 2.15% of the case population. Two reports implicating ZDHHC11 to involve in autism were identified (Marshall et al., 2008; Firouzabadi et al., 2016) however, elucidation of the molecular role in ASD remains. Additionally, genetic factors causing Cri du Chat Syndrome has been localized to this locus.

19q13.2 locus is present in 2.8% of the population but after screening against control datasets, none of the 48 genes outside control CNVs were present in more than 6 individuals of the SFARI ASD cohort. The two autism associated genes ACTN4 and CAPN12 that were present in just 2 and 3 individuals too failed to attain significance. However, ACTN4 gene was originally identified as an ASD candidate gene based on its enrichment in an autism-associated protein interaction module; sequencing of post-mortem brain tissue from 25 ASD cases resulted in the identification of significant non-synonymous variants in this gene with an expected false-positive rate at 0.1, confirming the involvement of this module with autism (Li et al., 2014). ACTN4's role in ASD is supported by 2 reports. CNVs within 13q12.12 region were found in seven cases from a study of 15,749 individuals from the International Standards for Cytogenomic Arrays (ISCA) consortium with unexplained developmental delay, intellectual disability, dysmorphic features, multiple congenital anomalies, autism spectrum disorders, or clinical features suggestive of a chromosomal syndrome (Kaminsky et al., 2011), as well as in an ASD case (Prasad et al., 2012). Additionally, Grisworld et al. (2012) observed deletions >1 Mb at 13q12.12 locus exclusively in cases.

Wenger et al. (2016) and Ousley et al. (2017) while Examining the Overlap between Autism Spectrum Disorder and 22q11.2 Deletion Syndrome concluded that strictly defined ASDs occur in a substantial proportion of individuals with 22q11.2DS and has a high rate of ASD at 14–25 %, among the highest of any genetic disorder. Therefore they recommend that all individuals with 22q11.2DS be screened for ASDs during early childhood. This locus has 6 autism genes within. Clathrin is the major protein of the polyhedral coat of coated pits and vesicles. Two different adapter protein complexes link the clathrin lattice either to the plasma membrane or to the trans-Golgi network. A homozygous mutation in the CLTCL1 gene was found to segregate perfectly with disease in a multiplex ASD family. No homozygotes for his mutation were observed in 1328 control chromosomes. Seven additional compound heterozygous mutations in the CLTCL1 gene were identified in ASD cases from the replication cohort that were not observed in 371 controls (Chahrour et al., 2012). Rare mutations in the GNB1L gene, including a translocation, duplications and missense variants,

have been identified in patients with ASD (Chen et al., 2011). This gene encodes a G-protein beta-subunit-like polypeptide which is a member of the WD repeat protein family. Members of this family are involved in a variety of cellular processes, including cell cycle progression, signal transduction, apoptosis, and gene regulation.

Two studies by Levy et al. and Sanders et al. published in *Neuron* in June 2011 report associations between mutations in several regions of 16p and autism. Researchers found two *de novo* and two inherited duplications at 16p13.2 and a duplication on 16p13.11. They also found seven deletions and four duplications in the 16p11.2 region and report that *de novo* mutations seem to recur at 16p13.2 and 16q13.3. Mostly Chromosome 16p13.3 duplication is linked to autism along with other primary diagnosis like developmental delay, speech delay, joint abnormalities, and characteristic facial features. Duplications involving RNPS1 were statistically enriched (P=0.003671) in a cohort of 57,356 patients with neurodevelopmental disorders compared to a cohort of 20,474 controls. A de novo 6.03 kb duplication encompassing the RNPS1 gene was identified in a female patient from the Signature Genomic Laboratories database (GC62115) with autistic disorder, developmental delay, dysmorphic features, seizure disorder, and multiple congenital anomalies (Nguyen et al., 2013). Rare mutations in the CACNA1H gene have been identified with autism. In one study, missense mutations in CACNA1H were found in 6 of 461 individuals with ASD (Splawski et al., 2006).

Although5q33.1 does not have a direct association with ASD, two genes USP7 and TSC2, within these loci have been evidenced to play major roles in causing autism. TSC2 gene within this locus causes Tuberous Sclerosis, a syndromic form of autism. *De novo* variants (six multigenic deletions, one nonsense variant) affecting the USP7 gene were identified in seven patients presenting with developmental delay/intellectual disability, with five of these cases having an additional diagnosis of ASD (Hao et al., 2015). Subsequent statistical analysis determined that there was a strong correlation between the seven patients having both a de novo USP7 variant and the expressed phenotype that was likely not due to chance ($p<0.0001$). TSC2 gene has been associated with syndromic autism, where a subpopulation of individuals with a given syndrome develop autism. In particular, genetic association has been found between autism and tuberous sclerosis (and hence the TSC1 and TSC2 genes as well). As well, association with TSC2 and autism has been found in an AGRE cohort (Serajee et al., 2003), and a rare mutation in TSC2 has been identified in an individual with ASD (O'Roak et al., 2012).

OBJECTIVE C

To construct and establish various genetic pathways and molecular interaction network affected by these DNA variants in ASDs.

C. i) Pathway Analysis for the CNV encompassed genes

Autism-CNVs pathways were constructed using IPA using the genes enriched in pathways analysis. We stipulated that at least two genes in a pathway must be disrupted for a pathway to be considered enriched, and used the minimal cut sets concept (MCSs) (Veerappa et al., 2013). The interconnected pathway was built from a list of most prevalent 25 causal genes in study populations. This pathway can be divided into seven sub-pathways: 1) Clusters of *CACNA1* genes consisting of *CANCA1H, CACNA1C, CACNA1B* and *CACNA1I,* involved in voltage gated calcium channel signaling has crucial functions that, when perturbed, may contribute to psychopathological susceptibility to autism. (Terwindt et al.,1998). 2) *DPP6* is shown to be involved in the physiological processes of brain function modulating the cell surface expression and the activity of the potassium channel KCND2. Rare mutations in the *DPP6* and *KCND2* have been identified with autism (Lee et al., 2014; Buxbaum et al., 2009). 3) *DLG1* encodes a multi-domain scaffolding protein required for normal development, having a role in septate junction formation, signal transduction, synaptogenesis. 4) *UBE3A* functions both as an E3 ligase and as a transcriptional co-activator in the ubiquitin proteasome pathway with preferential maternal-specific expression in brain and, more specifically, in neurons, (Dindot et al., 2008). Duplications in the 15q11-13 region bearing UBE3A are known to manifest autism (Scoles et al., 2011). 5) Protein encoded by *CHRNA7* forms a homo-oligomeric channel, displaying marked permeability to calcium ions and is a major component of brain nicotinic receptors. Haploinsufficiency of *CHRNA7* is causative for the majority of neurodevelopment phenotypes observed in the 15q11.2-13.3 micro-deletion in autism. 6) Given its involvement in the inhibition of excitatory neural pathways and its expression in early development, *GABRG3* is strongly implicated in the pathogenesis of autism. 7) *NIPA1* encodes a magnesium transporter associated with early endosomes and the cell surface in a variety of neuronal and epithelial cells which play a role in nervous system development and maintenance. Studies in mouse models point towards a contribution

of micro duplications at chromosome 15q11.2 to autism, and highlight *CYFIP1* and *NIPA1* as autism risk genes functioning in axonogenesis and synaptogenesis.

Pathway analysis of the autism gene-CNVs revealed a total of 12 miRNAs namely miR-499-3p, miR-879-5p, miR-4440, miR-4796-5p, miR-210-3p, miR-4724-5p, miR-501-5p, miR-513a-5p, miR-578, miR-1243, miR-4713-3p and miR-1238-3p (Figure 3). These miRNAs play a role in regulating genes involved in neurodevelopment, neurotransmission, and synapse. Various miRNAs were found regulating several autism genes pathways, for instance, targets of miR-499-3p are IMM2PL, UBE2G1, CACNA1B, APBA1 and DLGAP2 mRNAs while targets for miR-513a - 5p include CACNA1B, NIPA1, DUSP22, KCND2 and CACNA1C mRNAs.

