medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Noncoding de novo mutations in *SCN2A* are associated with autism spectrum disorders

Yuan Zhang[1], Mian Umair Ahsan[1], Kai Wang[1,2*]

1 Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

2 Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

*: Correspondence should be addressed to wangk@chop.edu

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Abstract

Coding *de novo* mutations (DNMs) contribute to the risk for autism spectrum disorders (ASD), but the contribution of noncoding DNMs remains relatively unexplored. Here we use whole genome sequencing (WGS) data of 12,411 individuals (including 3,508 probands and 2,218 unaffected siblings) from 3,357 families collected in Simons Foundation Powering Autism Research for Knowledge (SPARK) to detect DNMs associated with ASD, while examining Simons Simplex Collection (SSC) with 6383 individuals from 2274 families to replicate the results. For coding DNMs, *SCN2A* reached exome-wide significance ($p=2.06\times10^{-11}$) in SPARK. The 618 known dominant ASD genes as a group are strongly enriched for coding DNMs in cases than sibling controls (fold change=1.51, $p=1.13\text{x}10^{-5}$ for SPARK; fold change=1.86, $p=2.06\times10^{-9}$ for SSC). For noncoding DNMs, we used two methods to assess statistical significance: a point-based test that analyzes sites with a Combined Annotation Dependent Depletion (CADD) score ≥15, and a segment-based test that analyzes 1kb genomic segments with segment-specific background mutation rates (inferred from expected rare mutations in Gnocchi genome constraint scores). The point-based test identified *SCN2A* as marginally significant ($p=6.12\text{x}10^{-4}$) in SPARK, yet segment-based test identified *CSMD1*, *RBFOX1* and *CHD13* as exome-wide significant. We did not identify significant enrichment of noncoding DNMs (in all 1kb segments or those with Gnocchi>4) in the 618 known ASD genes as a group in cases than sibling controls. When combining evidence from both coding and noncoding DNMs, we found that *SCN2A* with 11 coding and 5 noncoding DNMs exhibited the strongest significance ($p=4.15\text{x}10^{-13}$). In summary, we identified both coding and noncoding DNMs in *SCN2A* associated with ASD, while nominating additional candidates for further examination in future studies.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by impaired social interaction and communication, and restrictive interests or repetitive behaviors[1-3]. ASD begins before the age of 3 years and can last throughout a person's life, though symptoms may improve over time[4]. A recently review conducted by Tomoya et al. reported that ASD affects approximately 2.3% of children aged 8 years and approximated 2.2% of adults in the US[5]. Additionally, it has a strong male bias, with boys being affected nearly 4 times more common than girls among children aged 8 years[6]. ASD is often accompanied by other psychiatric disorders, such as intellectual disability, attention-deficit hyperactivity disorder, anxiety, irritability and aggression[1]. With the increasing disease burden of ASD, early diagnosis and treatment become an urgent public health issue.

ASD exhibits extensive clinical and genetic heterogeneity with high heritability. Common variants are estimated to play an important role, accounting for ~40–80% of the overall liability for ASD[7-11]. Despite the evidence of a significant role for common variants in ASD risk, rare genetic variation (MAF<1%) confers higher individual risk[12,13]. The rare inherited variants account for a portion of the heritability[2,7], while *de novo* mutations (DNMs) identified from parent–offspring trios are the underlying cause for many cases of ASD, explain additional proportions of the overall liability[7]. Researchers have identified hundreds of high-confidence ASD genes enriched with likely deleterious protein-coding DNMs[14,15]. A large-scale exome sequencing study, using an enhanced analytical framework to integrate *de novo* and inherited rare coding variants, identified 102 putative ASD-associated genes (e.g., *CHD8*, *SCN2A*, *ADNP*)[14]. Furthermore, Zhou *et al*. performed a two-stage analysis of rare *de novo* and inherited

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

coding variants and identified 60 genes with exome-wide significance ($p < 2.5\times10^{-6}$), including five new risk genes (*NAV3*, *ITSN1*, *MARK2*, *SCAF1*, and *HNRNPUL2*)[15]. Additionally, DNMs in the non-coding genome can contribute to ASD risk[16-20]. Ryan et al. performed whole-genome sequencing of 200 ASD parent–child trios to characterize DNMs and found a significant enrichment of predicted damaging DNMs in ASD cases, of which 15.6% were non-coding[17]. Meanwhile, the authors revealed that non-coding elements most enriched for DNM were untranslated regions of genes, regulatory sequences involved in exon-skipping and DNase I hypersensitive regions. Kim *et al*. generated 813 whole-genome sequences from 242 Korean simplex families and found that target genes, including *ARHGEF2*, *BACE1*, *CDK5RAP2*, *CTNNA2*, *GRB10*, *IKZF1*, and *PDE3B*, affected by the non-coding DNMs in chromatin interactions led to early neurodevelopmental disruption implicated in ASD risk[20]. However, the pathogenic effects of noncoding DNMs related to ASD remain largely unexplored and poorly understood. Therefore, identifying noncoding DNMs that regulate gene function could provide important insights into ASD pathophysiology, which may have implications for targeted therapeutics.

Simons Foundation Powering Autism Research for Knowledge (SPARK) is an autism research initiative that aims to recruit and retain a community of 50,000 autistic individuals and their family members to advance understanding of the genetic basis of ASD[21]. As of March 2023, SPARK generated whole-genome sequencing data for 12,519 individuals from over 3000 families, offering the opportunity to assay the contribution of noncoding DNMs for ASD risk. Here, we performed whole-genome sequencing to identify both coding and noncoding DNMs in 3509 ASD trios comprised of affected probands and unaffected parents and in 2218 unaffected

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

sibling-parent trios from the SPARK cohort. Additionally, we replicated our findings in the Simons Simplex Collection (SSC) cohort with 6383 individuals from 2274 families [22].

Methods

Sample selection

The whole genome sequencing data of SPARK was made available via Simons Foundation Autism Research Initiative (SFARI) and can be requested through SFARI Base (https://www.sfari.org/resource/sfari-base/). All participants were recruited to SPARK under a centralized institutional review board (IRB) protocol (Western IRB Protocol no. 20151664). Written informed consent was obtained from all legal guardians or parents for all participants aged 18 and younger, as well as for those aged 18 and older who have a legal guardian. Assent was also obtained from dependent participants aged 10 and older. In total 3385 families were selected from the SPARK cohort. We excluded 28 families with missing data on paternal and/or maternal whole genome sequencing, leaving 3357 families (12,411 individuals) for analysis. Our study was approved by the Institutional Review Board of the Children's Hospital of Philadelphia.

The Simons Simplex Collection (SSC) cohort also represents an important data resource from the SFARI, to identify *de novo* genetic risk variants that contribute to the ASD[22]. Up to date, more than 2000 families with whole genome sequencing and clinical data have been collected. Our study included 6383 individuals (including 2274 affected probands and 1835 unaffected siblings) with whole genome sequencing from 2274 families to replicate results from the SPARK cohort. Probands were excluded who were younger than 4 years of age or older than 18. Informed consent was obtained at each data collection site included in the SSC.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Variant calling

Sequencing and genotyping of all samples for SPARK and SSC were performed at New York Genome Center (NYGC). Alignment of reads to the human reference genome version GRCh38, duplicate read marking, and Base Quality Score Recalibration (BQSR) were performed using the standard pipeline from the Centers for Common Disease Genomics (CCDG). Variants calls for all samples from the SPARK and SSC cohorts are provided from NYGC.

DeepVariant gVCFs are provided for all samples enrolled in SPARK. DeepVariant version 1.3.0 was used to call SNVs and INDELs to produce sample-level gVCFs using the default WGS configuration profile ("--model_type=WGS"). All samples were then jointly called using GLnexus version 1.4.1 and the default DeepVariant WGS configuration ("--config DeepVariantWGS"). Genomic "chunks" were processed in parallel for computational feasibility, then subsequent chunks were combined to form project VCFs (pVCFs) by chromosome. Post-calling BCFtools (version 1.17) norm was used to left-align and normalize indels.

