Identification of primary copy number variations reveal enrichment of Calcium, and MAPK pathways sensitizing secondary sites for autism

Swati Agarwala, Avinash M. Veerappa and Nallur B. Ramachandra

Abstract

Background: Autism is a neurodevelopmental condition with genetic heterogeneity. It is characterized by difficulties in reciprocal social interactions with strong repetitive behaviors and stereotyped interests. Copy number variations (CNVs) are genomic structural variations altering the genomic structure either by duplication or deletion. De novo or inherited CNVs are found in 5–10% of autistic subjects with a size range of few kilobases to several megabases. CNVs predispose humans to various diseases by altering gene regulation, generation of chimeric genes, and disruption of the coding region or through position effect. Although, CNVs are not the initiating event in pathogenesis; additional preceding mutations might be essential for disease manifestation. The present study is aimed to identify the primary CNVs responsible for autism susceptibility in healthy cohorts to sensitize secondary-hits. In the current investigation, primary-hit autism gene CNVs are characterized in 1715 healthy cohorts of varying ethnicities across 12 populations using Affymetrix high-resolution array study. Thirty-eight individuals from twelve families residing in Karnataka, India, with the age group of 13–73 years are included for the comparative CNV analysis. The findings are validated against global 179 autism whole-exome sequence datasets derived from Simons Simplex Collection. These datasets are deposited at the Simons Foundation Autism Research Initiative (SFARI) database.

Results: The study revealed that 34.8% of the subjects carried 2% primary-hit CNV burden with 73 singleton-autism genes in different clusters. Of these, three conserved CNV breakpoints were identified with ARHGAP11B, DUSP22, and CHRNA7 as the target genes across 12 populations. Enrichment analysis of the population-specific autism genes revealed two signaling pathways—calcium and mitogen-activated protein kinases (MAPK) in the CNV identified regions. These impaired pathways affected the downstream cascades of neuronal function and physiology, leading to autism behavior. The pathway analysis of enriched genes unravelled complex protein interaction networks, which sensitized secondary sites for autism. Further, the identification of miRNA targets associated with autism gene CNVs added severity to the condition.

Conclusion: These findings contribute to an atlas of primary-hit genes to detect autism susceptibility in healthy cohorts, indicating their impact on secondary sites for manifestation.

Keywords: Autism subgenome, CNVs, Autism-IPA, Autism pathway analysis, Primary-hit
Background
Autism is a genetic and neurodevelopmental condition with difficulties in reciprocal social interactions, abnormalities in verbal and nonverbal communication, strong repetitive behaviors, and stereotyped interests [1]. The most exclusive autism comorbidities are hypersensitivity, mood swings, impulsivity, agitation, and impairment in cognitive functions at different levels [2, 3]. The prevalence of these deficits in one or more functional domains result in autism onset, mostly before the age of three [3, 4]. Various studies have been conducted on autism starting from linkage studies, genome-wide association studies, single-nucleotide polymorphism (SNP) genotyping to present-day next-generation sequence analysis. In addition to these approaches, copy number variation (CNV) is one of the most promising studies, which adds another dimension to autism research. CNVs refer to the genomic structural variations with more than 1000 bases to many million bases in terms of size with alteration to the gene dosage. These variations can cause functional loss by disrupting regulatory elements, generating fusion proteins, or through position effect variegation. CNV occurrence can be limited to a single gene or a contiguous set of genes in a dosage-sensitive nature. Hence, the presence of these CNVs in genes can contribute to human phenotypic variability, complex behavioral traits, and disease susceptibility [5].

Various studies have addressed the impact of CNVs on autism. The first familial CNV-based study in autism identified de novo CNVs in 10% of the cases [6]. The conceptualization of CNV studies on autism has identified significant common and rare variants. These variants conferred differential effects on autism risk in the general population [7, 8]. Seventeen different loci, localized across 11 chromosomes, proposed a multigene model for CNV pathogenesis [6, 9–17]. Rare CNVs resulted in increased risk for autism by up to a 20-fold increment [17]. More than 40 recurrent autism CNVs have been identified [18]. CNV correlation has been established for multiple loci with significant autism genes namely SHANK2, SHANK3, NRXN1, NLGN4, PCDH10, DIA1, NHE926, and PARK2 [19]. Notably, 1q21.1 and 16p13.11 duplication/deletions, 15q11–q13 duplication, and 16p11.2 deletion have been the important contributors for recurrent autism CNVs [20].