When the MCSs were computed with respect to "physical and genetic" interactions in these pathways, the experimental block of essential genes inevitably lead to mutants. Thus, the establishment of autism genes CNVs enrichment pathway led to identification of some relatively new genes which show promises of its role in autism pathogenesis. These include *CASKIN1, KCNIP1, KCND2* and *miRNAs hsa-mir-513a-5p, hsa-mir-4796-5p, hsa-mir-501-5p, hsa-mir-4724-5p, hsa-mir-578, hsa-mir-1243, hsa-mir-1238-3p, hsa-mir-4671-3p, hsa-mir-4473, hsa-mir-879-5p, and hsa-mir-499-3p,* which are closely linked to known autism causal genes and may be strong indicators as candidate genes.

Identification of recurrent CNVs in the normal cohorts compared to rare pathological CNVs in cases provide another dimension to assess the role of primary sites towards the sensitizing of the second-site for manifesting autism.

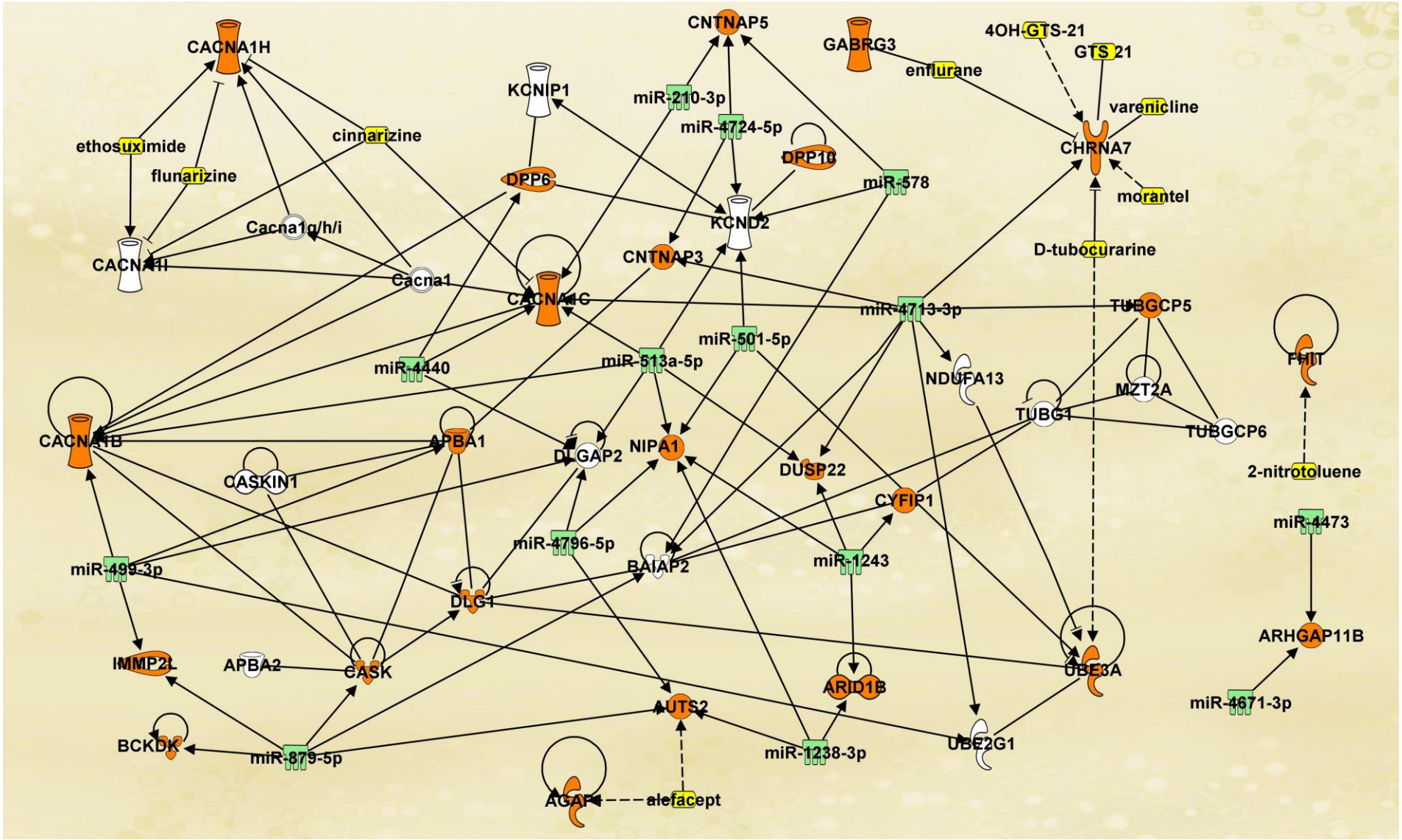


Figure 4: Ingenuity Pathway Analysis of autism genes enriched under CNVs.

ii. Role of *KMT2C* and *CNTNAP2* in autism pathophysiology

Various loss of function mutations have been reported for nucleus localized gene *KMT2C*. It is a chromatin marker associated mediating mono- and tri-methylation of histone H3 at lysine 4 (H3K4me1 and H3K4me3). Various studies have investigated the role of *KMT2C* in specific knockdown of 'trr,' in its *Drosophila* ortholog pointed out neuropathological aspects such as impaired short memory, intellectual disability and the mis regulated genes in underlying molecular processes: neuronal migration and synaptic plasticity (Koemans et al., 2017).

Another key player in synaptic plasticity is *CNTNAP2*, localized at myelinated axons associated with potassium channels. It functions in the nervous system of vertebrates as cell adhesion molecules and receptors. Rare variants in CNTNAP2 gene, including deletions and nonsynonymous changes, have indicative roles in autism, Intellectual Disability, Developmental Delay and language impairment (Li et al., 2010). Several studies provide genetic evidences for CNTNAP2 gene to be closely related to altered gene expression in autism brain (Sampath et al., 2013). *CNTNAP2* directly interacts with *FOXP2* suggesting a link between language impairment and autism manifestation circuital pathway (Vernes et al., 2008). Therefore, rare mutations in *KMT2C, CNTNAP* and *CACNA1H* could be the key players in manifestation of autism. These genes could be looked int detail at the molecular level for further insights in the underlying pathways.

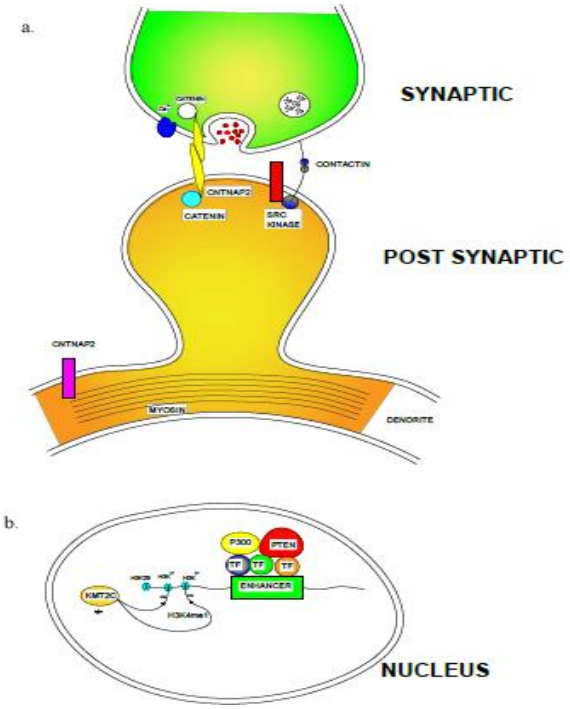


Figure 5: Pathway Prediction for the damaging mutant genetic variants *CNTNAP2* & *KMT2C*

iii. Delineating the autism pathway using candidate genes identified in 15 whole exome sequence data from Indian autism subjects

The present study comprises a detailed study of the whole exome sequence data in 15 autistic subjects of Indian Origin. The probands showed variants to be present in different concentration in terms of chromosome. The highest number of variants was observed in chromosome 1. Chromosome 15 and 16 which are known to habour the loci of autism genes also contains considerable number of variants. Also, the variants type present in the autistic subject comprised of missense, insertion deletions, 3' and 5' UTR mutations, stop gain- stop loss mutations. Out of all the variant types identified, the focus of the work was on stop gain-stop loss and missense mutations as these are the most damaging mutations having adverse implications on the manifestation of any disease. Further filtering and applying stringent criteria to identify high confidence genes, several deleterious and damaging single base and Insertion-deletion mutations in 33 autism genes were obtained. Out of which 3 main genes were studied in detail. *PTEN* gene showed stop-gain mutations in all family members. Similarly, *CHD8* gene showed stop-loss mutation in all family members. *CNTNAP2* gene showed missense mutation in the proband and the unaffected sibling.

