Mapping structural variants to rare disease genes using long-read whole genome sequencing and trait-relevant polygenic scores

Cas LeMaster[1], Carl Schwendinger-Schreck[1], Bing Ge[2], Warren A. Cheung[1], Rebecca McLennan[1], Jeffrey J. Johnston[1], Tomi Pastinen[1], Craig Smail[1]

[1] Genomic Medicine Center, Children's Mercy Research Institute and Children's Mercy Kansas City, Kansas City, MO, USA

[2] McGill University, Montreal, Quebec, Canada

ABSTRACT

Recent studies have revealed the pervasive landscape of rare structural variants (rSVs) present in human genomes. rSVs can have extreme effects on the expression of proximal genes and, in a rare disease context, have been implicated in patient cases where no diagnostic single nucleotide variant (SNV) was found. Approaches for integrating rSVs to date have focused on targeted approaches in known Mendelian rare disease genes. This approach is intractable for rare diseases with many causal loci or patients with complex, multi-phenotype syndromes. We hypothesized that integrating trait-relevant polygenic scores (PGS) would provide a substantial reduction in the number of candidate disease genes in which to assess rSV effects. We further implemented a method for ranking PGS genes to define a set of core/key genes where a rSV has the potential to exert relatively larger effects on disease risk. Among a subset of patients enrolled in the Genomic Answers for Kids (GA4K) rare disease program (N=497), we used PacBio HiFi long-read whole genome sequencing (lrWGS) to identify rSVs intersecting genes in trait-relevant PGSs. Illustrating our approach in Autism (N=54 cases), we identified 22,019 deletions, 2,041 duplications, 87,826 insertions, and 214 inversions overlapping putative core/key PGS genes. Additionally, by integrating genomic constraint annotations from gnomAD, we observed that rare

duplications overlapping putative core/key PGS genes were frequently in higher constraint regions compared to controls ($P = 1x10^{-03}$). This difference was not observed in the lowest-ranked gene set ($P = 0.15$). Overall, our study provides a framework for the annotation of long-read rSVs from lrWGS data and prioritization of disease-linked genomic regions for downstream functional validation of rSV impacts. To enable reuse by other researchers, we have made SV allele frequencies and gene associations freely available.

INTRODUCTION

Structural variants (SVs) are a significant source of genetic diversity and have been increasingly recognized as contributors to rare and complex disease (Groza et al. 2024; Kainer et al. 2023; Merker et al. 2018). Typical SVs are genetic alterations longer than 50 bp in length and classified as deletions, duplications, insertions, and inversions. Considering such lengths, the analysis of SVs has been constrained by the limitations of short-read sequencing, which often fails to accurately resolve complex and repetitive genomic rearrangements (Merker et al. 2018; Marx 2023). In recent years, new long-read high fidelity (HiFi) sequencing technologies have enabled the detection and characterization of SVs with greater resolution and accuracy (Hon et al. 2020; Kucuk et al. 2023).

Despite improved SV detection, in a rare disease context it remains highly challenging to identify candidate pathogenic SVs. One solution to this challenge is to first understand the landscape of disease-associated genes for a given rare disorder. One such approach is in integrating trait-relevant polygenic scores (PGS), which quantify inherited genetic predisposition for a given phenotype by estimating disease effects across multiple genetic loci (Simona et al. 2023). In a rare disease context, increased polygenic liability has been observed for many disease cohorts, including in schizophrenia (Davies et al. 2020), severe neurodevelopmental disorders (Niemi et