For SSC cohort, the SNVs and indels in families were called using four different callers: GATK HaplotypeCaller v.3.5.0, FreeBayes v1.1.0, Platypus v0.8.1, and Strelka2 v2.9.2. In addition, multi-nucleotide variants were called using FreeBayes and Platypus. Post-calling BCFtools (version 1.3.1) norm was used to left-align and normalize indels. They partitioned the genome into the high-quality regions, consisting of unique space as well as ancient repeats, and the recent repeat regions, which consisted of repeats <10% diverged from the consensus in RepeatMasker. Variants were only assessed in high quality portions of the genome and those in recent repeat

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

regions were removed from the study. Our study selected project VCFs (pVCFs) generated by GATK HaplotypeCaller to replicate *de novo* analysis.

DNMs detection

We developed a custom pipeline for DNM analysis. Candidate DNMs, defined as variants present in the offspring and absent in both parents, were identified from per-family VCFs generated by DeepVariant. Bcftools and Bedtools (v2.31.0) were used to further filter the candidate DNMs. The filtering criteria for DNMs as follows: 1) Allelic depth (AD) ≥ 6 in the offspring, 2) Genotype quality (GQ) ≥ 25 in the offspring and GQ ≥ 20 in the parents, 3) Read depth (DP) > 8 in the parents, 4) Annotated the candidate DNMs using a custom pipeline based on ANNOVAR[23] (human genome hg38), and filtering the calls with an allele frequency (genomAD) < 0.1%, 5) Removed DNMs located in low mappability regions, 6) Filtering the fraction of reads supporting the alternate allele (AB.ALT) at 0.25~0.75 in the offspring, 7) removed variants located in regions known to be difficult for variant calling (e.g., HLA gene and MUC gene), and 8) when an individual carries multiple DNMs within 100 bp in the same gene, only one variant with the most severe effects was included in the analysis. Combined Annotation-Dependent Depletion (CADD, v1.6)[24] was used to score the deleteriousness of noncoding mutations. To identify potentially pathogenic variants, we extracted noncoding DNMs with CADD score thresholds of 15, which rank among the top ~1.8% of variants with the most severe predicted effect. Pipelines for analysis were blind with respect to affected probands and unaffected siblings.

Enrichment of coding DNMs

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

We further determined whether any individual genes carry an excess of DNMs. To compute the expected number of coding DNMs (including exonic and canonical splicing regions) in the cohort, we used pre-computed tabulation of the probability of DNMs arising in each gene based on RefSeq transcript definitions[25]. The number of expected coding DNMs equals to a gene's inferred DNMs mutation rate multiply by the number of trios and by 2 (for the number of chromosomes for autosomes). Given that the number of DNMs per trio follows a Poisson distribution[26], we use the Poisson test to evaluate the excesses of *de novo* events:

$$ppois(q = observed\ coding_DNMs - 1,$$

$$lambda = probability_gene * n_trios * 2,$$

$$lower.tail = FALSE)$$

Enrichment of noncoding DNMs

For noncoding DNMs, we used two methods to assess statistical significance: a point-based test that analyzes sites with a CADD score ≥15, and a segment-based test that uses genomic non-coding constraint of haploinsufficient variation (Gnocchi) constraint scores in 1kb genomic segments to infer the background mutation rates.

(1) Point-based test: to compute the expected number of noncoding DNMs, we assumed that the probability of the noncoding mutations within a given gene aligns with that of coding DNMs. Rodriguez-Galindo et al. supported that the *de novo* mutation rate is similar in exons compared to introns in the germline, after accounting for trinucleotide sequence composition and an excess of nonsynonymous exonic variation arising from sampling bias[27]. Given that gene length is an

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

obvious factor in a gene's mutability, the expected number of noncoding DNMs is calculated as follows: multiply the gene's DNM mutation rate by the number of trios, then by 2 (for autosomes), and finally by the ratio of the noncoding gene length to the coding gene length. Here, the noncoding variants are classified as those in genic (intronic, upstream, downstream, ncRNA exonic, ncRNA intronic, ncRNA splicing, UTR5, and UTR3) regions. We separately calculated statistical significance for intergenic regions by assigning noncoding variants to the nearest genes. Among the noncoding DNMs, we further classified mutations as severe if they had a CADD score greater than 15. Consequently, the noncoding gene length were calculated based on mutations with CADD score ≥15. Finally, we computed the p value for the observed number of noncoding DNMs compared to the expected number of noncoding DNMs:

$$ppois(q = observed\ noncoding_DNMs - 1,$$

$$lambda = probability_gene * n_trios * 2 * \frac{noncoding_gene_length}{coding_gene_length}$$

$$lower.tail = FALSE)$$

(2) Segment-based test: Chen et al. built a genome-wide genomic constraint map (Gnocchi) per 1-kb genomic windows by utilizing the Genome Aggregation Database (gnomAD version 3.1.2)[28]. The authors derived Gnocchi score by comparing the observed variation to an expectation from1,984,900 tiling 1kb genomic windows (passing all quality control checks) on autosomes. We annotated 1kb genomic windows and excluded variants located within coding and intergenic regions, leaving 876,086 1kb noncoding bins. We used Gnocchi genome constraint in 1-kb genomic segments to calculate the expected number of noncoding mutations in our cohorts as follows:

$$Ecpected\ DNM = \text{average_count_1kb} * \beta$$

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Where average count 1kb is the average number of DNM per 1kb, which is calculated as the average number of noncoding DNMs in a test dataset among 876,086 noncoding 1 kb genomic windows. β is inferred directly from Gnocchi calculation [28] which includes a score of the expected number of rare variants per 1kb genomic segments based on genomic contexts, and it is calculated as the ratio of the counts of the expected mutations in this 1kb bin divided by the average number of expected mutations in all 1kb bins in gnomAD.

P value is calculated as Poisson distribution using expected and observed counts in all bins assigned to each given gene. A Bonferroni corrected $p < 2.5 \times 10^{-6}$ (adjusting for ~20,000 genes) was regarded as statical significant given the theoretical number of genes tested.

To characterize the enrichment of groups of genes with DNMs contributing to ASD risk, we defined 3 gene sets, including 3054 LoF constrained genes (pLI>0.9), 1339 known NDD genes from Developmental Disorders Genotype-to-Phenotype database (DDG2P)[29], and 618 known dominant ASD genes (Supplementary Table 1). This set of 618 known ASD genes are compiled in Zhou et al[15], and they encompass known NDD genes from DDG2P, high-confidence ASD genes collected by the SFARI, and dominant ASD genes included in the SPARK genes list. Additionally, for highly constrained noncoding 1kb regions based on Gnocchi threshold (for example, Gnocchi>4), we calculated the total counts of expected and observed DNMs with score over this threshold.

Functional prediction of noncoding variants

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Genome browser (https://genome.ucsc.edu) was used to visualize and browse noncoding DNMs with a CADD score ≥15, and candidate cis-regulatory elements (cCREs) derived from ENCODE[30] integrated DNase-based sequencing (DNase-seq) and chromatin immunoprecipitation with sequencing (ChIP–seq) data. Genes with an pLI > 0.9 are defined as LoF constrained genes. We annotated potential splice site variants using SpliceAI[31], which predicts splice site gain or loss events, and we set delta scores of ≥0.8 as the high precision threshold.

Results

Cohort characteristics and study workflow

Our custom DNMs analytic pipeline was shown in **Fig 1**. A total of 3508 affected probands and 2218 unaffected sibling control from 3357 family trios were included in SPARK cohort (March 2023 release). Over 65% of the families are quartets. The mean (standard deviation) age of the SPARK cohort in this analysis was 9.0 years (5.9 years) for affected probands, 7.9 (4.5) for unaffected siblings, and 40.1 (8.4) for parents. The breakdown of sex in the full SPARK cohort was 58.2% male and 41.8% female, while among ASD cases, the breakdown was 79.7% male and 20.3% female. The characteristics of the individuals included in our analysis was shown in **Table 1**. Additionally, the SSC cohort, comprising of 6383 individuals (including 2274 affected probands and 1835 unaffected siblings) from 2274 families (over 80% were quartets), were applied to replicate the *de novo* variants analysis.