High risk autism genes were obtained and tabulated from extensive literature survey and various autism databases. A total of 62 genes are known to be damaging and deleterious which play a key role in autism manifestation. Various pathways are impaired in this case for say, DNA helicases, voltage gated channels, methyltransferases, chromatin remodeling pathway and many more. Each of these pathways is involved in manifestation of autism at varied points in the pathway network. Multiple mutations belonging to the same pathway makes the process go hay away completely and impairs the downstream process completely. Chromosome 2, 3 and 15 have the highest burden of mutations which is in par with the global chromosome burden for autism. Out of the 62 genes identified as high risk autism genes, 24 genes carried mutations in the whole exome sequences of autistic subjects of Indian origin. *PTEN, ANK2, ANKRD11* and *CNTNAP2* are present in all subjects having various mutations as single point and multivariant polymorphism along with structural variants and copy number variants. When looked into all of these mutations individually, it results in a list of homozygous and heterozygous mutations both novel and known in literature. These mutations converge into similar pathway clusters and results into autism manifestation. Presence of Single base , multi base, structural and copy number variations in the same gene

PTEN make it damaging and a master gene which when impaired collapses the whole system and no downstream process takes place anymore.

Indian samples variant list shown a deviation from the high risk autism genes identified in global whole exome sequence datasets of autism cases with similar phenotype. This is an interesting finding which helps the author hypothesize that Indian subjects have unique gene mutations leading to autism and a study is the need of the hour to identify them in order to diagnose and treat and manage autism better.

All the genes were each looked for pathway relevant to autism and its functionality impaired due to mutation in relevance to autism and a network was constructed and checked if they interacted physically. All the genes obtained were converged into three main pathways.

The study was focused after all the screening and filtering if genes based of haploinsufficiency, pathogenicity and other factors and a question was posed: Is PTEN the master gene for autism manifestation and how does the pathophysiology of the same work. The PTEN gene structure was studied in detail and the protein domains were identified and gene mutation regions were delineated and characterized in silico. Physically interacting functional proteins of PTEN were looked into and many of them were involved in autism directly or involved in pathways relevant to autism.

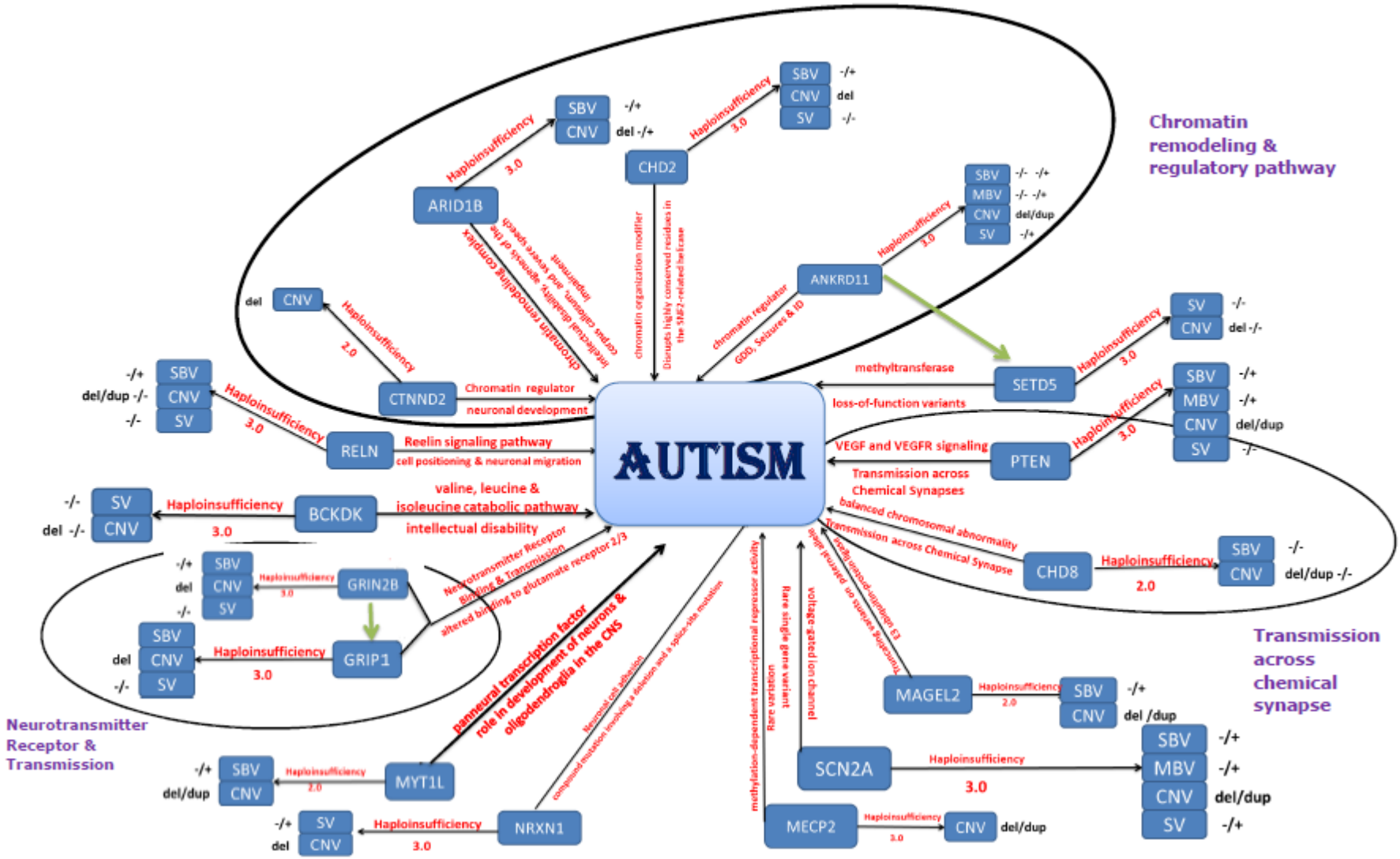


Figure 6: Pathway for recurrently mutated genes in Autism

OBJECTVE D

Validation and annotation of the newly identified probable candidate genes obtained from the genetic pathways and molecular interaction network studies in ASD subjects.

D. i) Role of *CNTNAP2* in the manifestation of autism outlines the regulation of signalling between neurons at the synapse

The present study reports the identification of 58 high risk genetic variants in the whole exome sequence datasets out of 63 high risk autism genes. The variants type present in the autistic subject comprised of non-synonymous, nonframeshift substitution, stopgain, frameshift deletion, frameshift insertion, stoploss and missense mutations. Out of all the variant types and genes identified, the focus of the work was on the non-synonymous mutations of *CNTNAP2* gene which was studied in detail, as it encodes a cell adhesion protein that regulates signaling between neurons at the synapse, which is one of the vital processes impaired in autism. Out of 180 whole exome sequence datasets, *CNTNAP2* gene variants were present in 164 whole exome sequence datasets. The common polymorphism at the same location was identified among 163 cases in *CNTNAP2* in which the alteration of the nucleotide A to C at the position 1765 in the exon 11, results in the amino acid change from Threonine to Proline at 589 positions. Similar gene has been mutated in 15 Indian samples, but the exact gene variant is missing. The variant was identified in the exon 3, alteration from A to G at the position 286, results in amino acid change from Serine to Glycine at 96 positions.