52 al. 2018), and autism spectrum disorder (Gaugler et al. 2014). Through integrating PGS, we can
53 gain enhanced resolution into genes with impacts on rare diseases.
54
55 In this study, we leveraged PacBio long-read HiFi genomes available in the Genomic Answers for
56 Kids (GA4K) study at Children's Mercy Research Institute to investigate the role of rare structural
57 variants (rSVs) in autism spectrum disorder. The GA4K study is a large-scale, phenotypically
58 diverse pediatric rare disease cohort of over 13,000 individuals (Cohen et al. 2022). At the time
59 of our study, there were 497 rare disease probands with HiFi long read whole genome sequences
60 (lrWGS) completed. Further integrating open-source PGS available from PGS Catalog (Lambert
61 et al. 2021), we computed individual PGS liability for autism spectrum disorder and additional
62 traits and phenotypes. Our study focused on rare deletions, duplications, insertions, and
63 inversions overlapping trait-relevant PGS genes. We further conducted a genomic constraint
64 analysis of rSVs, providing insights on their evolutionary tolerance and potential pathogenicity.
65
66 Overall, we found an enrichment of deletions, duplications and insertions at trait-relevant genes
67 and a significantly higher constraint of duplications in individuals with autism. Furthermore, we
68 trace SV-overlapped genes back to autism and implicate anterograde trans-synaptic signaling
69 using ontology enrichment. Our study utilizes the intersection of rSVs, trait-relevant PGS, and
70 genomic constraint to explore the structural variant landscape in a large-scale rare disease cohort,
71 with a focus on genes associated with autism spectrum disorder.
72
73 **RESULTS**
74 ***Structural variant type and frequency show prevalence of deletions and insertions***
75 We analyzed 497 probands with HiFi long read whole genome sequences (lrWGS) in the
76 Children's Mercy GA4K database (N = 254 male; N = 243 female). We assessed ancestry using
77 Somalier, which predicted 437 European, 58 Admixed American, and 2 Asian individuals.

78 Structural variants (SVs) were called using PacBio SV (PBSV) after alignment with human
79 genome GrCh38 (Pedersen et al. 2020). All individuals were assessed for total number of SVs
80 (**Supplemental Table 1, Supplemental Figure 1**). We developed a python application called
81 *GA4K-SV-FINDER* where researchers can query these SVs, including allelic frequencies and
82 overlapping genes, by inputting user-specified genomic regions or gene names (**Data**
83 **Availability**).
84
85 To identify potential large-effect rare structural variants (rSVs) impacting rare disease phenotypes
86 we removed common SVs from the dataset (>0.5% MAF; **Methods**) (**Supplemental Figure 1**).
87 This reduced a total of 10,631,806 SVs to 1,017,180 rSVs (**Figure 1A**). The mean number of SVs
88 per individual was 21,521, and 2,047 for rSVs (**Figure 1A**).
89
90 Representing the lowest frequency (0.5%) of rSV type, inversions (INVs) had the largest mean
91 length (25,022 bp) compared to deletions (DEL, 20.3%, 879 bp), duplications (DUP, 2.4%, 8,000
92 bp), and insertions (INS, 76.8%, 885 bp) (**Figure 1**). When comparing SV type across sexes,
93 frequencies were similar and within 5% variability between males and females for length and
94 count. The frequency and length of rSVs in our study closely correspond with population
95 frequencies seen in other studies (Kosugi et al. 2019; Guo et al. 2021).