Identification of coding DNMs for ASD

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Through our custom pipeline designed for *de novo* analysis, we identified an average of 1.27 coding DNMs per affected offspring and 1.21 per unaffected siblings in the final call set (**Table 2**). In the ASD case group, 9.5% (424/4,446) of the coding DNMs were classified as likely gene disrupting (LGD), and 61.4% (2,732/4,446) were characterized as missense. While in the unaffected sibling control group, 7.4% (196/2,650) of the DNMs were classified as LGD, and 62.6% (1,660/2,650) were missense mutations. The number of coding DNMs observed per trio approximately follows Poisson distribution (**Fig. 2**). SSC was used to replicated DNM analysis. The average of coding DNMs is 1.38 for probands and 1.30 for unaffected siblings. Additionally, 9.3% of the coding DNMs were classified as LGD in the ASD group, compared to 6.1% of the coding DNMs classified as such in the unaffected sibling control group. The distribution of coding DNMs from SSC was shown in Supplementary Fig 1.

We applied *de novo* enrichment analysis to compare the observed number of DNMs to the expected numbers in ASD case trios using a one tailed Poisson test. After excluding genes that showed evidence of fold change (FC) enrichment less than one for candidate DNMs in ASD cases compared to unaffected siblings, we identify 116 genes (Supplementary Table 1) that harboring an excess of coding DNMs (p <0.05, DNM count ≥ 3). Of note, two genes with the excess of coding DNVs met Bonferroni corrected significance ($p < 2.5\times10^{-6}$), including *SCN2A* and *CCDC168*. When coding DNMs were categorized into missense and LGD variants, 33 genes were enriched for missense DNMs (e.g., *SCN2A*, *BRD4*, *KDM5B*, *CHD2*) and 11 genes were enriched for LGD DNMs (e.g., *SCN2A*, *BRSK2*, *CCDC168*, *ADNP*, *PRICKLE2*, *RAI1*, *KDM6B*, *SHANK3*, *AUTS2*, *SRCAP*, and *WDFY3*; p <0.05, FC > 1, DNM count ≥ 3). Among these, *SCN2A*, *BRSK2*, *CCDC168*, *ADNP*, and *PRICKLE2* with LGD DNMs met Bonferroni corrected

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

significance. Meanwhile, we identify 72 genes with an excess of coding DNMs (P <0.05, FC > 1, DNM count ≥ 3) from the SSC cohort, three of which (*CHD8*, *AHNAK2*, and *FLG2)* reached the Bonferroni corrected significance. Furthermore, of the 72 genes examined, 35 were identified as constrained. Nevertheless, 23 genes with missense DNMs and 3 genes with loss of function DNMs reach suggested significance (p <0.05, enrichment FC > 1, DNM count ≥ 3), of which, *CHD8* harboring loss of DNMs met Bonferroni corrected significance. Additionally, the overlapping genes that are enriched for coding DNMs between SPARK and SSC cohorts includes *SCN2A*, *ANKRD11*, *GRIN2B*, *KDM6B*, *ARID1B*, *SPEN*, *PTPRF*, and *DNMT3A*, with *PTPRF* being a previously unreported candidate gene.

Identification of noncoding DNMs for ASD

Besides coding DNMs, we also utilized our custom pipeline to identify noncoding DNMs. By comparing each affected and unaffected offspring to their parents, 302,603 noncoding DNMs (148,716 in genic regions and 153,887 in intergenic regions) were identified from 3508 affected offspring trios, while 196,898 DNMs (95,875 in genic regions and 101,023 in intergenic regions) were identified from 2218 unaffected sibling trios in the SPARK cohort. The average of noncoding DNMs is 86.3 per proband and 90.0 per unaffected sibling.

Among the noncoding DNMs, we further classified 5,343 mutations in proband trios and 3,424 in unaffected sibling trios as likely functional, based on a CADD score ≥15, which account for the top ~1.8% of variants. The distribution of noncoding DNMs is shown in **Fig 2**. The average number of noncoding DNMs had a CADD ≥ 15 is 1.52 per proband and 1.56 per unaffected sibling. We further replicated our pipeline in SSC, identifying 209,396 noncoding DNMs in 2274

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

affected offspring trios and 171,258 in 1835 unaffected sibling trios. The average of noncoding DNMs is 92.1 per proband and 93.3 per unaffected sibling. After filtering CADD score ≥ 15, we identified an average of 1.60 noncoding DNMs per ASD case and 1.59 per unaffected siblings in the final call set. The distribution of noncoding DNMs for SSC shown in Supplementary Fig 1.

We further evaluated the excesses of noncoding *de novo* events over expectation and identified 21 genes with suggestive evidence ($p < 0.05$, CADD score ≥15, FC >1, DNM count ≥3) by point-based test. For further replication, we examined the genes in the SSC cohort, and found that five out of 21 genes (*TSHZ2*, *SCN2A*, *NELL1*, *ZEB2*, and *AGMO*) had higher FC in ASD case than unaffected siblings in SSC cohort, though none reached statistical significance. Among them, *TSHZ2* and *NELL1* were not reported as candidate genes before.

Finally, by combining evidence from case-only tests for both coding and noncoding DNMs, we identified 89 genes enriched for DNMs ($p < 0.05$). Notably, *SCN2A*, with 11 coding DNMs and 5 noncoding DNMs in probands and none in unaffected siblings (**Table 3**), shows the most significant DNM burden ($p = 4.15\times10^{-13}$). In the SSC cohort, 65 genes had an excess of DNMs ($p < 0.05$). Among them, *SCN2A* had one noncoding DNM in probands and none in unaffected sibling control in SSC. Furthermore, the overlapping genes that reached threshold of 0.05 between SPARK and SSC cohorts include *SCN2A*, *DNMT3A*, *ARID1B*, *GRIN2B*, and *KDM6B*. The top genes in SPARK and SSC cohorts showed in **Table 4**.

In addition to point-based test, we also used a segment-based test to evaluate the observed number noncoding DNMs over expectation. We identified 627 genes exhibiting excesses of

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

noncoding DNMs ($p < 0.05$, DNM count ≥3; Supplementary Table 2). Within this group, four genes (*CSMD1*, *WWOX*, *RBFOX1*, and *CDH13*) achieved Bonferroni corrected significance ($p < 2.5\times10^{-6}$). We applied SSC for further replication and found that 535 genes are enriched for noncoding DNMs ($p < 0.05$, DNM count ≥3) and four of these genes (*CSMD1*, *CDH13*, *RBFOX1*, *SGCZ*) met exome-wide significance (Bonferroni $p < 2.5\times10^{-6}$). Taken together, 47 genes (e.g. *AGMO*, *GRIN2A*, *TEK*, *WWOX*, *MEMO1*) have excesses of noncoding DNMs in both SPARK and SSC cohorts. Notably, *CDH13*, *CSMD1*, and *RBFOX1* reached exome-wide statistical significance (Bonferroni $p < 2.5\times10^{-6}$) across both cohorts (Table 5).