Table 6: Varied *CNTNAP2* gene variants present across the Indian & Global samples

Sample Type	**Ref/Alt**	**AA Change**	**Allele Frequency**
Indian	A/G	CNTNAP2:NM_014141:exon3:c.A286G:p.S96G	0.0008
Global	A/C	CNTNAP2:NM_014141:exon11:c.A1765C:p.T589P	0.216

Discussion

The whole exome sequence analysis revealed variations in the protein coding regions in relevant autism genes. A custom developed pipeline was used that stringently filtered and provided significant variants having higher probability of causing autism. A total of 58 high

risk autism genes were found which have damaging/ deleterious mutations. Among them, *CNTNAP2* gene variant have non synonymous SNV mutations in 164 cases. Out of which, 163 cases has the common polymorphism at the same location in the nucleotide position 1765 where change from A>C, resulted in amino acid change from threonine to proline. Among 15 Indian samples, the variant was in the nucleotide position 286 where change from A>C, resulted in amino acid change from serine to glycine.

Mutations in genes for autism were found mostly in synaptic plasticity process. Several risk genes that are key regulators of synaptic plasticity have been implicated in autism. Indeed, many of the risk genes that have been linked to autism encode synaptic scaffolding proteins, receptors, cell adhesion molecules or proteins that are involved in chromatin remodelling, transcription, protein synthesis or degradation, or actin cytoskeleton dynamics (Rubeis et al., 2014). Changes in any of these proteins can increase or decrease synaptic strength or number and, ultimately, neuronal connectivity in the brain. In addition, when deleterious mutations occur, in efficient genetic buffering and impaired synaptic homeostasis may increase an individual's risk for *ASD.*

One such notable gene *CNTNAP2* found in 164 cases, contained novel damaging mutations. *CNTNAP2* is one of the largest mammalian genes that spans more than 2.3 Mb at the chromosomal region 7q35-q36.1, which encodes CAM that regulates signaling between neurons, highly expressed in neurons that control language and language development difficulties. It encodes CASPR2 with expression restricted to neurons, which is a trans membrane scaffolding protein belonging to the neurexin superfamily that clusters voltage gated potassium channels at the Nodes of Ranvier. It plays a major role in ‘language development’ in autism and it is highly expressed in a cortico–striato–thalamic circuit that is involved in diverse higher order cognitive functions. Pathway analysis revealed synaptic pathways which have direct implication to autism. The direct physical interactions between the new genes identified and the known candidate genes indicate a possible involvement of these genes in regulation of function of the candidate genes. Hence, the identified novel genes might have a role to play in the development of autism. Autism is not just phenotypically heterogenous but also genotypically varied and complex. This project attempted to decipher this genotypic heterogeneity using multiple approaches such as Whole Genome genotyping using SNP Array to identify Single Nucleotide Polymorphisms (SNPs) and Copy Number Variations (CNVs), Whole Exome Sequencing using Next Generation Sequencing technology to fish out polymorphisms in coding regions as well as the Un-Translated Regions (UTRs) of the genome

and additionally the homozygosity explored the impact of Gut-Microbiome on the behavioral manifestations of ASD.

ADDITIONAL STUDY E

Homozygosity mapping analysis on autism cases mapped against controls

Homozygosity mapping identifies autism candidate genes in causative pathways (Manuscript summitted)

Sample Selection: The study included a total of 426 unrelated probands analyzed along with 232 parents (116 trios). The samples were obtained from an international public repository Gene Expression Omnibus Database (GEO) that archives and freely distributes microarray datasets (accession GSE9222) genotyped for unrelated ASD index cases using the Affymetrix GeneChip® 500K single nucleotide polymorphism (SNP) mapping array. DNA derived from blood or lymphopblasts was hybridized to Affymetrix GeneChip® 500K mapping arrays [Early Access Array - The GeneChip® Human Mapping 500K Array Set is comprised of two arrays (Nsp/Sty arrays)]. Nsp and Sty arrays were analyzed for CNV using a combination of DNA Chip Analyzer (dChip), Copy Number Analysis for GeneChip (CNAG), and Genotyping Microarray based CNV Analysis (GEMCA) (Marshall et al. 2008).

Categorization of samples: Out of 426 probands and 232 parents, 334 probands and 122 parents (61 trios) were scrutinized based on the type of microarrays used to perform Genome variation profiling and SNP genotyping, only the samples genotyped using Nsp and Sty arrays were selected and analyzed.

Results

An integrated approach of whole genome genotyping and homozygosity mapping on 334 autism cases mapped against 167 controls revealed the presence of 38 homozygous regions. The genome wide homozygosity data displays scored blocks of haplotype in bar chart in comparison with controls highlighted in red bars. It depicted the most promising genomic haplotype regions which have been passed on across generations. Analyzing the collated data from the homozygous blocks of the cases meeting the criteria for (a) identifying novel autism candidate gene expression in brain, (b) should participate in neuronal development, (c) interaction with known autism genes, (d) non-homozygous in controls, (e) de novo in origin, (f) overlapping in two or more unrelated samples, (g) recurrent in two or more unrelated samples,

(h) involved in expression of brain development and (i) participating in neuronal migration, axon growth, neuritis outgrowth, synaptic plasticity and cell adhesion were selected for present study. We observed a total of 38 homozygous regions bearing 297 genes participating in various processes such as Adherens junction, Dipeptidase activity, Platelet-derived growth factor and receptor binding, Vitamin digestion and absorption, syntrophin complex, delta-catenin binding and many physiological pathways relevant to autism. While the remaining were pseudo genes, however the homozygosity blocks analysis in cases revealed the presence of recessive genes partaking in neuronal development. These homozygous regions contained 52 autism candidate genes.

Identification of risk homozygous haplotypes

Homozygosity mapping for 61 trios with 122 controls on Affymetrix Mapping 250K Nsp SNP Array revealed a total of 12 homozygous regions with a maximum score of 5714, within the maximum block length of 1000 through the Homozygosity Mapper. These homozygous regions were found bearing 29 genes in chromosomes 2, 4, 7, 8, 11 and 12. Applying the criteria for candidate gene selection revealed the presence of 6 homozygous regions on chromosomes 2, 8, 11 and 12 bearing *SNTG1, KITLG, GLULP6, FOLH1, PTPRJ, CEP290* and *KITLG* genes which are involved in autism with homozygosity score of 5053, 5714, 4645, 4603, 4592 and 5714. Similarly, homozygosity mapping on genotype of 61 trios with 122 controls using Affymetrix Mapping 250K Sty SNP Array revealed a total of 11 homozygous regions revealing a maximum score of 4827, within the maximum block length of 1000 in the chromosomes 2, 8, 9, 10 and 16 comprising of 102 genes. Interpretation of the obtained result and applying the conditions for candidate gene selection revealed the presence of 6 homozygous regions of chromosomes 2, 8, 10 and 16 bearing *ANKRD26P1, HERC2P4, HERC2P8, SHCBP1, SLC6A10P, TP53TG3, TP53TG3B, UBE2MP1, ZNF267, IGHV3OR16-13, ABCC12, NETO2, FNTA, GLULP6* and *P4HA1* genes related to autism with homozygosity score of 4053, 4217, 3970, 4478, 4821 and 4827.

Homozygosity stretches resulting in a maximum score of 26206, within the maximum block length of 1000 were generated from the Homozygosity Mapper for 273 autism cases with 45 HapMap Yoruba controls using Affymetrix Mapping 250K Nsp SNP Array. There were nine homozygous regions in the chromosomes 8, 9 and 11 comprising of 28 genes with homozygosity score of 26206, 24275, 23087, 21801, 21778, 21627, 21374, 21335, 21155, 26206, 24275 and 23087. With respect to the criteria for candidate gene selection revealed the

presence of *SNTG1, PTPRJ, OR4B1, FOLH1, OR4X2, OR4X1, OR8H2 and TRIM51* genes which plays an significant role in autism. Analyzing the chromosomal regions in neuronal development revealed two homozygous regions in chromosome 1.