CNV analysis in healthy cohorts acts as a frontier for disease susceptibility, which is evident in various research studies conducted over the last two decades. These findings highlight the role of the CNV burden in healthy groups with added contributions from other factors. It aids in the development of biomarkers for the diagnosis and prognosis of neurodevelopmental phenotypes [21, 22]. Besides, various researchers have reported a significantly higher burden of rare CNVs involving functional genes in diseases [7, 8, 17, 23]. It is hypothesized that the CNVs might not be the initiating event in the pathogenesis, and additional preceding mutations may be necessary to induce the condition [23]. A study conducted by Girirajan et al. [23] has put forth a two-hit model for disease manifestation with two promising findings. Affected individuals with a microdeletion on chromosome 16p12.1 are more likely to have additional significant CNVs than healthy individuals. The second finding of 16p12.1 (CDR2) deletion, 3q29 (DLG1) duplication, and rare copy number variants in affected individuals indicates an association of CNVs with the occurrence of severe intellectual disability and neuropsychiatric diseases due to a variable set of outcomes [23]. These outcomes are described as primary and secondary hits in the two-hit model. Susceptible gene variants present in a healthy individual, which predisposes them to disease susceptibility, are described as primary-hit. These primary-hits might/might not result in disease manifestation. For a definite progression of the disease, the occurrence of another gene variant in the individual is necessary, which are stated as secondary-hits. Combinatorial effect of a primary-hit and secondary-hit in an individual results in disease development [7]. This two-hit model could be applied to autism as well, owing to its genetic heterogeneity.

The current investigation is aimed at identifying primary autism gene CNVs in the healthy cohorts. The identification of the primary-hit CNVs in the healthy cohorts would help in uncovering the autism susceptibility loci. These primary-hit CNVs would act as molecular biomarkers for recognition of secondary-hits to minimize disease progression.

Methods
The study included a sample cohort of 1715 normal healthy individuals belonging to different ethnicities. Firstly, it included 270 HapMap samples with 30 both-parent-and-adult-child trios from Yoruba people in Ibadan, Nigeria (HapMap YRI) as well as CEPH/Utah Collection (HapMap CEU), and 45 unrelated HapMap individuals, from Tokyo Japan (HapMap) as well as Han Chinese in Beijing Japanese (HapMap CHB) populations [24]. Secondly, 155 Chinese and an equal number of 472 each, from Ashkenazi Jews replicate 1 (AJI), as well as Ashkenazi Jews, replicate 2 (AJII) populations were selected. Thirdly, 184 individuals from Taiwan, 41 from the New World population (Totonacs and Bolivians), 53 from Australia, and 31 Tibetan samples were recruited [25]. These sample datasets were obtained from the Array Express Archive of the European Bioinformatics Institute. Case registries were referred for the exclusion of subjects, wherein the samples with pre-diagnosed autism and autistic symptoms were excluded.

Thirty-eight individuals from twelve families residing in Karnataka, India, with an age group of 13–73 years
were selected for the comparative CNV analysis. Ethical approval was obtained by the Institutional Human Ethics Committee (IHEC) of the University of Mysore, Karnataka, India. Written informed consent was obtained from each subject as per the IHEC approved procedure. Informed consent for minor subjects was obtained from guardians/parents.

Five milliliter of blood was collected in K$_2$+ EDTA vacutainer tubes from the Indian study group. Genomic DNA extraction was carried out using the Promega Wizard® Genomic DNA purification kit. Visualization of isolated and quantified DNA was performed using biophotometer and gel electrophoresis.

Genome-wide genotyping was performed using the Affymetrix Genomewide Human SNP Array 6.0 chip and Affymetrix CytoScan High-Density array. The array contained 1.8 million SNP and 2.6 million CNV markers with a median inter-marker distance of 500–600 bases. These array-based studies provided the highest physical coverage and maximum panel power for the genome.

BirdSuite algorithm (https://www.broadinstitute.org/birdsuite/birdsuite-analysis) was implemented to detect commonly known copy number polymorphisms (CNPs) based on curated literature. It detected rare and common CNVs using the hidden Markov model (HMM) algorithm from Affymetrix SNP 6.0 array data. For the HMM algorithm, the hidden state mapped a specific individual to its genomic copy number. The observed states indicated the normalized intensity measurements for each array probe. This approach identified the sample-specific variable copy number regions. Collation of sample-wise CNV calls was performed from Canary and BirdSuite algorithms using the outputs from the previous step. The selection criteria were for filtering the obtained CNV calls was postulated. This criteria suggest to include BirdSuite CNV calls with a log10 of odds score ≥ 10 for an approximate false discovery rate (FDR) of ~ 5% for further analysis. For copy number (CN) states, all calls to be included except for those with CN state as 2 and differential CNP calls with CN states, in comparison to the population model.

Classification of copy number changes was performed using CNV Finder of Welcome Trust Sanger Institute with a varying quality score in the provided data. This method was based on two assumptions: firstly, the majority of data points were normalized around a log2 ratio of zero, and secondly, the data points localized outside of centralized log2 ratio distribution, indicative of a difference in the CN between reference and test genome.