Examination of noncoding DNMs in known ASD risk genes via case-sibling comparison

To evaluate the contribution of noncoding DNMs within established ASD risk genes, we examined 618 well-documented dominant ASD genes (Supplementary Table 3). Among these, 22 genes (e.g., *SCN2A*, *NRXN3*, and *MEIS2*) that harbored three or more noncoding DNMs had higher enrichment in probands compared to unaffected siblings (CADD score ≥15, FC >1, and DNM count ≥3), including *SCN2A*, *NRXN1*, *BCL11A*, *CASZ1*, *MEIS2*, *PBX1*, *EBF3*, *MEF2C*, *FOXP2*, *HDAC4*, *CUX2*, *SOX5*, *SATB1*, *EPB41L1*, *ZEB2*, *GLI3*, *NRXN3*, *RARB*, *MAGI2*, *CAMTA1*, *MACF1*, and *TRPM3*. Remarkably, three out of these 22 genes—*SCN2A*, *HDAC4*, and *ZEB2*—are enriched for noncoding DNMs ($p < 0.05$). For further replication, we applied the finding to the SSC. This corroborative analysis revealed that, five genes (*MEIS2*, *ZEB2*, *NRXN3*, *RARB*, and *MAGI2*) exhibited increased enrichment in SSC cohort (FC >1 and DNM count ≥3). Furthermore, among the three genes (*SCN2A*, *HDAC4*, and *ZEB2*) identified with a significant burden of noncoding DNMs in the SPARK, two genes (*SCN2A* and *ZEB2)* showed a higher burden in SSC, though not reaching statistical significance.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

When we applied a segment-based test to evaluate the excesses of noncoding DNMs for 618 known ASD risk genes, 23 genes are enriched for noncoding DNMs in SPARK and 22 genes are enriched in SSC ($p < 0.05$). Moreover, the overlapping genes that are enriched for noncoding DNMs between SPARK and SSC cohorts includes *MAGI2* and *GRIN2A* (**Table 5;** Supplementary Table 4).

Gene set analysis by case vs sibling control comparisons

We next characterized the enrichment of gene set with DNMs contributing to ASD risk in probands compared to unaffected sibling controls. Constrained genes (pLI>0.9) as a group are enriched for coding DNMs in cases than sibling controls (fold change=1.14, $p = 0.018$ for SPARK, fold change=1.44, $p = 1.21\times10^{-7}$ for SSC; **Table 6**). Moreover, the set of 618 known ASD genes had higher enrichment in both cohorts (fold change=1.51, $p = 1.13\times10^{-5}$ for SPARK, fold change=1.86, $p = 2.06\times10^{-9}$ for SSC). However, these three gene sets (LoF constrained, NDD, and known ASD genes) do not showed higher burden of noncoding DNMs in cases versus sibling controls (CADD≥15).

We further evaluate the enrichment of noncoding DNMs identified by segment-based test. We failed to find any significant enrichment of noncoding DNMs in cases over sibling controls in any of the gene sets. Additionally, when we focus on 1kb segments with Gnocchi score>4, we still do not see increased burden of noncoding DNMs in cases over controls (**Table 6**).

Function prediction of noncoding DNMs in SCN2A, ZEB2, and AGMO

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Three noncoding DNMs located in *SCN2A* (chr2:165296095, G>C; chr2:165297139, T>C; and chr2:165370303, A>T) are situated less than 20 base pairs (bp) away from an exon and are highly conserved (**Fig. 3**). The probability that the position 2:165296095 (5 bp away from an exon) being splice donor loss is 0.98 (delta score ≥ 0.8 is high precision). Likewise, the probability that the position 2:165370303 (4 bp away from exon) being splice donor loss is 0.81. In addition, the noncoding DNM found at position chr2:165294336 (T>C) of *SCN2A* is located within a candidate cis-Regulatory Element (cCRE) with a proximal enhancer like signature.

In the *ZEB2* gene, 10 noncoding DNMs were identified in probands and 3 in unaffected sibling controls (2.07-fold). Three of the 10 noncoding DNMs in probands were identified as cCREs with a distal enhancer-like signature. One noncoding DNM, located at position chr2:144404388 (T>C), showed higher enrichment of the H3K27Ac histone mark on HSMM Cells, as determined by a ChIP-seq assay. Another one, located at position chr2:144504145 (G>A), exhibited higher enrichment of the H3K4Me1 histone mark on K562 Cells. Additionally, one noncoding DNM (chr2: 144396373, G>C) is situated 39 bp away from the exon, while the probability that the position being a splice site is 0.

The *AGMO* gene harbored 4 noncoding DNMs in ASD cases. One noncoding *de novo* mutation, located at position chr7:15556988 (T>C), shows characteristic of a distal enhancer-like signature and higher enrichment of the H3K4Me1 histone mark on HUVEC cells (Supplementary Fig. 4). Additionally, mutation in chr7:15502483 (G>A) shows slightly higher enrichment of the H3K4Me1 histone mark on HUVEC cells.

Discussion

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

In the current study, we developed a custom pipeline to identify both coding and noncoding DNMs across 3357 families (12,411 individuals) with whole genome sequencing data from the SPARK cohort. We identified 116 genes that are enriched for coding DNMs in cases, of which *SCN2A* and *CCDC168* reached exome-wide significance. Furthermore, when integrating evidence from both coding and noncoding DNMs, *SCN2A* exhibited the most significant enrichment in SPARK with replication in SSC.

ASD is characterized by its clinical and etiological heterogeneity, which makes it difficult to elucidate the neurobiological mechanisms underlying its pathogenesis. Recently, DNMs have been recognized as strong source of genetic causality, and the characterization of DNMs allows additional ASD risk genes to be identified. DNMs in non-coding regions have become of interest in recent years. Previous whole exome sequencing studies were unable to detect these variants due to the lack of coverage and sequencing depth across non-coding regions. However, there is evidence that ASD genes harbor hotspots of hypermutability in non-coding regions and besides, deleterious mutations across them are subjected to strong negative selection just like the loss of function mutations located in the coding region[12]. Werling et al. present an analytical framework to evaluate rare and *de novo* noncoding mutations from whole genome sequencing of 519 ASD families unable to demonstrate a rare noncoding variant contribution to ASD risk, but found that noncoding *de novo* indels category showed a greater number of nominally significant results than expected[19]. The authors[19] demonstrated that the contribution of *de novo* noncoding variation is probably modest compared to *de novo* coding variants. Although whole genome sequencing now allows the identification of noncoding DNMs in affected probands, their contribution to risk

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

remains relatively unexplored and demands further investigation. Our study suggested that noncoding DNMs of *SCN2A* were associated with ASD risk.

The *SCN2A* gene encodes the voltage-gated NA^+ channel Nav1.2, one of the major neuronal sodium channels that play a role in the initiation and conduction of action potentials[32]. Pathogenic mutations in *SCN2A*, have been associated with a spectrum of epilepsies and neurodevelopmental disorders[33-37]. Moreover, multiple studies have confirmed the contribution of *de novo SCN2A* mutations to the risk for ASD. Stephan et al. using whole-exome sequencing of 928 individuals, including 200 phenotypically discordant sibling pairs confirmed that *de novo* coding mutations across *SCN2A* were associated with ASD risk[37] . Our discovery of coding DNMs events in *SCN2A* related to the risk of ASD are supporting previous studies. Intriguingly, three noncoding DNMs events observed in *SCN2A* are located within 20 base pairs of exons, further implicating them in ASD risk. Notably, two of these noncoding DNMs demonstrated a high likelihood of affecting splice donor/loss sites, potentially leading to significant alterations in gene expression. Also, one noncoding DNM in *SCN2A* displays characteristic of a proximal enhancer like signature. This suggests a critical role for both coding and proximal noncoding regions of *SCN2A* in ASD pathogenesis, highlighting the intricate interplay between genetic variants and their potential impact on splicing and gene function.

CHD8 encodes chromodomain helicase DNA-binding protein 8 and its mutation is a highly penetrant risk factor for ASD[14,38]. While DNMs of *CHD8* are associated with ASD risk in the SSC cohort, but not in the SPARK cohort. One possible explanation may be the different sample characteristics and family fractures. In SSC cohort, over 80% family types are quartet families

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

and around 20% are simplex. While in SPARK cohort, 65% family types are quartet families and 33% are simplex families. Additionally, the stochastic nature of rare mutations may also explain the general lack of DNMs in *CHD8* in the SPARK cohort.