Homozygosity stretches resulting in a maximum score of 22373, within the maximum block length of 1000 were generated from the Homozygosity Mapper for 273 cases with 45 HapMap Yoruba controls using Affymetrix Mapping 250K Sty SNP Array. There were six homozygous regions in the chromosomes 3, 11 and 16 comprising of 145 genes with homozygosity score of 22373, 22313, 18529, 18449, 18268 and 18022. Focusing on the chromosomal regions involved in neuronal development revealed 3 homozygous regions in chromosome 3. Applying the criteria for candidate gene selection revealed the presence of 15 genes *IQCF2, IQCF1, RRP9, PARP3, GPR62, PCBP4, ABHD14B, ABHD14A, ACY1, RPL29, DOCK3, NETO2, UBE2MP, SHCBP1* and *ANKRD26P1* involved in autism.

Table 7: Comparison of Homozygous Haplotype regions shared in autism simplex families and autism cases

Sample Type	**Chr**	**Start(bp)**	**End(bp)**	**Start(SNP)**	**End(SNP)**	**Genes**	**Score***
Autism Trios using Affymetrix Mapping 250k Nsp SNP Array	8	50280717	50921942	rs1914995	rs13270712	*SNTG1*	5714
	12	88808147	89252667	rs7488412	rs11105085	*KITLG*	5154
	2	194665045	195006184	rs11687738	rs7587288	*GLULP6*	5053
	11	49038697	49228613	rs7112529	rs202675	*FOLH1*	4645
	11	48086151	48110043	rs12292520	rs17789481	*PTPRJ*	4603
	12	88459309	88681867	rs2471510	rs1558737	*CEP290,KITLG*	4592
Autism cases Yoruba population genotype as Controls using Affymetrix Mapping 250k Nsp SNP Array	8	50280717	50808334	rs1914995	rs1914995	*SNTG1*	24275
	11	49038697	49295867	rs7112529	rs7112529	*FOLH1*	21801
	11	48168664	48293931	rs3741410	rs3741410	*PTPRJ*	21778
	4	33158269	33222880	rs9917850	rs9917850	*PTPRJ*	23087
Autism Trios using Affymetrix Mapping 250k Sty SNP Array	16	31716601	46710869	rs9938037	rs9938037	*SLC6A10P, TP53TG3, TP53TG3B*	4827
	16	47190596	48492861	rs12051064	rs12051064	*ABCC12*	4821
	2	194458906	195180067	rs7582395	rs7582395	*GLULP6*	4053
Autism cases, Yoruba population genotype as Controls using Affymetrix Mapping 250k Sty SNP Array	3	51880614	52047421	rs7615151	rs7615151	*PARP3,GPR62, PCBP4,ABHD14B,* ***ABHD14A****,ACY1*	18268

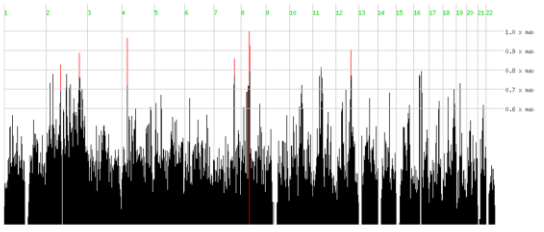
* Threshold level of significance (P=0.05) is scored at 4000 and above

Table 8: eQTL analysis for the homozygous regions identified across samples

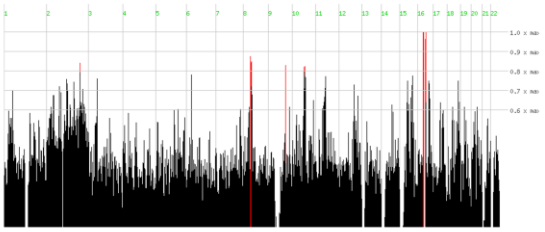
Sample Type	rs ID	SNP Position	RDB Score	Ref/Alt	Bound proteins	Affected Autism Genes
Autism Trios using Affymetrix Mapping 250kNSp SNP Array	rs11104868	chr12:88828441	1d	G/A	CBX3,FOS,BACH1,MYC, JUND,MAFK,POU2F2,RBBP5 ,RUNX3, SMARCA4	*CEP290*
	rs7979666	chr12:88843826	1f	C/T	-	*CEP290*
	rs11104939	chr12:88932363	3a	T/C	CEBPB	*KITLG*
	rs12292520	chr11:48086150	2b	G/T	POLR2A	*PTPRJ*
	rs17789481	chr11:48110042	2b	A/G	EP300,FOXA1,GATA3,AR	*PTPRJ*
Autism cases Yoruba population genotype as Controls using Affymetrix Mapping 250kNSp SNP Array	rs1681625	chr11:48020542	2b	A/G	HNF4A,ELF1,HNF4G,MYBL2,NFIC, POLR2A,SP1,SREBF1,CDX2,FOXA1, JUND,USF1,POLR2A,TBP	*PTPRJ*
	rs12292520	chr11:48086150	2b	G/T	POLR2A	*PTPRJ*
Autism Trios using Affymetrix Mapping 250k Sty SNP Array	rs7078127	chr10:74847633	1f	A/C	-	***P4HA1, NUDT13, MRPS16***
	rs11866251	chr16:48227861	3a	G/A	-	*ABCC11*
	rs16945839	chr16:48160409	3a	C/T	CTCF	*ABCC12*
Autism cases Yoruba population genotype as Controls using Affymetrix Mapping 250k Sty SNP Array	rs17051043	chr3:50693998	1d	C/T	GATA2	*HEMK, HEMK1*
	rs17051041	chr3:50757396	1f	A/G	EZH2	*DOCK3,UBA7, MAPKAPK3*
	rs2675835	chr3:50862740	1f	G/A	-	*DOCK3,MAPKAPK3*

2.

a).



b).



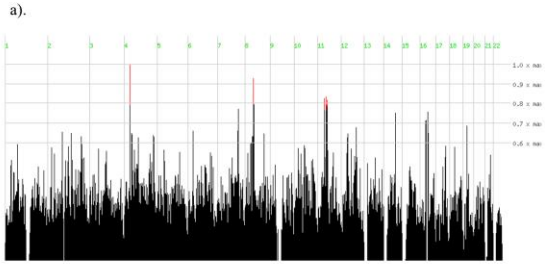
c).

Sample	Cases	Controls	Mapping Array
I	61	122	Affymetrix Mapping 250K Nsp SNP Array
II	61	122	Affymetrix Mapping 250K Sty SNP Array

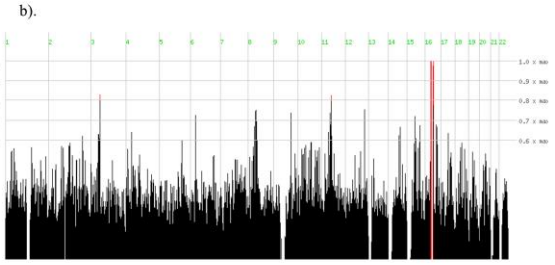
Figure 7: Genome-wide homozygosity mapping with significant homozygous regions marked in red for 61 cases mapped against 122 controls using Affymetrix Mapping 250K Nsp and Sty SNP Array

2.

a).



b).



c).

Sample Type	Cases	Controls	Mapping Array
I	273	Unrelated 45 (Hapmap:Yoruba)	Affymetrix Mapping 250K Nsp SNP Array
II	273	Unrelated 45 (Hapmap:Yoruba)	Affymetrix Mapping 250K Sty SNP Array

Figure 8: Genome-wide homozygosity mapping with significant homozygous regions marked in red for 273 cases mapped against HapMap Yoruba using Affymetrix Mapping 250K Nsp and Sty SNP Array

eQTL analysis of Homozygosity Blocks: eQTL analysis of the homozygous blocks revealed the presence of several polymorphisms affecting many downstream genes, with scores ranging 1a- 1f affecting the binding of regulatory proteins. These deregulated the expression of downstream genes important for brain development. eQTL analysis of the homozygous blocks in a) 61 trios

with 122 controls on Affymetrix Mapping 250K Nsp SNP Array revealed the presence of several polymorphisms including rs11104868, rs7979666, rs6963478, rs12292520 and rs17789481 with a RegulomeDB score of 1d, 1f and 2b affecting the binding of few regulatory proteins deregulating the *CEP290, KITLG and PTPRJ* genes.