CNP analysis was performed to obtain CN state calls in genomic regions using the Canary algorithm. Computation of single intensity summary statistics within the CNP region was completed manually using selected probes. An aggregative comparison of these intensity summaries has been used to assign individual CN state call across all samples, compared to those previously observed in training data.

Genotyping console selected quality control (QC) passed samples in CEL file format to call genotypes using the Birdseed algorithm. It detected CNVs with a threshold parameter of > 1 kb size and > 5 probes.

Genome-wide CNV study was carried out using Affymetrix Genotyping Console software as per standard protocol. The results were visualized using SVS Golden Helix Version 7. After employing Bonferroni correction for multiple testing, the corrected data output was used for CNV testing. For population-wise genotyped data, the threshold for the Bonferroni method was set between 1 × 10$^{-7}$ and 7 × 10$^{-8}$ for α = 0.05 on the Affymetrix 6.0 platform.

The stringency of CNP calls was met with a log10 of odds score ≥ 10 and FDR of 5%. These values corresponded to collated data output obtained from BirdSuite and Canary algorithms. All the called-SNPs with a QC call rate of > 97% were entered into the CNV analysis across subjects. Filters on call rates were used to identify call rates obtained from poor quality DNA for the overall SNPs. In the present study, contrast QC of > 2.5 with robust strength was observed across all samples. To control the possibility of spurious or artifact CNVs, the Eigenstrat approach of Price et al. was referred [26]. The principal components of the correlations among gene variants were obtained and accordingly corrected. Fifty-five individuals were extreme outliers with ≥ 1 significant Eigenstrat axes. These were excluded from the study group. Failure to meet the stipulated QC threshold resulted in the dropping of 543 CNVs in the selected individuals. Validation of CNVs was established based on ≥ 50% reciprocal overlapping with the reference set. Relative values between the comparisons of algorithms/platforms/sites were quite informative, even though the sensitivity of Jaccard statistics to the CNVs calls by each algorithm was considered. All the overlap analyses performed, handled losses and gains separately except when otherwise stated and conducted hierarchically. The algorithmic calls, called in both Canary and Birdsuite, were not considered; instead, they were collated for informative relative values between the different comparisons in terms of algorithms/platform/sites.

The reference autism gene list was prepared using two-point approaches. It was performed through an extensive well-defined PUBMED literature search matrix and SFARI database, based on inclusion-exclusion gene selection criteria. Inclusion-exclusion of gene selection was included in the criteria such as: should be an autism candidate gene expressed in brain; participate in neuronal development; interact with known autism genes; non-homozygous in controls; de novo in origin; overlap...
in two or more unrelated samples; recurrent in two or more unrelated samples; and involve in the expression of brain development and participate in neuronal migration, axon growth, neuritis outgrowth, synaptic plasticity, and cell adhesion (Fig. 1). Associated genes and genes with lower significance in terms of the \( p \) value, pathogenicity scoring, number of studies performed, and those without validations were excluded from the final gene list. The gene list was used for the overall analysis, and the CNVs were accordingly filtered. Consistently replicated genes found across populations were selected. The shared map of autism genes under CNVs was generated for all chromosomes using the Circos software package.

Function-based gene categorization for the identified CNV autism genes was performed following GO classification: Biological process, cellular component, and molecular function using WEB-based Gene Set Analysis Toolkit (WEBGASTALT). Multiple-test adjustment was applied using the hypergeometric statistical method following the Benjamini-Hochberg procedure. The significance level and \( p \) value with FDR were calculated for the top seven genes. KEGG pathways were identified using classified genes based on two criteria. It included the pathway associations and quantification of genes in each pathway with \( p \)-values and its enrichment significance. In each generated pathway map, genes from the gene list were highlighted in red.

The pathways and molecular interactions were generated through the Ingenuity Pathway Analysis (www.ingenuity.com). IPA was used to identify the interaction between genes, protein-protein interactions, biological mechanisms, location, and target gene functionality. Genes and the chemical-based search were used to explore the information on protein families, protein signaling, normal cellular protein activity, and associated metabolic pathways. Localized genes and their protein products have been interconnected through edges. An edge (line) represented the relationship between two nodes. Each network edge was described using a knowledge base of pathways and the curated literature available within IPA software. The cascades of protein-protein interaction, protein binding, activation, upregulation, downregulation, and mRNA expression by targeting a mature miRNA network were observed in pathway enrichment [27].

The validation of the four recurrent CNV breakpoints was performed by amplification using polymerase chain reaction (PCR) in our laboratory and published elsewhere [28].