To our best knowledge, this study was among the first to evaluate the significance of non-coding DNMs implicated in the risk of ASD from whole genome sequencing data across more than 3300 families. Our findings also suggested the contribution of noncoding DNMs in known ASD risk genes, especially *SCN2A*. This study also has several limitations. The sample size from SPARK is relatively small, and the size of the ASD cases was around double that of the unaffected sibling controls, which may hinder the discovery of ASD risk genes. Furthermore, the probability of *de novo* mutations per gene may differ between coding and noncoding regions. To compute the expected number of noncoding DNMs, we assume that the mutation rate of the noncoding DNMs occurring within the same gene is comparable to that of coding DNMs[39,40]. Rodriguez-Galindo et al. demonstrated that the rate of generation of new genetic variants, the mutation rate, does not significantly vary between exons and adjacent introns when accounting for sequence context[27]. However, Sankar et al. revealed mutation rates in exons are 30%–60% higher than in noncoding DNA due to the relative overabundance of synonymous sites involved in CpG dinucleotides[41]. Studies also reported that mutation rate in non-coding regions is highly heterogeneous and can be affected by local sequence context as commonly modelled in gene constraint metrics as well as by a variety of genomic features at larger scales[42,43]. Therefore, we used two different methods (point-based and segment-based) to calculate the noncoding DNMs because of the probability of DNMs in noncoding regions still limited. In the near future, repeating these analyses with large-scale whole genome sequencing data, potentially utilizing

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

long-read sequencing technologies, and investigating inherited variants will aid in identifying additional ASD risk genes.

Data availability

Whole genome sequencing data was made available via SFARI and can be requested through SFARI Base (https://www.sfari.org/resource/sfari-base/).

The SPARK data is accessible as follow:

SPARK_iWGS_v1.1 includes whole genome data from 12519 individuals from 3417 families.

The SSC data is accessible as follow:

SFARI_SSC_WGS_2a includes whole genome data from 6383 individuals from 2274 families.

Acknowledgements

We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, E. Wijsman). We are grateful to all of the families in SPARK, the SPARK clinical sites and SPARK staff. We appreciate obtaining access to genetic data on SFARI Base. Approved researchers can obtain the SSC population dataset described in this study by applying at https://base.sfari.org. Approved researchers can obtain the SPARK population dataset described in this study by applying at https://base.sfari.org. We appreciate obtaining access to recruit participants through SPARK research match on SFARI Base. We thank Dr. Yufeng Shen (Columbia University) for his

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

insightful comments on the analysis strategies. The study is supported in part by a Foerderer grant, NIH grant HG013031 and the CHOP Research Institute.

Competing Interests

The authors declare no competing interests.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Reference

1. Lord C, Elsabbagh M, Baird G, Veenstra-Vanderweele J. Autism spectrum disorder. *Lancet.* 2018;392(10146):508-520.
2. Wang L, Wang B, Wu C, Wang J, Sun M. Autism Spectrum Disorder: Neurodevelopmental Risk Factors, Biological Mechanism, and Precision Therapy. *Int J Mol Sci.* 2023;24(3).
3. Alpert JS. Autism: A Spectrum Disorder. *Am J Med.* 2021;134(6):701-702.
4. Centers for Disease Control and Prevention. 2024(https://www.cdc.gov/ncbddd/autism).
5. Hirota T, King BH. Autism Spectrum Disorder: A Review. *JAMA.* 2023;329(2):157-168.
6. Maenner MJ, Shaw KA, Bakian AV, et al. Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2018. *MMWR Surveill Summ.* 2021;70(11):1-16.
7. Gaugler T, Klei L, Sanders SJ, et al. Most genetic risk for autism resides with common variation. *Nat Genet.* 2014;46(8):881-885.
8. Havdahl A, Niarchou M, Starnawska A, Uddin M, van der Merwe C, Warrier V. Genetic contributions to autism spectrum disorder. *Psychol Med.* 2021;51(13):2260-2273.
9. Sandin S, Lichtenstein P, Kuja-Halkola R, Hultman C, Larsson H, Reichenberg A. The Heritability of Autism Spectrum Disorder. *JAMA.* 2017;318(12):1182-1184.
10. Bai D, Yip BHK, Windham GC, et al. Association of Genetic and Environmental Factors With Autism in a 5-Country Cohort. *JAMA Psychiatry.* 2019;76(10):1035-1043.
11. Sandin S, Lichtenstein P, Kuja-Halkola R, Larsson H, Hultman CM, Reichenberg A. The familial risk of autism. *JAMA.* 2014;311(17):1770-1777.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

12. Michaelson JJ, Shi Y, Gujral M, et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell.* 2012;151(7):1431-1442.

13. He X, Sanders SJ, Liu L, et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet.* 2013;9(8):e1003671.

14. Satterstrom FK, Kosmicki JA, Wang J, et al. Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell.* 2020;180(3):568-584 e523.

15. Zhou X, Feliciano P, Shu C, et al. Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. *Nat Genet.* 2022;54(9):1305-1319.

16. Hodge JC, Mitchell E, Pillalamarri V, et al. Disruption of MBD5 contributes to a spectrum of psychopathology and neurodevelopmental abnormalities. *Mol Psychiatry.* 2014;19(3):368-379.

17. Yuen RK, Merico D, Cao H, et al. Genome-wide characteristics of de novo mutations in autism. *NPJ Genom Med.* 2016;1:160271-1602710.

18. Zhou J, Park CY, Theesfeld CL, et al. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat Genet.* 2019;51(6):973-980.

19. Werling DM, Brand H, An JY, et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat Genet.* 2018;50(5):727-736.

20. Kim IB, Lee T, Lee J, et al. Non-coding de novo mutations in chromatin interactions are implicated in autism spectrum disorder. *Mol Psychiatry.* 2022;27(11):4680-4694.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

21. SPARK Consortium. SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research. *Neuron.* 2018;97(3):488-493.

22. Fischbach GD, Lord C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron.* 2010;68(2):192-195.

23. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38(16):e164.

24. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 2019;47(D1):D886-D894.

25. Samocha KE, Robinson EB, Sanders SJ, et al. A framework for the interpretation of de novo mutation in human disease. *Nat Genet.* 2014;46(9):944-950.

26. Neale BM, Kou Y, Liu L, et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature.* 2012;485(7397):242-245.

27. Rodriguez-Galindo M, Casillas S, Weghorn D, Barbadilla A. Germline de novo mutation rates on exons versus introns in humans. *Nat Commun.* 2020;11(1):3304.

28. Chen S, Francioli LC, Goodrich JK, et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature.* 2024;625(7993):92-100.

29. Wright CF, Fitzgerald TW, Jones WD, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet.* 2015;385(9975):1305-1314.

30. Consortium EP, Moore JE, Purcaro MJ, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature.* 2020;583(7818):699-710.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

31. de Sainte Agathe JM, Filser M, Isidor B, et al. SpliceAI-visual: a free online tool to improve SpliceAI splicing variant interpretation. *Hum Genomics.* 2023;17(1):7.

32. Liao Y, Deprez L, Maljevic S, et al. Molecular correlates of age-dependent seizures in an inherited neonatal-infantile epilepsy. *Brain.* 2010;133(Pt 5):1403-1414.

33. Kobayashi K, Ohzono H, Shinohara M, et al. Acute encephalopathy with a novel point mutation in the SCN2A gene. *Epilepsy Res.* 2012;102(1-2):109-112.

34. Horvath GA, Demos M, Shyr C, et al. Secondary neurotransmitter deficiencies in epilepsy caused by voltage-gated sodium channelopathies: A potential treatment target? *Mol Genet Metab.* 2016;117(1):42-48.

35. Schwarz N, Hahn A, Bast T, et al. Mutations in the sodium channel gene SCN2A cause neonatal epilepsy with late-onset episodic ataxia. *J Neurol.* 2016;263(2):334-343.

36. Li J, Cai T, Jiang Y, et al. Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. *Mol Psychiatry.* 2016;21(2):290-297.

37. Sanders SJ, Murtha MT, Gupta AR, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature.* 2012;485(7397):237-241.

38. Bernier R, Golzio C, Xiong B, et al. Disruptive CHD8 mutations define a subtype of autism early in development. *Cell.* 2014;158(2):263-276.