Several polymorphisms were identified for rs7106648, rs1681625 and rs12292520 in b) 61 trios with 122 controls on Affymetrix Mapping 250K Sty SNP Array. These revealed the presence of several polymorphisms including rs7078127, rs8050306, rs7500312, rs11866251 and rs16945839 holding a RegulomeDB score of 1f, 2b and 3a. These affected the binding of few regulatory proteins deregulating the *P4HA1, NUDT13, MRPS16, ABCC11 and ABCC12* genes. eQTL analysis in the homozygous blocks of c) 273 autism cases with 45 HapMap Yoruba controls using Affymetrix Mapping 250K Nsp SNP Array with a regulomeDB score of 2b affecting the binding of few regulatory proteins, deregulating one of the autism candidate gene PTPRJ. eQTL analysis in the homozygous blocks of d) 273 autism cases with HapMap Yoruba controls using Affymetrix Mapping 250K Sty SNP Array revealed the presence of several polymorphism including rs17051043, rs17051043 and rs2675835 with a regulomeDB score of 1d, 1f and 2b affecting the binding of few regulatory proteins. These deregulated the *HEMK, HEMK1, DOCK3, UBA7 and MAPKAPK3* genes which were found to express in the brain from the developmental stage.

These were expressed in developmental stages of the brain. Identified candidate genes and the affected downregulating genes containing SNPs were also playing a crucial role in some of the physiological and neuronal developmental pathways such as dipeptidase activity, Adherens junction, Vitamin digestion and absorption, Platelet-derived growth factor receptor binding and syntrophin complex pathways. Significant overrepresentation of these genes was observed in several physiological pathways relevant to autism such as Arginine and proline metabolism, protein farnesylation, protein geranylation, protein prenylation pathways important for neuronal development, aminoacylase activity, hydrolase activity, Urea cycle and metabolism of amino groups, GTPase binding. The cases display a much higher degree of haplotype sharing within overlapping homozygous regions. Excess haplotype sharing often indicates the presence of a disease locus, an observation that forms the basis of the current study.We observed that these cases shared a significantly higher number of homozygous segments bearing an accountable number of recessive genes which are related to autism.

Candidate Autism Genes

Clustering of mutated genes for autism on several chromosomes have been identified with various binding sites for regulatory proteins BX3, FOS, BACH1, MYC, JUND, MAFK, POU2F2, RBBP5, RUNX3, SMARCA4 etc. regulating various downstream genes which are known to be candidate for autism manifestation. Among the regulatory proteins CBX3, BACH1 are functioning as Repressors while FOS, BACH1, MYC, JUND, MAFK, POU2F2, RBBP5, RUNX3, SMARCA4 are functioning as Activators which are involved in the activation of various genes which perform various cellular functions such as chromatin organization, GPCR signaling, Cell Cycle regulation, homeostasis, signaling pathways, cellular stress response and so on.

Pathway analysis for the homozygous blocks

Pathway analysis of the major genes identified various pathways and processes viz., migration of Purkinje cells, morphology, formation and hypoplasia of brain, development of head and central nervous system and pervasive development disorder relevant to autism. Several gene hits such as *AUTS2, JMJD1C, PCGF5, CEP152, ABHD14A, CEACAM21, A4GALT, OPN1MW, CELSR1, RYK, ADGRA2, FZD3, FZD3, NF608,* and *RERE* were localized in various levels of these pathways.

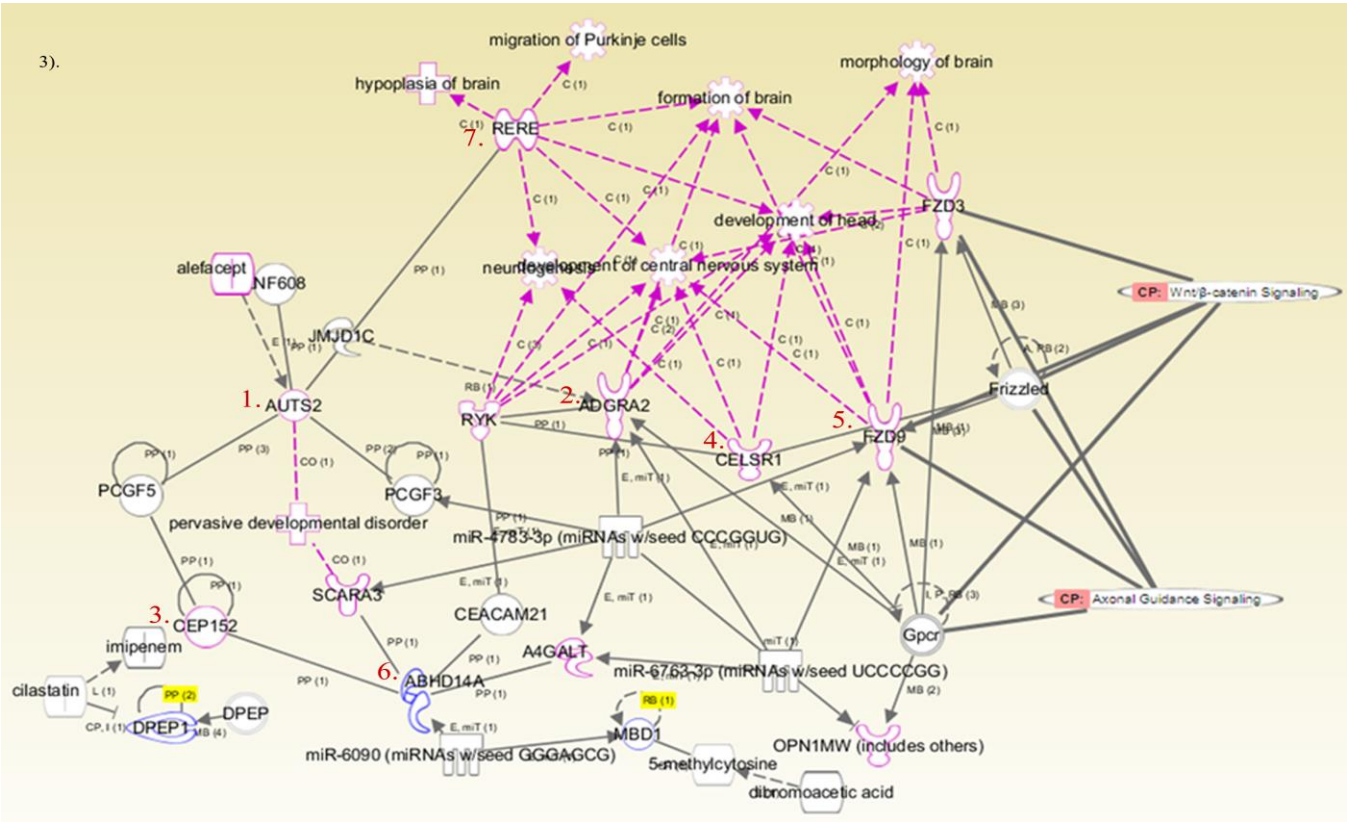


Figure 9: Pathway Analysis of enriched autism genes in risk homozygous haplotype blocks

Discussion

Susceptible genetic loci identification is crucial for better understanding the underlying mechanisms of autism, and thus aiding towards the development of their treatment and management. Owing to high heritability of autism, although various common genetic risk variants have surfaced, yet a long way to go to find rare variations showing significance of heritability in autism. Approach of homozygous haplotype mapping is complimentary to GWAS and NGS sequencing in studying the complexity and heterogeneity of autism. We report the presence of risk Homozygous Haplotypes (rHH) mapping study to identify candidate autism gene variants particularly recessive gene loci involved in autism manifestation. The idea was to apply the concept of homozygosity mapping to the trios sample cohort and understand the role of haplotype blocks in unrelated subjects.

Study group used in this study comprises of both males and females of average intellectual level, with no major emotional/behavior problems. In 61 trios, where matched parental samples were available, the confidence of the data is considerably higher than the 273 samples which were analyzed against 45 HapMap Yoruba as controls. In cases where case-controls were available, 23 haplotype blocks were found. These were located in homozygous haplotypes that were significantly more common in autism subjects than in controls. Importantly, we identified 57 previously reported autism genes. Multiple population clusters showed many significant rHH genes, suggesting the existence of common risk genes but population-specific risk alleles.