**Results**

Of the total of 1715 normal healthy subjects, 34.8% of the individuals showed significant CNVs in the autism-specific subgenome. These CNVs were ranged between 8.88 and 49.05% (Fig. 2a). The highest and lowest CNV frequencies of 49% and 8% were identified in Australia and HapMap YRI respectively (Table 1). This covered 2% autism gene-specific CNV burden across the 12 populations under study. CNVs in autism genes were seen in all the chromosomes except chromosomes 4, 14, 22, and Y (Fig. 2a). CNVs present in these chromosomes were identified by 90716 SNP and CNV combined markers with an average size and count of 261.35 kb and 40.91%.
148.60 kb for autism CNV burden respectively (Supplementary Table 1, Supplementary Figure 1). Of the 2% CNV burden, duplication CNVs (73%) were predominant over the deletion CNVs (27%) (Table 1). The 2% CNV autism gene landscape contained 73 singleton autism genes [54–117] (Supplementary Table 2). The notable causal autism genes mapped for these CNV regions were ARHGAP11B, DUSP22, CHRNA7, CYFIP1, NIPA1, TUBGCP5, CACNA1H, and CELF4. The frequency of these genes was highest in ARHGAP11B, followed by DUSP22 and CHRNA7 in the CNVs regions (Fig. 2b).

Two prominent findings were comprehended for two autism causal gene clusters—CYFIP1, NIPA1, TUBGCP5 and ARHGAP11B, DUSP22, CHRNA7. These clusters were present in conserved CNV regions in multiple loci across many populations under study. The cluster containing CYFIP1, NIPA1, and TUBGCP5 was under recurrent CNV events on 15q11.2 loci. The other cluster with ARHGAP11B, DUSP22, and CHRNA7 showed significantly conserved CNV breakpoints across many populations under study. For example, DUSP22 showed one start and two end breakpoints on the chromosome 6. The start breakpoint “257341” was present in nine populations and two end breakpoints, “381131” and “382897”, were present in six populations under study. DUSP22 and ARHGAP11B genes were represented in 5.65% and 2.85% of the identified CNVs, respectively. These were marked as highly recurrent genes for autism CNV burden.

All these autism genes were under the influence of CNVs with varying CN states (Supplementary Figure 2a). CN state can be studied to estimate the expression levels of proteins. Hence, baseline brain expression level in silico analysis was performed for DUSP22 and ARHGAP11B. ARHGAP11B showed an expression level of one transcript (ENST00000602616) out of nine with a cut-off value of 0.4. Similarly, DUSP22 showed a dosage level of 13 with an expression of two transcripts (ENST00000419235 and ENST00000344450) out of 16 (Table 2). DUSP22 was under tremendous CNV burden with a frequency range of 0.09–0.16% under 70 autism genes-CNV breakpoints, both within and across populations under study. Pair-wise clustering of shared autism genes (in %) across all chromosomes and 12 populations are presented in the Circos image (Supplementary Figure 2b).

A total of 146 CNVs bearing 40 autism causal genes were limited to distinct populations with a frequency of 0.25–2.73%. There were 18 genes in 113 CNVs–AJI and AJII samples, five genes in 6 CNVs–Taiwan samples, three genes in 6 CNVs–Australia samples, three genes in 3 CNVs–Indian samples, and two genes in 2 CNVs–Tibetan samples. Similarly, CNV burden of one gene in 1 CNV–HapMap CEU, two genes in 5 CNVs–New World, four genes in 7 CNVs–China and CHB, and one gene in 3 CNVs–HapMap YRI were seen. HapMap JPT did not show any specific gene under the CNV burden. Overall, sex bias was absent for the CNV burden with CNV distribution of 51.36% in males and 48.47% in females. However, negligible male and female biases were
observed in the majority of the populations under study with regards to the percentage of CNVs present.

For further analysis, highly prevalent seven autism gene CNVs in each population ranging from 4 to 53% were chosen. Based on gene ontology (GO) study, they were classified into major categories of biological, functional processes, and location using the WEBGASTALT tool. The encoded proteins from these genes were localized primarily in the intracellular region, followed by organelle lumen and intracellular organelle lumen, while the rest were found in cell projection and cytoskeleton regions. Further, ARID1B, DBH, UBE3A, CDKL5, HLA-DRB1, and VPS13B genes were identified through enrichment analysis of autism-gene-CNVs.

The genes identified through pathway analysis were functional in neurodevelopment, neurotransmission, and synapse formation. These genes were recognized as targets of multiple miRNAs. For instance, targets of miR-499-3p were IMM2PL, UBE2G1, CACNA1B, APBA1, and DLGAP2 mRNAs. Similarly, the targets for miR-513a-5p included CACNA1B, NIPA1, DUSP22, KCND2, and CACNA1C mRNAs (Fig. 3). CNVs in autism genes CACNA1C, CACNA1H, CACNA1I, DUSP22, and CHRNA7 were enriched with calcium and MAPK signaling pathways across all populations. However, CHRNA7 and CHRM3 were enriched for neuroactive ligand-receptor interactions and present exclusively in the Chinese population (Table 3; Fig. 4).