39. He B, Gao P, Ding YY, et al. Diverse noncoding mutations contribute to deregulation of cis-regulatory landscape in pediatric cancers. *Sci Adv.* 2020;6(30):eaba3064.

40. Keightley PD, Gaffney DJ. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc Natl Acad Sci U S A.* 2003;100(23):13402-13406.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

41. Subramanian S, Kumar S. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome Res.* 2003;13(5):838-844.

42. Seplyarskiy VB, Sunyaev S. The origin of human mutation in light of genomic data. *Nat Rev Genet.* 2021;22(10):672-686.

43. Seplyarskiy VB, Soldatov RA, Koch E, et al. Population sequencing data reveal a compendium of mutational processes in the human germ line. *Science.* 2021;373(6558):1030-1035.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Tables

Table 1. Characteristics of individuals included in the SPARK cohort.

Characteristic	Autism case (n=3508)	Sibling control (n=2188)	Parents (n=6714)
Age at enrollment in years, mean (SD)	9.0 (5.9)	7.9 (4.5)	40.1 (8.4)
Sex at birth, n (%)			
Male	2796 (79.7)	1069 (48.9)	3358 (50.0)
Female	712 (20.3)	1119 (51.1)	3356 (50.0)
Genetically inferred race/ethnicity, n (%)			
White	2648 (75.48)	1656 (75.69)	5121 (76.28)
African American	168 (4.79)	96 (4.39)	325 (4.84)
Asian	195 (5.56)	123 (5.62)	384 (5.72)
Hispanic	382 (10.89)	240 (10.97)	740 (11.02)
Other	115 (3.28)	73 (3.34)	143 (2.13)

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Table 2. Summary of *de novo* mutations (DNMs) identified from whole-genome sequencing

	SPARK cohort		SSC cohort	
Categories	ASD cases (n=3508)	Unaffected sibling controls (n=2188)	ASD cases (n=2274)	Unaffected Sibling controls (n=1835)
Coding DNMs	4446	2650	3143	2390
Likely gene disrupting [a]	424	196	292	146
Missense	2732	1660	1925	1501
Coding DNMs per trio	1.27	1.21	1.38	1.30
Non-coding DNMs	302,603	196,898	209,396	171,258
With CADD ≥15	5343	3424	3645	2920
Genic	148,716	95,875	99,960	80,981
With CADD ≥15	3090	2010	2119	1703
Intergenic	153,887	101,023	109,436	90,277
With CADD ≥15	2253	1414	1526	1217
Noncoding DNMs per trio	86.3	90.0	92.1	93.2

[a]Loss-of-function mutations including frameshift deletion, frameshift insertion, startloss, stopgain, stoploss and splicing.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Table 3. A list of DNMs in *SCN2A* from whole-genome sequencing of the SPARK cohort

Subject ID	Pos (hg38)	Ref.	Alt.	Function	Exonic Function	Base change	CADD
SP0302927	2:165314008	A	G	Exonic	Nonsynonymous	c.A1283G	28.9
SP0165280	2:165326850	A	C	Splicing	-	c.2017-2A>C	33
SP0242051	2:165354671	G	C	Exonic	Nonsynonymous	c.G3399C	33
SP0028263	2:165373267	G	T	Exonic	Stopgain	c.G3892T	47
SP0251020	2:165374728	ATGTACTTCTGGTTTGTCTGATC	A	Exonic	Frameshift deletion	c.4017_4038del	34
SP0256108	2:165374956	T	C	Exonic	Nonsynonymous	c.T4244C	29.9
SP0273995	2:165377651	G	A	Splicing	-	c.4308+1G>A	34
SP0333694	2:165386818	T	C	Exonic	Nonsynonymous	c.T4624C	29
SP0112003	2:165386919	TG	T	Exonic	Frameshift deletion	c.4726delG	34
SP0198912	2:165388935	T	A	Exonic	Nonsynonymous	c.T5129A	28
SP0184979	2:165389304	A	AACCC	Exonic	Frameshift insertion	c.5498_5499insACCC	33
SP0157054	2:165275094	T	G	Intronic	-	-	16.05
SP0137285	2:165294336	T	C	Intronic	-	-	16.6
SP0191905	2:165296095	G	C	Intronic	-	c.267+5G>C	25.5
SP0225518	2:165297139	T	C	Intronic	-	c.386+4T>C	20.7
SP0103495	2:165370303	A	T	Intronic	-	c.3849+4A>T	24.8

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Table 4. Top genes with DNMs in the SPARK and SSC cohortsGenes

	Count of coding DNVs in case	Count of coding DNVs in control	Count of noncoding DNVs with CADD ≥15 in case	Count of noncoding DNVs with CADD ≥15 in control	Case-only p value for coding DNMs	Case-only p value for noncoding DNMs	Combined p values for both coding and noncoding DNMs	SFARI gene score
Top 10 genes in SPARK cohort								
SCN2A	11	0	5	0	2.06E-11	6.12E-04	4.15E-13	1
CCDC168	4	0	0	0	3.49E-07	1	5.53E-06	-
PIEZO1	3	0	1	0	2.13E-04	0.02	4.56E-05	-
PTEN	4	0	0	0	4.34E-06	1	5.79E-05	1
C16orf96	3	0	0	0	7.30E-06	1	9.37E-05	-
BRD4	6	0	0	0	9.95E-06	1	1.25E-04	2
ADNP	5	0	0	0	1.42E-05	1	1.72E-04	1
DDX3X	3	0	2	0	1.39E-03	0.03	4.70E-04	1
KDM6B	5	0	2	0	4.91E-04	0.11	5.60E-04	1
SOX1	4	1	0	0	7.46E-05	1	7.84E-04	-
Top 10 genes in SSC cohort								
FLG2	8	5	0	0	7.05E-09	1	1.39E-07	-
CHD8	7	0	0	0	4.02E-07	1	6.32E-06	1
AHNAK2	10	5	0	0	1.11E-06	1	1.63E-05	-
SYNGAP1	5	1	1	2	1.94E-05	0.34	8.48E-05	1
AMZ1	4	1	0	0	9.49E-06	1	1.19E-04	-
SCN2A	5	0	1	0	3.53E-05	0.35	1.52E-04	1
PLIN4	5	4	0	0	2.45E-05	1	2.85E-04	-
WDFY4	3	0	0	0	3.15E-05	1	3.58E-04	2
CDC42BPB	3	1	2	0	8.06E-03	0.01	5.49E-04	2
ALMS1	5	1	1	1	6.38E-04	0.11	7.65E-04	-
Suggestive significant (p<0.05) in both cohorts (SPARK/SSC)								
SCN2A	11/5	0/0	5/1	0/0	2.60E-14	2.02E-03	2.41E-15	1
KDM6B	5/3	0/0	2/1	0/1	5.52E-05	0.14	1.23E-04	1
GRIN2B	5/3	0/0	2/0	0/0	1.91E-05	0.97	6.47E-04	1
DNMT3A	3/3	0/0	0/2	0/0	1.19E-04	0.55	1.14E-03	1
ARID1B	4/4	0/0	1/1	1/2	1.16E-04	0.93	2.66E-03	1

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Table 5. Top genes with noncoding DNMs in the SPARK cohort, with validation in the SSC cohort, using point-based and segment-based statistical tests.