Establishment of autism risk gene loci enrichment pathway led to identification of some previously identified autism genes which show promising role in autism pathogenesis. These include which are closely linked to known autism causal genes and may be strong indicators as candidate genes viz., *AUTS2, JMJD1C, PCGF5, PCGF3, CEP152, ABHD14A, CEACAM21, A4GALT, OPN1MW, CELSR1, RYK, ADGRA2, FZD3, FZD9, NF608* and *RERE*. The genes were encompassed and involved in various pathways and processes and formed clustering at varied levels and affected downstream processes.

CONCLUSION

Identifying causal genetic factors in complex diseases are a great challenge particularly in case of neuropsychiatric condition such as Autism Spectrum Disorders (ASD). This complexity is reflected in ASD's genetic architecture where genetic players do not adhere to any traditional routes of segregation following Mendelian patterns. Instead they act independently as rare variants of large effect or in a combinatorial fashion comprising of large number of common variations of small effect. Autism is not just phenotypically heterogenous but also genotypically varied and complex. This project attempted to decipher this genotypic heterogeneity using multiple approaches such as Whole Genome genotyping using SNP Array to identify Single Nucleotide Polymorphisms (SNPs) and Copy Number Variations (CNVs), Whole Exome Sequencing using SNP micro-array and Next Generation Sequencing technology to fish out polymorphisms in coding regions as well as the Un-Translated Regions (UTRs) of the genome and additionally haplotype blocks using homozygosity mapping methods to explore the impact transmitted haplotypes in autism.

The highlights of our study are provided below:

The summary of findings can be subdivided into identifying different kinds of genetic variations like CNVs, SNPs in global subjects with and without autism as well as Indian autistic individuals; identifying regulatory roles of autism genes and miRNAs; identification of haplotype blocks through homozygosity mapping and finally identification of pathways affected by autism genes occurring either independently or within regions of CNV burden.

A. i) CNV analysis on 1746 individuals across 12 different population identified primary CNV hits such as:

- **Notable autism genes and miRNA under CNVs:** *MIR 650, MIR 1324, MIR 3180, MIR 3179, ARHGAP11B, DUSP22, CHRNA7, KCNIP1* and *KCND2.*
- **CNV Burden:** CNV burden is prominent in chromosome 6, 15 and 16 across 12 populations while chromosome 4, 14, 21 and Y have total absence of any CNVs.
- **Comparison of results found in normal cohorts with 200 whole exome sequence datasets:** Revealed 24 autism risk genes found in autism cases but absent in normal cohorts, and were identified to outline the CNVs found in normal cohorts encompassing autism genes, making them the primary hits which can manifest into autism only via a secondary hit.

ii). SNP Analysis from 1000 genome data:

- The eQTL study has identified 16 genes which are known to cause autism. Various autism genes namely, *PITX1*, *BST1*, *PRKCB* and *DLX2* are found to be involved in multiple pathways causing severity in autism cases like circadian rhythms, and learning and memory, Serine/threonine-specific protein kinase that mediates B cell activation, T cell migration, antigen-presenting cell function and cytokine release, pre-B-cell growth etc.

B. i) Deciphering of high risk genes in global whole exome sequences of autism subjects:

- Whole exome sequence datasets performed 15 autistic subjects with homogenous phenotypic traits characterized by repetitive behaviours with social and communication deficits, mild to moderate intellectual disability, delayed speech language and global developmental delay.
- Focusing on Deleterious/Damaging mutations resulted in identification of 1849 mutations that disrupted the normal gene function. Utilizing custom developed pipeline with stringent filters revealed 24 significant disease-causing variants. Thus, screening these variants led to the identification of rare mutations in *KMT2C, CNTNAP2, PTEN* and *CACNA1H.*

ii). Meta-analysis of Copy Number Variations (CNVs) in individuals with Autism Spectrum Disorder (ASDs) using Global ASD case and control datasets:

- The study identified the role of common chromosome locus in causing ASD and identified that most of the loci previously implicated in ASD are present in greater than 2% of the SFARI ASD cohort.
- Several genes causal to ASD are present in less than 0.5% of the population. Therefore, there is a need to use a more inclusive approach to perform genetic analysis of ASD and concluded that gene or locus prioritization should not be done based on frequency of occurrence.
- ASD subjects have larger CNVs than controls and genes within them are expressed in different brain regions implicated in Autism.
- Most of the autism CNVs identified were under the burden of segmental duplication that are hot spots for genomic mutations with evolutionary significance.

C. Pathways identified for autism genes within CNV regions:

- Calcium Signaling Pathway

- MAPK Signaling Pathway
- Various pathways are impaired such as, DNA helicases, voltage gated channels, methyltransferases, chromatin remodeling pathway etc involved in manifestation of autism at varied points in the pathway network.

D. Delineating the autism pathway using candidate genes identified in 15 whole exome sequence data from Indian autism subjects

- A total of 62 genes are known to be damaging and deleterious and play a key role in autism manifestation based on selection criteria.
- Chromosome 2, 3 and 15 have the highest burden of mutations which is in par with the global chromosome burden for autism.
- *PTEN, ANK2, ANKRD11* and *CNTNAP2* have various mutations either as single nucleotide and multi-nucleotide polymorphism along with Insertion-Deletion variants and copy number variants.

E. Homozygosity mapping analysis on autism cases mapped against controls

- Homozygosity mapping on 334 autism cases mapped against 167 controls revealed the presence of 38 homozygous regions based on autism candidate gene selection criteria.
- A total of 297 genes participated in various processes such as Adherens junction, Dipeptidase activity, Platelet-derived growth factor and many physiological pathways relevant to autism.
- eQTL analysis of homozygous blocks revealed presence of several polymorphisms affecting many downstream genes in brain development. Autism Gene clusters have been identified with various binding sites for regulatory proteins *BX3, FOS, BACH1, MYC, JUND, MAFK, POU2F2, RBBP5, RUNX3, SMARCA4* etc.
- Multiple population clusters showed significant risk Homozygous Haplotypes (rHH) genes, suggesting the existence of common risk genes but population-specific risk alleles.

REFERENCES

1. ADVISER: Annotation and Distributed Variant Interpretation SERver. PLoS ONE 2015;10(2): e0116815.
2. Autism Speaks. (2018). CDC increases estimate of autism's prevalence by 15 percent, to 1 in 59 children
3. Bourgeron, Thomas. "From the genetic architecture to synaptic plasticity in autism spectrum disorder." *Nature Reviews Neuroscience* 16.9 (2015): 551
4. Buxbaum, J.D. (2009). Multiple rare variants in the etiology of autism spectrum disorders. *Dialogues in clinical neuroscience*, *11*(1): 35.
5. Chakravarty S., Varadarajan R. Residue depth: a novel parameter for the analysis of protein structure and stability. Structure. 1999; 7:723–732.
6. Chang X, Wang K. wANNOVAR annotating genetic variants for personal genomes via the web. J Med Genet 2012; 49(7):433-436.
7. Chen, Jianling, et al. "Synaptic proteins and receptors defects in autism spectrum disorders." *Frontiers in cellular neuroscience* 8 (2014): 276.
8. Codina-Solà, Marta, et al. "Integrated analysis of whole-exome sequencing and transcriptome profiling in males with autism spectrum disorders." *Molecular autism* 6.1 (2015): 21
9. Cook, E.H Jr. (2001) Genetics of autism. *Child and Adolescent Psychiatric Clinics of North America*., 10(2): 333-350.
10. Cook, E.H. (1998) Genetics of autism. *Mental retardation and developmental disabilities research reviews*. 4: 113–120.
11. Cukier, Holly N., et al. "Exome sequencing of extended families with autism reveals genes shared across neurodevelopmental and neuropsychiatric disorders." *Molecular autism* 5.1 (2014): 1
12. Dang, Vinh T et al. "Identification of human haploinsufficient genes and their genomic proximity to segmental duplications." European Journal of Human Genetics 16.11 (2008): 1350.
13. De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Cicek, A. E., & Singh, T. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, *515*(7526), 209-215.
14. Dindot, S.V., Antalffy, B.A., Bhattacharjee, M.B., Beaudet, A.L. (2008) The Angelman syndrome ubiquitin ligase localizes to the synapse and nucleus, and maternal deficiency results in abnormal dendritic spine morphology. *Human Molecular Genetics* 17(1): 111-8.
15. Freitag, Christine M., et al. "Genetics of autistic disorders: review and clinical implications." *European child & adolescent psychiatry* 19.3 (2010): 169-178
16. Girirajan, S., Rosenfeld, J.A., Coe, B.P., Parikh, S., Friedman, N., Goldstein, A., Filipink, R.A., McConnell, J.S. et al. (2012) Phenotypic Heterogeneity of Genomic Disorders and Rare Copy-Number Variants. *The New England Journal of Medicine*, 367(14): 1321–1331.
17. Gross, M. (2011) Copy Number counts for autism. *Current Biology*, 21(15): 571–573.
18. Grove et al. "Common risk variants identified in autism spectrum disorder." BioRxiv (2017): 224774.
19. Haraksingh, R. R., & Snyder, M. P. (2013). Impacts of variation in the human genome on gene regulation. *Journal of Molecular Biology*, 425(21): 3970-3977.
20. Holt et al. "Linkage and candidate gene studies of autism spectrum disorders in European populations." European Journal of Human Genetics 18.9 (2010): 1013.