**Discussion**

CNVs are genomic structural variations that contribute to the disease pathogenesis through gene function disruptions. Several studies on primary CNVs have been indicative of their role in the manifestation of conditions such as asthma, nondisjunction, Parkinson’s disease, diabetes, migration, and olfactory receptors [7, 29–33]. CNVs in the form of duplications and deletions manifest the gain and loss of function in a gene [34], which disrupt the protein structure and alter its transcriptional activity in the regulatory regions [35] in the autism subgenome.

**Table 1** Distribution of autism CNV duplication and deletion regions present across 12 populations

| Sl. No | Population     | Total individuals analyzed | No. of individuals with CNVs (%) | Overall autism-specific CNVs (%) | CNVs in autism genes (%) |
|-------|----------------|----------------------------|----------------------------------|---------------------------------|--------------------------|
|       |                | Male                      | Female                          | Male                           | Female                   |
|       |                | Duplication               | Deletion                        | Duplication                    | Deletion                 |
| 1     | HapMap YRI     | 90                        | 9                                | 42.86                          | 57.14                    |
| 2     | CEU            | 90                        | 9                                | 25                             | 75                       |
| 3     | Ashkenazi Jews I and II | 944             | 44                               | 20.83                          | 79.17                    |
| 4     | China and CHB  | 199                       | 15                               | 43.75                          | 56.25                    |
| 5     | Tibet          | 31                        | 48                               | 71.43                          | 28.57                    |
| 6     | India          | 38                        | 18                               | 100                            | 0                        |
| 7     | HapMap JPT     | 45                        | 36                               | 33.33                          | 66.67                    |
| 8     | Australia      | 53                        | 49                               | 82.5                           | 17.5                     |
| 9     | New World      | 41                        | 41                               | 65.21                          | 34.78                    |
| 10    | Taiwan         | 184                       | 31                               | 92                             | 8                        |
| Total |                | 1715                      | 35                               | 73.19                          | 26.81                    |

**Table 2** Baseline expression dosage of ARHGAP11B and DUSP22 in the brain using Human Gene Atlas

| Gene with baseline expression dosage | CNV type | CNV state | No. of individuals | Entire/Partial | Protein dosage |
|-------------------------------------|----------|-----------|--------------------|----------------|----------------|
| ARHGAP11B                           | Deletion | 1         | 9                  | Partial        | 0.2            |
|                                     |          | 1         | 18                 | Entire         | 0.2            |
|                                     | Duplication | 3     | 21                 |                | 0.6            |
|                                     |          | 4         | 1                  |                | 0.8            |
| DUSP22                              | Deletion | 1         | 87                 | Entire         | 6.5            |
|                                     |          | 3         | 16                 |                | 19.5           |

*CN copy number, EBI European Bioinformatics Institute
The present study establishes the autism-CNV atlas, prioritizing autism-specific CNV regions in healthy cohorts. It uncovers the primary-hit CNVs in the autism sub-genome which has been mainly unexplored. A similar trend is reported in the inherited CNVs with SHAN K2 deletion, mutations with duplication in CHRNA7, and deletions in CYFIP1 loci, which are indicative of putative multi-hit model for autism [16]. A similar trend is consistently identified in the present investigation.

Identified CNVs are present in autism-specific sub-genome with a mean average of 34%. Investigation of CNV size has been limited to ≥ 100 kb due to maximum signal to noise ratio for CNVs below 100 kb. The majority of discovered CNVs belong to a size range of 100–500 kb. The frequency of CNV events declined beyond 500 kb. Higher CNV burden is observed in autism-specific chromosomal regions 6, 15, 16, and 18, following a similar trend as mentioned in Girirajan et al. [23]. The chromosomes with autism genes are more susceptible to CNV accumulation. CNV distribution in terms of size, count, type, and state showed a different percentage for inter and intra populations, consistent with previous studies.

Further, autism gene-CNV duplications outnumber the deletion regions, as evident in AJI, AJII, Australia, and Taiwan. This can be because the genome can withstand duplications better than deletions. Loss of function is more damaging and hence it results in higher dosage and early disease manifestation. These findings are found in accordance with a previous study on autism in the European population and healthy cohorts [23, 32]. HapMap YRI and HapMap CEU contain an equal
number of duplications and deletions, suggesting that these population-specific CNVs are random events. Numerous studies advocate a similar balanced contribution from deletion and duplication in these populations [36]. Hence, studies in a larger size cohort would be needed to confirm the findings.