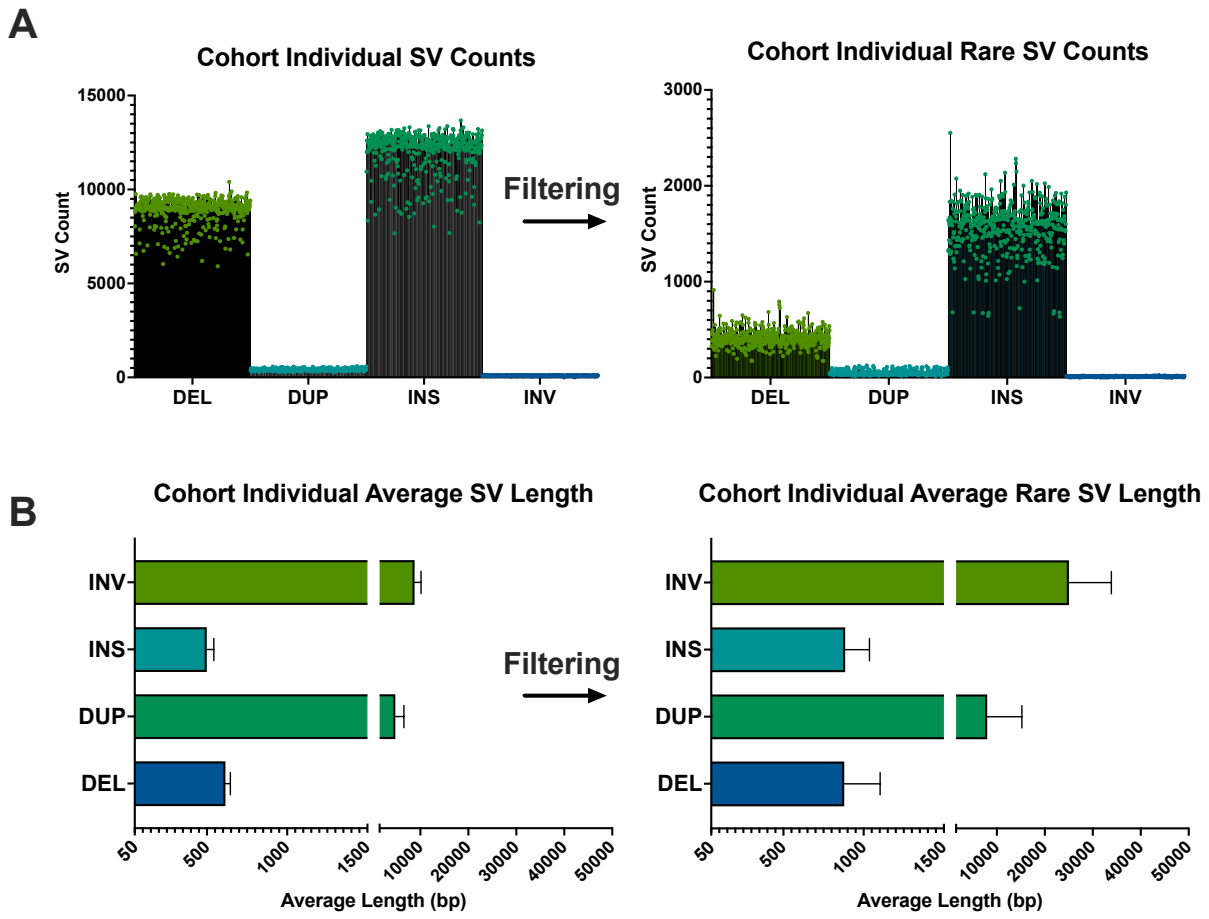


Figure 1: SV type frequency and average length in study population. (**A**) Two dot-bar graphs representing SV type (Deletions (DEL), duplications (DUP), insertions (INS), inversions (INV)) counts before and after filtering for rare variants. (**B**) Two horizontal bar graphs with outlier whiskers displaying the average lengths for SVs before and after filtering.

Cohort phenotypes aligned with polygenic scores in five GA4K phenotypes

We calculated PGS for all probands across various rare phenotypes (short stature/height; global developmental delay/fluid intelligence; autism/autism; hypotonia/grip strength; and seizure/epilepsy). Probands were then grouped by their phenotype to construct case/control cohorts. Probands not possessing a given phenotype were considered controls in each mapping. We chose to analyze the top-five most frequent phenotypes in GA4K that had accompanying lrWGS. At the time of assessment, the largest phenotype was Global Developmental Delay at 138 cases, followed by Seizures (86), Hypotonia (62), Autism (54), and Short Stature (45) (**Figure**

2A). We observed significant differences in Autism and Short Stature PGS, with scores higher for Autism and lower for height, respectively (**Figure 2A**).