21. Huang, J., Lin, A., Narasimhan, B., Quertermous, T., Hsiung, C. A., Ho, L. T., & Olshen, R. A. (2004). Tree-structured supervised learning and the genetics of hypertension. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(29): 10529-10534.
22. Huang, Ni, et al. "Characterising and predicting haploinsufficiency in the human genome." PLoS genetics 6.10 (2010): e1001154.
23. Koemans, Tom S., et al. "Functional convergence of histone methyltransferases EHMT1 and KMT2C involved in intellectual disability and autism spectrum disorder." PLoS genetics 13.10 (2017): e1006864.
24. Lee, H., Meng-chin, A.L., Kornblum, H I., Papazian, D.M., & Nelson, S.F. (2014). Exome sequencing identifies de novo gain of function missense mutation in KCND2 in identical twins with autism and seizures that slows potassium channel inactivation. *Human Molecular Genetics*, ddu056.
25. Lerer, E., Levi, S., Salomon, S., Darvasi, A., Yirmiya, N., & Ebstein, R. P. (2008). Association between the oxytocin receptor (OXTR) gene and autism: relationship to Vineland Adaptive Behavior Scales and cognition. *Molecular Psychiatry*, *13*(10): 980-988.
26. Li, Jun, et al. "Schizophrenia related variants in CACNA1C also confer risk of autism." PloS one 10.7 (2015): e0133247.
27. Marshall, C.R., Scherer, S.W. (2012) Detection and characterization of copy number variation in autism spectrum disorder. *Methods in Molecular Biology*. 838:115-35.
28. McBride KL, Varga EA, Pastore MT, et al. Confirmation study of PTEN mutations among individuals with autism or developmental delays/mental retardation and macrocephaly. Autism Res. 2010;3:137
29. Nikitin A, Egorov S, Daraselia N, et al. Pathway studio—the analysis and navigation of molecular networks. Bioinformatics 2003; 19(16):2155–2157.
30. Parks, L. K., Hill, D. E., Thoma, R. J., Euler, M. J., Lewine, J. D., & Yeo, R. A. (2009). Neural correlates of communication skill and symptom severity in autism: A voxel-based morphometry study. *Research in Autism Spectrum Disorders*, *3*(2): 444-454.
31. Peñagarikano, Olga, and Daniel H. Geschwind. "What does CNTNAP2 reveal about autism spectrum disorder?." *Trends in molecular medicine* 18.3 (2012): 156-163.
32. Pham PH, Shipman WJ, Erikson GA, Schork NJ, Torkamani A. Scripps Genome
33. Pinto, Dalila, et al. "Functional impact of global rare copy number variation in autism spectrum disorders." *Nature* 466.7304 (2010): 368.
34. Pinto, Dalila, et al. "Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants." *Nature biotechnology* 29.6 (2011): 512.
35. Pinto, Dalila, et al. "Convergence of genes and cellular pathways dysregulated in autism spectrum disorders." *The American Journal of Human Genetics* 94.5 (2014): 677-694.
36. QIAGEN Ingenuity® Variant Analysis™ software (www.qiagen.com/ingenuity) QIAGEN Redwood City.
37. Russo, Anthony J. "Decreased epidermal growth factor (EGF) associated with HMGB1 and increased hyperactivity in children with autism." *Biomarker insights* 8 (2013): BMI-S11270.
38. Sampath, Srirangan, et al. "Defining the contribution of CNTNAP2 to autism susceptibility." *PLoS One* 8.10 (2013): e77906
39. Schwender, H., Ickstadt, K., & Rahnenfuehrer, J. (2008). Classification with High-Dimensional Genetic Data: Assigning Patients and Genetic Features to Known Classes. *Biometrical Journal*, *50*(6): 911-926.

40. Scoles, H.A., Urraca, N., Chadwick, S.W., Reiter, L.T., LaSalle, J.M. (2011) Increased copy number for methylated maternal 15q duplications leads to changes in gene and protein expression in human cortical samples. *Molecular Autism* 2(1): 19
41. Shishido, E., Aleksic, B., & Ozaki, N. (2014). Copy-number variation in the pathogenesis of autism spectrum disorder. *Psychiatry and clinical neurosciences*, 68(2): 85-95.
42. Sparrow, S. S., & Cicchetti, D. V. (1985). Diagnostic uses of the vineland adaptive behavior scales. *Journal of Pediatric Psychology*, *10*(2): 215-225.
43. Terwindt, G.M., Ferrari, M.D., Frants, R.R., Ophoff , R.A. (1998) Genetics and pathology of voltage-gated Ca2+ channels. *Histology & Histopathology* 13(3): 827-36.
44. Vaishnavi, Varadharajan, Mayakannan Manikandan, and Arasambattu Kannan Munirajan. "Mining the 3′ UTR of autism-implicated genes for SNPs perturbing microRNA regulation." Genomics, proteomics & bioinformatics 12.2 (2014): 92-104.
45. Veerappa AM, Murthy MN, Vishweswaraiah S, et al. Copy number variations burden on miRNA genes reveals layers of complexities involved in the regulation of pathways and phenotypic expression. PLoS One 2014; 9(2): e90391.
46. Veerappa AM, Saldanha M, Padakannaya P, et al. Family-based genome-wide copy number scan identifies five new genes of dyslexia involved in dendritic spinal plasticity. J Hum Genet 2013; 58(8):539-47.
47. Veerappa, A.M., Vishweswaraiah, S., Lingaiah, K., Murthy, M., Manjegowda, D.S., Nayaka, R., Ramachandra, N.B. (2013a) Unraveling the Complexity of Human Olfactory Receptor Repertoire by Copy Number Analysis across Population Using High Resolution Arrays. *PLoS ONE* 8(7): e66843.
48. Vernes, Sonja C., et al. "A functional genetic link between distinct developmental language disorders." New England Journal of Medicine 359.22 (2008): 2337-2345.
49. Vijayakumar, N. Thushara, and M. V. Judy. "Autism spectrum disorders: Integration of the genome, transcriptome and the environment." *Journal of the neurological sciences* 364 (2016): 167-176.
50. Vorstman, J. A. S., et al. "Identification of novel autism candidate regions through analysis of reported cytogenetic abnormalities associated with autism." *Molecular psychiatry* 11.1 (2006): 18.
51. Yasukawa, M., Bando, S., Dolken, G., Sada, E., Yakushijin, Y., Fujita, S., Makino, H. (2001) Low frequency of BCL-2/J(H) translocation in peripheral blood lymphocytes of healthy Japanese individuals. *Blood* 98:486–488