Autism genes with a 2% CNV burden show overlapping mutations for 73 singleton genes with previously reported autism genes. This is based on relevance, research findings in various autism cohort consortiums, and SFARI gene scoring [37, 38] (Supplementary Table 2). Out of these, 14 autism genes have been mapped to SFARI gene scoring 1. These are termed as high confidence genes with clear implications for autism. These are known to have at least three de novo gene disrupting mutations reported with a rigorous threshold FDR of < 0.1. A total of 15 genes are scored 2 and referred as strong candidate genes with two de novo gene disruptive mutations. These have been implicated by genome-wide significance or replicative in multiple studies with strong evidence. Further, 25 autism genes are scored 3 with suggestive evidence. These genes contain single de novo mutations identified from significant and non-replicated association studies. These have been reported through non-association or rare-inherited case studies with no comparative statistical study in controls. Three genes are scored syndromic with the risk of autism susceptibility. An extensive literature study shows evidence for 16 genes with no scoring and confirmed as specific to autism. Hence, these genes confirmed through curated literature have been considered in the singleton gene list (Supplementary Table 2).

The selection of seven genes for downstream analysis is based on relevance to autism and recurrent CNVs present across populations. Two prominent gene clusters are identified. The first imprint gene cluster CYFI P1, NIPA1, and TUBGCP5 is associated with changes in brain-behavior, morphology, and cognitive functions, which are key phenotypes in autism [39]. This gene cluster impacts the molecular control of synaptogenesis and neuronal connectivity in a dosage-sensitive manner [40]. Various de novo autism-specific mutations have been reported with this cluster. Further, mutations in this cluster have been marked as recurrent pathogenic CNV regions for neurodevelopmental disorders such as autism. Either side of its flanking regions contains autism genes such as UBE3A and ATP10A [41].

In the second gene cluster, ARHGAP11B and DUSP22 are under the influence of CN states 1, 3, 4, and 1, 3, respectively. The data points for CN states 1 and 3 depict a mirror image (when halved). Populations with duplications are on the higher side and deletions on the diametrically opposite lower side. CNVs with ARHGAP11B are more frequent in the Tibetan population, resulting in varied protein dosage. Higher CN states (> 2) also alter the expression level of ARHGAP11B, prominent in AJI, AJII, and Taiwan. This is in conjunction with similar CNV studies performed for asthma, nondisjunction, Parkinson’s disease, diabetes, and miRNA gene regulation in healthy cohorts [29–33]. Further, multiple mutations in ARHGAP11B include the recurrent 15q duplication and point mutations. These mutations result in early truncation and induce the proliferation of basal progenitors in the cranial neocortex [42]. ARHGAP11B, in such situations, triggers enhanced brain stem cell formation, which is a prerequisite for enlarged brain. This provides an advantage to ARHGAP11B to incur a prominent phenotype of an enlarged brain in autism [42]. DUSP22 shows recurrent breakpoints on chromosome 6 across populations. The entire protein product is affected by a protein dosage of 19.5 for deletions and duplication variants. Similarly, the DUSP22 gene results in the formation of excess neurons in the prefrontal brain in autism, which is the warehouse of social, language, and cognitive functions [43]. Thus, the presence of primary-hit CNVs for these genes increases susceptibility toward autism upon secondary-hit, either through point mutations or other gene mutations.

The population-specific CNVs are identified in varying frequencies across the sample cohort. The diversity and

| Population       | Pathways                  | Autism gene-specific CNVs |
|------------------|---------------------------|---------------------------|
| Ashkenazi Jews I | Calcium signaling         | CHRNA7, CACNA1I, CACNA1H  |
| and II          | MAPK signaling            |                           |
| Taiwan          | Calcium signaling         | DUSP22                    |
| New World       | Calcium signaling         | CACNA1I, CACNA1H          |
| Australia       | Calcium signaling         | CACNA1I, CACNA1H, CACNA1C |
| China and CHB   | Neuroactive ligand-receptor interaction | CHRNA7, CHRM3 |

Table 3 Pathway enrichment analysis of autism gene CNVs across 12 populations
Fig. 4 (See legend on next page.)
exclusivity can be either because of variable sample size or random events. Sex bias interpretation for autism CNV regions could not be conclusive due to limited information. All these CNV genes are autosomal. Hence, it is challenging to infer sex bias-based interpretations of the study. None of the established sex bias genes for autism are identified. Therefore, sex bias is ruled out and considered as balanced across populations.

GO analysis for the highly prevalent seven genes (>50%) pinpoints relevant autism gene functionality in each population. The majority of the identified autism genes are involved in the regulation of the cellular process, cellular response to organic substances, and regulation of cellular signaling. In biological processes, 90% of the genes are under the cellular process regulation, response to organic substances, and regulation of signaling. Under the molecular function category, genes encoding for cation binding and metal ion binding are significantly high. Neuronal stability and plasticity are regulated by actin and microtubule regulation present in cytoskeletal regions. These play a key role in brain functionality through neurite outgrowth and dendritic spine formation [44, 45]. Out of these enriched genes, VPS13B and ARID1B contribute to seizures and neurological speech impairments. These are causal for autism and result in its early onset. One recent study has identified an intragenic and multie exon deletion in the VPS13B gene [46]. The subsequent gene product impairs the adaptive functionality, resulting in autism on partial inactivation [46]. ARID1B is a gene with high statistical significance and an FDR value of 0.01. It has strong gene-based de novo mutational evidence for autism with absence or low mutational frequency in controls [47, 48].