	SPARK cohort			SSC cohort			
Genes	Count of noncoding DNMs in cases	Expected noncoding DNMs in cases	Case-only p value for noncoding DNMs	Count of noncoding DNMs in cases	Expected noncoding DNMs in cases	Case-only p value for noncoding DNMs	SFARI gene score
Point-based test: noncoding mutations with a CADD score ≥15							
Known ASD risk genes (*Zhou et al.*)							
SCN2A	5	0.66	6.12E-04	1	0.43	0.35	1
ZEB2	10	4.89	2.81E-02	4	3.17	0.39	-
AGMO	4	1.17	3.09E-02	1	0.76	0.53	2
Newly found candidate ASD risk genes							
TSHZ2	9	2.96	3.51E-03	4	1.92	0.13	-
NELL1	7	2.10	5.88E-03	1	1.36	0.74	-
Segment-based test: Gnocchi genome constraint in 1kb genomic segments							
Known ASD risk genes (*Zhou et al.*)							
MAGI2	126	100.86	0.0087	86	64.41	0.0059	-
GRIN2A	48	33.50	0.0106	30	21.39	0.045	1
Newly found candidate ASD risk genes							
CSMD1	142	69.10	1.12E-14	104	44.13	2.98E-19	2
RBFOX1	208	120.15	2.44E-13	124	76.74	4.38E-07	2
CHD13	146	94.35	5.07E-07	110	60.26	5.86E-09	2

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Table 6. Enrichment of DNMs in ASD cases across gene sets by case/control comparisons.

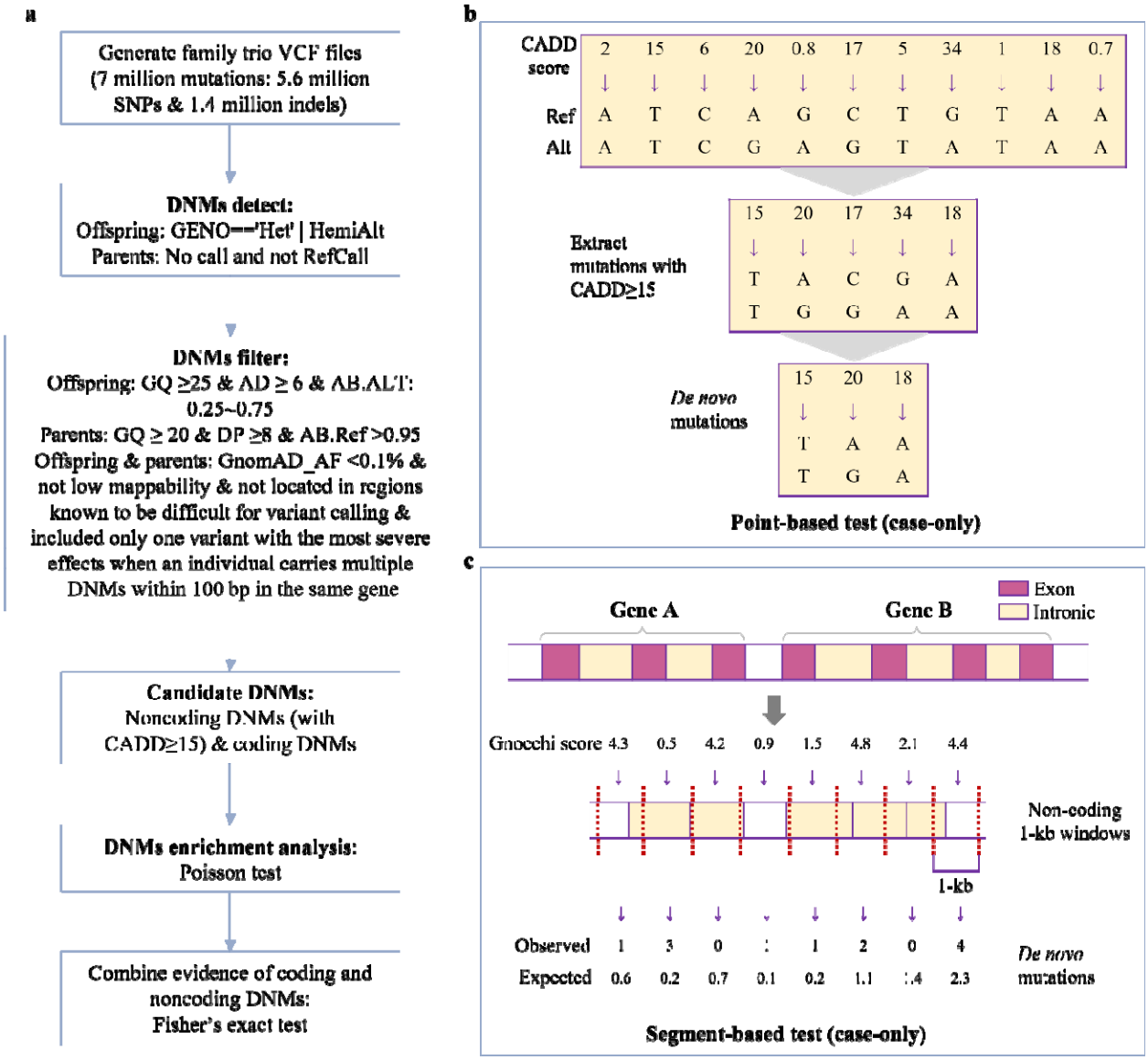
Gene sets were grouped as 3054 LoF constrained gene (pLI >0.9), 1339 neurodevelopmental disorders (NDD) genes and 618 known ASD genes.

		SPARK cohort				**SSC cohort**	
Gene sets	Case/control (# subjects)	Fold change	P value		Case/control (# subjects)	Fold change	P value
Genes with coding DNMs							
All genes	2484/1528	1.05	0.225		1717/1330	1.17	0.015
LoF Constrained genes	1000/567	1.14	0.018		726/451	1.44	1.21×10^{-7}
NDD genes	320/201	1.00	0.552		245/166	1.21	0.037
Known ASD genes	378/162	1.51	1.13×10^{-5}		303/140	1.86	2.06×10^{-9}
Genes with noncoding DNMs (point-based test with CADD>15)							
All genes	2014/1297	0.93	0.922		1358/1093	1.01	0.472
Constrained genes	969/623	0.96	0.766		633/546	0.91	0.917
NDD genes	199/125	0.99	0.550		131/125	0.84	0.926
Known ASD genes	293/166	1.11	0.163		177/164	0.86	0.917
Genes with noncoding DNMs (segment-based test)							
All genes	3508/2188	1	1		2274/1835	1	1
Constrained genes (231,967 1kb windows)	3501/2183	1.15	0.516		2271/1828	2.90	0.098
NDD genes (48,339 1kb windows)	2622/1663	0.93	0.865		1655/1387	0.87	0.957
ASD known risk genes (47,389 1kb windows)	2662/1663	0.99	0.553		1738/1391	1.04	0.333
Genes with noncoding DNMs (segment-based test with Gnocchi score ≥4)							
All genes (12,792 1kb windows)	1195/747	1.00	0.535		762/612	1.01	0.470
Constrained genes (2,993 1kb windows)	351/231	0.94	0.763		233/180	1.05	0.341
NDD genes (935 1kb windows)	113/67	1.05	0.401		64/67	0.76	0.945
ASD known risk genes (767 1kb windows)	92/54	1.06	0.394		65/54	0.97	0.601

medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

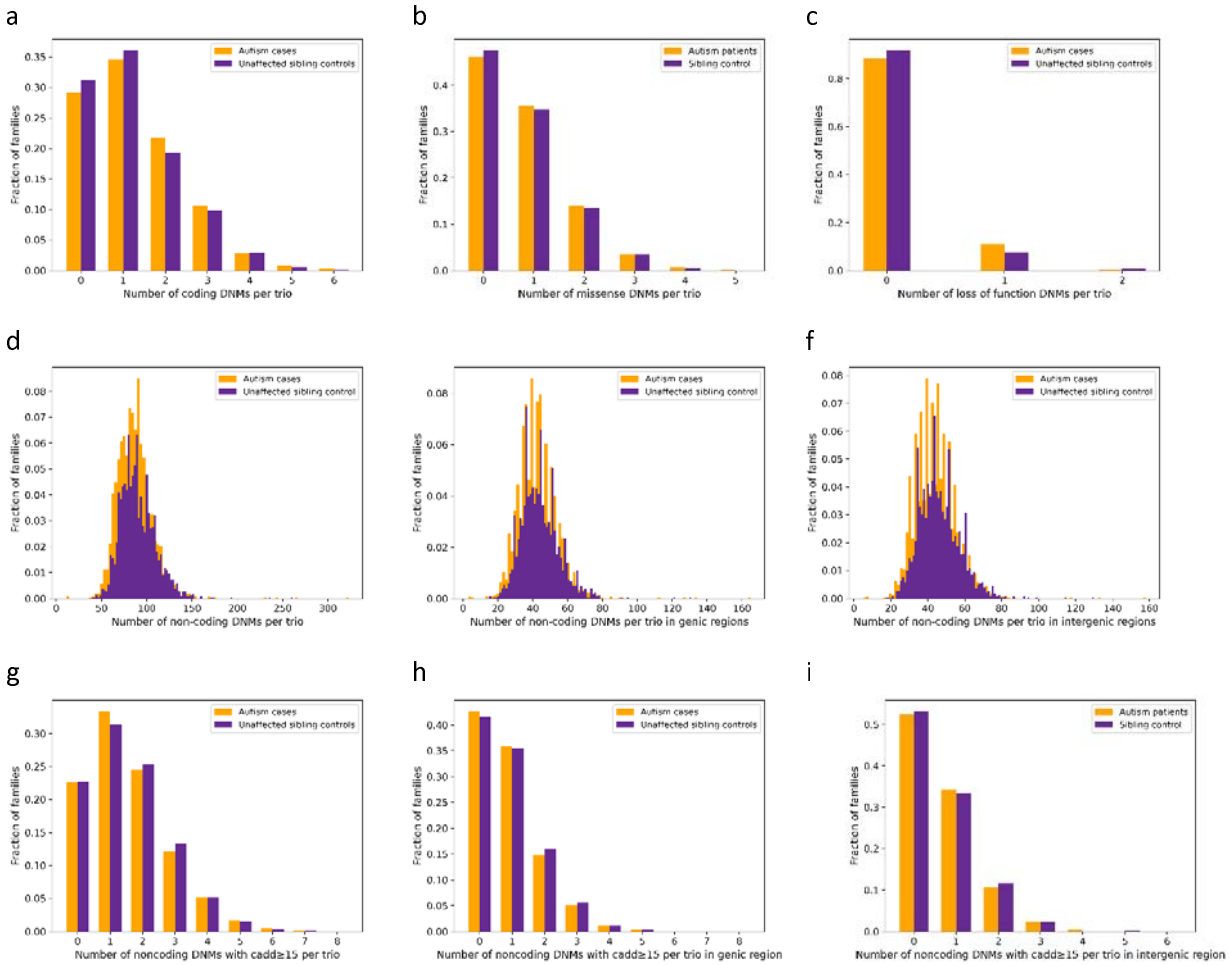
Figures

Fig 1. Analysis workflow. (a) DNMs analytic pipeline. For noncoding DNMs, this study used two methods to assess statistical significance **(b)** point-based test that analyzes sites with a Combined Annotation Dependent Depletion (CADD) score ≥15, and **(c)** segment-based test that uses Gnocchi genome constraint scores in 1kb genomic segments to infer background mutation rates.



medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Fig 2. Distribution of DNMs in SPARK. (a) distribution of coding DNMs. **(b)** distribution of missense DNMs. **(c)** distribution of loss of function DNMs. **(d)** distribution of noncoding DNMs. **(e)** distribution of gene-relate noncoding DNMs. **(f)** distribution of intergenic noncoding DNMs. **(g)** distribution of noncoding DNMs have a CADD score ≥15. **(h)** distribution of gene-relate noncoding DNMs have a CADD score ≥15. **(i)** distribution of intergenic noncoding DNMs have a CADD score ≥15.



medRxiv preprint doi: https://doi.org/10.1101/2024.05.05.24306908; this version posted May 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

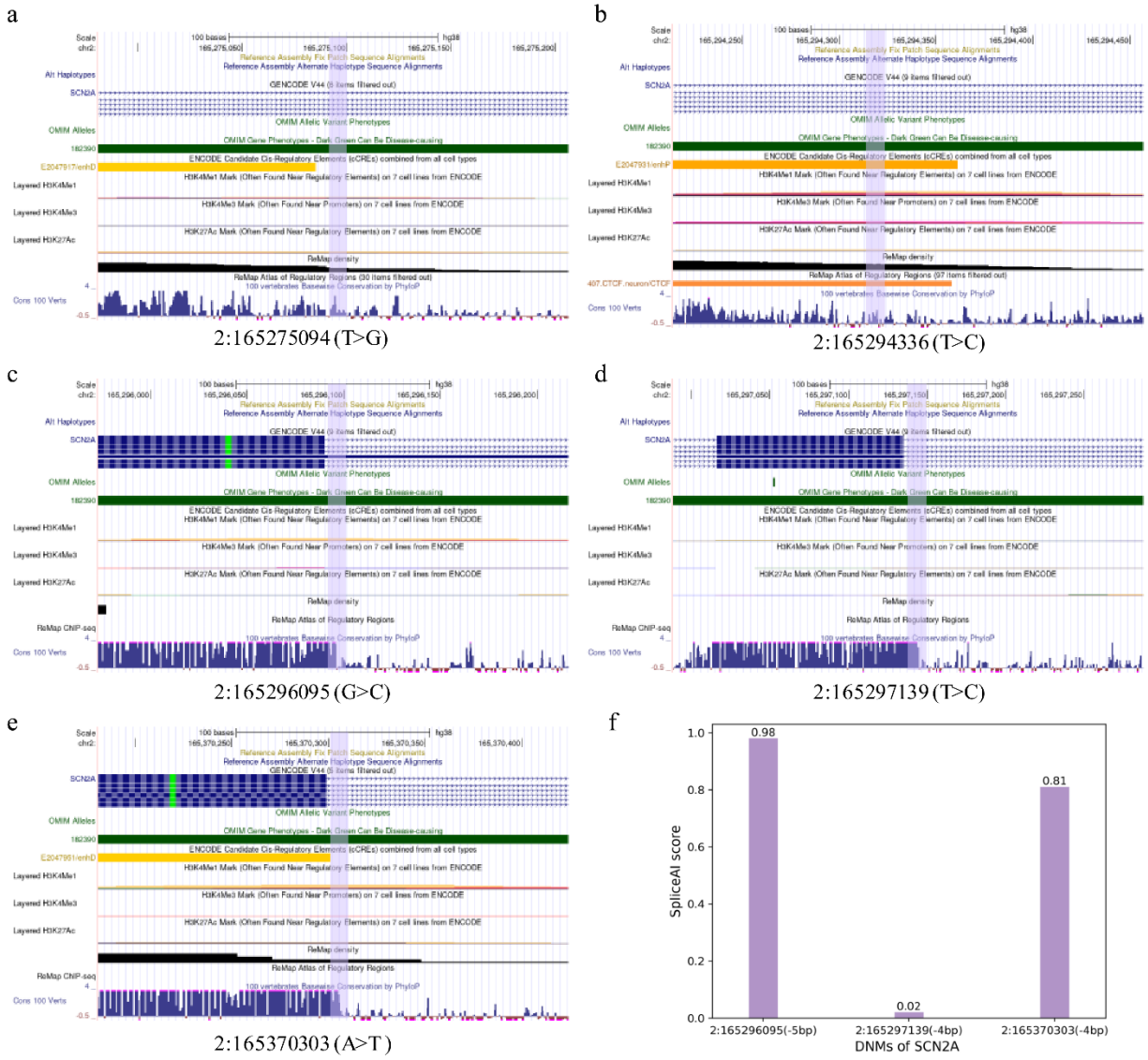



Fig 3. **Point variation of the noncoding DNMs for the *SCN2A* gene. (a)** point variation of chr2:165275094 (T>G). **(b)** point variation of chr2:165294336 (T>C). **(c)** point variation of chr2:165296095 (G>C). **(d)** point variation of chr2:165297139 (T>C). **(e)** point variation of chr2:165370303 (A>T). **(f)** probability of the mutations situated less than 20 base pairs (bp) distant from the exon being splice-altering.