Seven autism CNV genes are enriched for various pathways such as calcium signaling, MAPK signaling, and neuroactive ligand-receptor interactions. The autism-specific genes for calcium signaling pathway—CHRNA7, CACNA1H, CACNA1C, and CACNA1I—are enriched across all 12 populations. The influx of Ca²⁺ from the environment or release from internal stores causes a rapid increase in cytoplasmic calcium concentration. This dysregulated modulation of Ca²⁺ concentration results in impaired neuronal function leading to autism [49]. Products of CHRNA7 and CACNA1H are neurotransmitters and voltage channels responsible for the influx of Ca²⁺. These are involved in the regulation of the downstream cascade of reactions in the cellular pathways [50, 51]. CHRNA7 microduplication has been detected in a subject with autism and moderate cognitive impairment [46]. Mutations in these genes impair the protein product formed, which in turn affects various downstream signaling pathways. CACNA1H, DUSP22, CACNA1C, and CACNA1I are known to regulate the MAPK signaling pathway [52]. CACNA1H and CACNA1I contribute to CNV burden in most populations, while those for CACNA1C and DUSP22 are confined to a few populations [53]. These genes are expressed in four distinct MAPK groups; extracellular signal-related kinases 1/2, Jun amino-terminalse kinases 1/2/3, p38 proteins, and ERK5. These genes are involved in various cellular functions such as cell proliferation, differentiation, and migration.

The establishment of the enrichment pathway for autism-gene CNVs has identified significant genes for autism pathogenesis initiation. The minimal cut sets are computed for physical and genetic interactions. As a result, the experimental block of essential genes inevitably leads to mutants. These sets include CASKIN1, KCNIP1, and KCND2 genes. These are closely linked to known autism causal genes and might be reliable indicators as autism candidate genes.

The primary-hit autism gene CNVs identified in 1715 individuals are cross-analyzed against 179 autism whole-exome sequence datasets with identification of overlapping regions for CHRNA7 and CYFIP1. The co-occurrence of a loss of one copy of SHANK2 and CYFIP1 increases the risk of abnormal synaptic function in autistic subjects [16]. These autism genes CNVs contain 24 autism risk genes, resulting in autism manifestation, not found in a healthy cohort. Therefore, it can be inferred that the secondary-hit by the autism risk genes results in autism manifestation in autism cases, unlike the unaffected healthy cohorts, which escape contracting the condition.

Conclusion
Identification of recurrent CNVs in the healthy cohorts provides another dimension to assess the role of primary-hits toward the sensitization of the secondary-hits for the manifestation of autism. These primary-hits are vulnerable randomly, causing disease pathogenesis upon secondary-hits. Therefore, understanding susceptible loci in a healthy cohort would help in identifying...
the soft spots to avoid increasing the probability of autism manifestation.

Limitation of the study
A detailed study in larger cohorts must be warranted to identify ethnicity-specific markers. Overlapping studies could be performed on similar datasets on other platforms like next-generation sequencing, if possible, for in silico validations.

Supplementary information
Supplementary information accompanies this paper at https://doi.org/10.1186/s43042-020-00091-3.

Additional file 1: Supplementary Figure 1. Overview of a). CNV distribution in terms of count. The counts deviated significantly across population and revealed contrasting contributions towards the CNV. Chromosome 6, 15 and 16 showed higher CNV count burden, consistent with previous autism studies b). CNV distribution in terms of size. It has an average size of 2319 kbs across populations. The CNV size corresponds to the CNV count burden to chromosomes 6, 15 and 16. c). Percentage of genes under CNV burden across populations.

Additional file 2: Supplementary Figure 2. (a) Copy number states of autism genes across populations. The data points for CN states 1 and 3 depicts a mirror image (when halved), with populations showing duplications on the higher side and lower deletions on the diagnostically opposite side for all populations. (b) The pairwise clustering of shared autism genes in percent across all chromosomes for populations under study. DUSP22 is identified to be under a tremendous CNV burden across all populations. DUSP22 is present in varying frequency (0.09-0.16%) under several autism genes-CNV breakpoints, both within and across populations.

Additional file 3: Supplementary Table 1. Chromosome wise enrichment of autism gene CNV count (in percentage) across 12 populations

Additional file 4: Supplementary Table 2. Singleton genes identified under CNVs with overlapping CNV studies and functional relevance to autism.

Abbreviations
AJ: Ashkenazi Jews replicate 1; AJII: Ashkenazi Jews replicate 2; CEU: CEPH/Utah Collection HapMap; CBH: Han Chinese in Beijing; Japanese HapMap; CNV: Copy number polymorphism; CNVs: Copy number variations; FDR: False discovery rate; GO: Gene ontology; HapMap; JPT: HapMap Japan; HapMap YRI: HapMap Yoruba; HMM: Hidden Markov model; IHEC: Institutional Human Ethics Committee; IPA: Ingenuity Pathway Analysis; Kilobases: Kb; MAPK: Mitogen-activated protein kinase; PCR: Polymerase chain reaction; QC: Quality control; SFARI: Simons Foundation Autism Research Initiative; SNP: Single-nucleotide polymorphism; WEBGESTALT: WEB-based Gene Set Enrichment Analysis Toolkit

Acknowledgements
We thank the subjects and their families for participating in this study; University of Mysore for their help and encouragement; and also Department of Studies in Genetics and Genomics, University of Mysore for providing facility to conduct this work; research colleges from our laboratory, Department of Studies in Genetics and Genomics, University of Mysore for their support.

Authors’ contributions
All authors have read and approved the manuscript. SA have made substantial contributions to conception and design, and analysis and interpretation of data; involved in drafting the manuscript and reviewing the manuscript critically. AMV made substantial contributions to conception and design, acquisition of data and reviewing the manuscript critically. NBR have made substantial contributions to conception and design; revised the manuscript critically for important intellectual content and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding
Department of Science and Technology—Health Science, Government of India, New Delhi (DST/INSPIRE/160260) and University Grants Commission-Major Research Project (MRP-MAJOR-JENE-2013-19809) helped in the funding needed for the study and validations.

Availability of data and materials
Sample cohort datasets have been obtained from The International HapMap Consortium, 2003 and the Array Express Archive of the European Bioinformatics Institute.

Ethics approval and consent to participate
Written consent was obtained from all participants involved in this study and the Institutional Human Ethical Committee (IHEC No.3/9/2008-09) approved the consent procedure.

Consent for publication
Consent has been taken from all participants, the data provider, and the authors for the said publication.

Competing interests
The authors declare that they have no competing interests.

Received: 3 April 2020 Accepted: 14 August 2020
Published online: 15 December 2020

References
1. Ousley O, Cermak T (2014) Autism spectrum disorder: defining dimensions and subgroups. Curr Dev Disord Rep. https://doi.org/10.1007/s40474-013-0003-1
2. Cordeiro Q, Vallada H (2005) Genetics of autism [S]. Rev Bras Psiquiatr 27: 257. https://doi.org/10.1590/S1516-44662005000300022
3. Tye C, Runcieles AK, Whitehouse AJO, Alvareas GA (2018) Characterizing the interplay between autism spectrum disorder and comorbid medical conditions: An integrative review. Front Psychiatry. https://doi.org/10.3389/fpsyg.2018.00751
4. Cook EH Jr (2001) Genetics of autism. [Review] [128 refs]. Child Adolesc Psychiatr Clin North Am
5. Sener EF (2014) Association of copy number variations in autism spectrum disorders: a systematic review. Chinese J Biol. https://doi.org/10.1155/2014/713109
6. Sebat J, Lakshebn B, Malhotra D, Troge J, Lese-Martin C, Walsh T et al (2007) Strong association of de novo copy number mutations with autism. Science (80- ). https://doi.org/10.1126/science.1138659
7. Girirajan S, Brkanac Z, Coe BP, Baker C, Vives L, Vu TH et al (2011) Relative burden of large CNVs on a range of neurodevelopmental phenotypes. PLoS Genet. https://doi.org/10.1371/journal.pgen.1002334
8. Marshall CR, Scherer SW (2012) Detection and characterization of copy number variation in autism spectrum disorder. Methods Mol Biol. https://doi.org/10.1007/978-1-61779-507-7_5
9. Jacquemont ML, Sanlivielle D, Redon R, Raoul O, Cormier-Daire V, Lyonnet S et al (2006) Array-based comparative genomic hybridisation identifies high frequency of cryptic chromosomal rearrangements in patients with syndromic autism spectrum disorders. J Med Genet. https://doi.org/10.1136/jmg.2006.043166
10. Sztatman P, Paterson AD, Zwagenbaum L, Roberts W, Brian J, Liu XQ et al (2007) Mapping autism risk loci using genetic linkage and chromosomal rearrangements. Nat Genet. https://doi.org/10.1038/ng.1985
11. Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Staug J et al (2008) Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet. https://doi.org/10.1016/j.ajhg.2007.12.009
12. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R et al (2008) Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med. https://doi.org/10.1056/NEJMoa075974