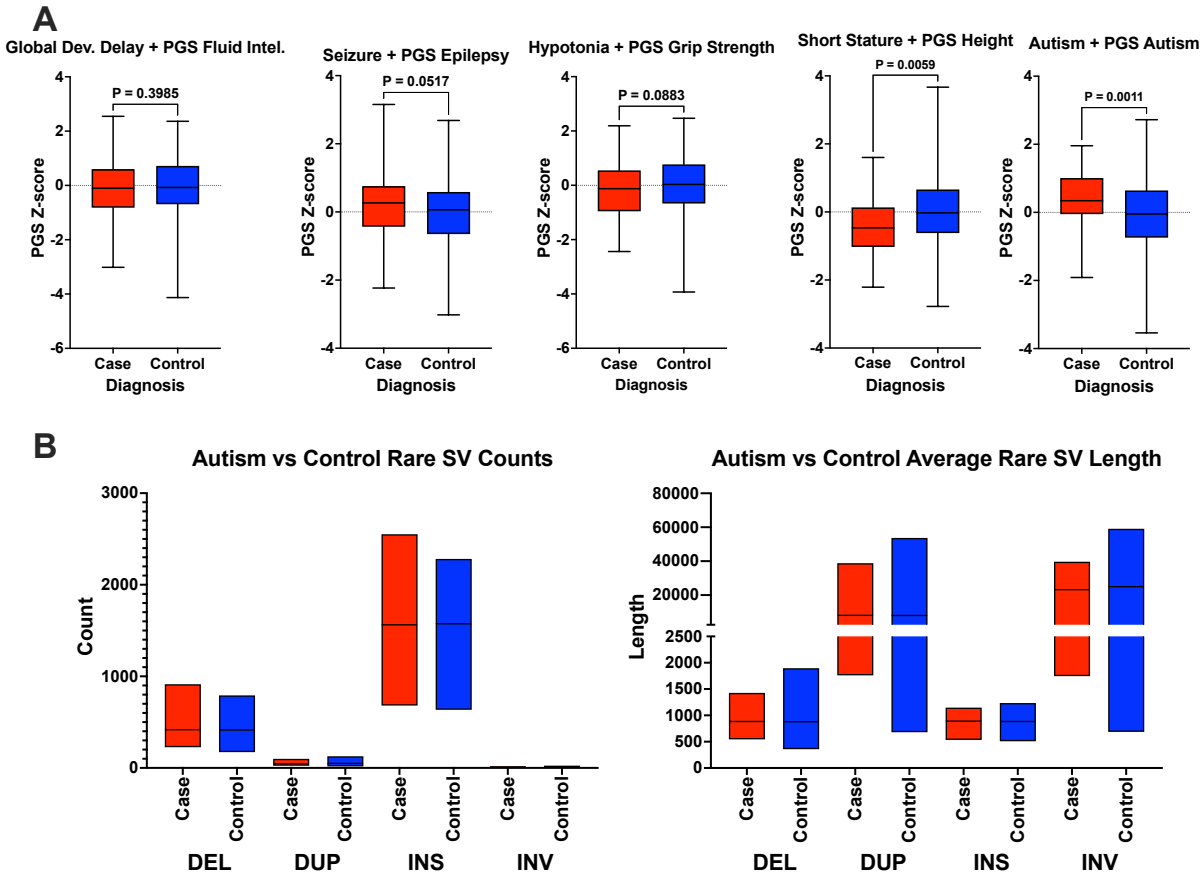


Figure 2: Box plots show significant mappings for Autism and Short Stature. (**A**) Box and whisker plots showing differences in selected PGS mapped to five rare disease phenotypes in GA4K for cases and controls. Red indicates case individuals. Blue indicates control individuals. PGS Z-score is indicated on the Y axis (N cohort = 497, N GDD cases = 138, N Seizure = 86, N Hypotonia = 62, N Autism = 54, N Short Stature = 45). P-values were computed using a Mann-Whitney test. (**B**) Box plots showing the count of rSV type (DEL, DUP, INS, INV) (left) and rSV type length (right) across individuals for cases (N = 54) and controls (N = 443) in the Autism cohort.

Focusing on the larger case number of the two significant phenotypes, we then compared rSVs in the Autism group (N cases = 54) group with controls (N controls = 443). We observed that probands with Autism had slightly higher median counts of DUPs and INVs, and longer mean lengths, except for INVs, when compared with controls (**Figure 2B**).

127 ***High weight PGS genes reveal rSV enrichment in Autism***

128 Due to the complex genetic nature of Autistic disorders, we decided to further focus on individuals
129 in our cohort with Autism. We evaluated genes harboring variants included in a PGS for Autism
130 (PGS Catalog ID: PGS000327). Each variant is weighted according to its relative association with
131 a phenotype (**Methods**). We intersected the PGS variant coordinates with the coordinates of
132 genes. This revealed 5,830 genes intersected by at least one overlapping PGS variant. We then
133 intersected cohort rSVs with these PGS-linked genes and found at least one overlap in 4,271
134 genes (**Supplemental Table 2**). Among this gene set, Ras Suppressor Protein 1 (RSU1) had the
135 highest PGS weighting (gene weight = 0.139). The lowest weighted gene was Importin 7 (IPO7)
136 (gene weight = 0.023) (**Supplemental Table 2; Figure 3**). Additionally, there are 262 genes
137 associated with the Human Phenotype Ontology (HPO) ID for autism spectrum disorder
138 (HP:0000717) (Kohler et al. 2021); of those genes, 84 were found to have overlap with rSVs
139 (average gene weight = 0.05).

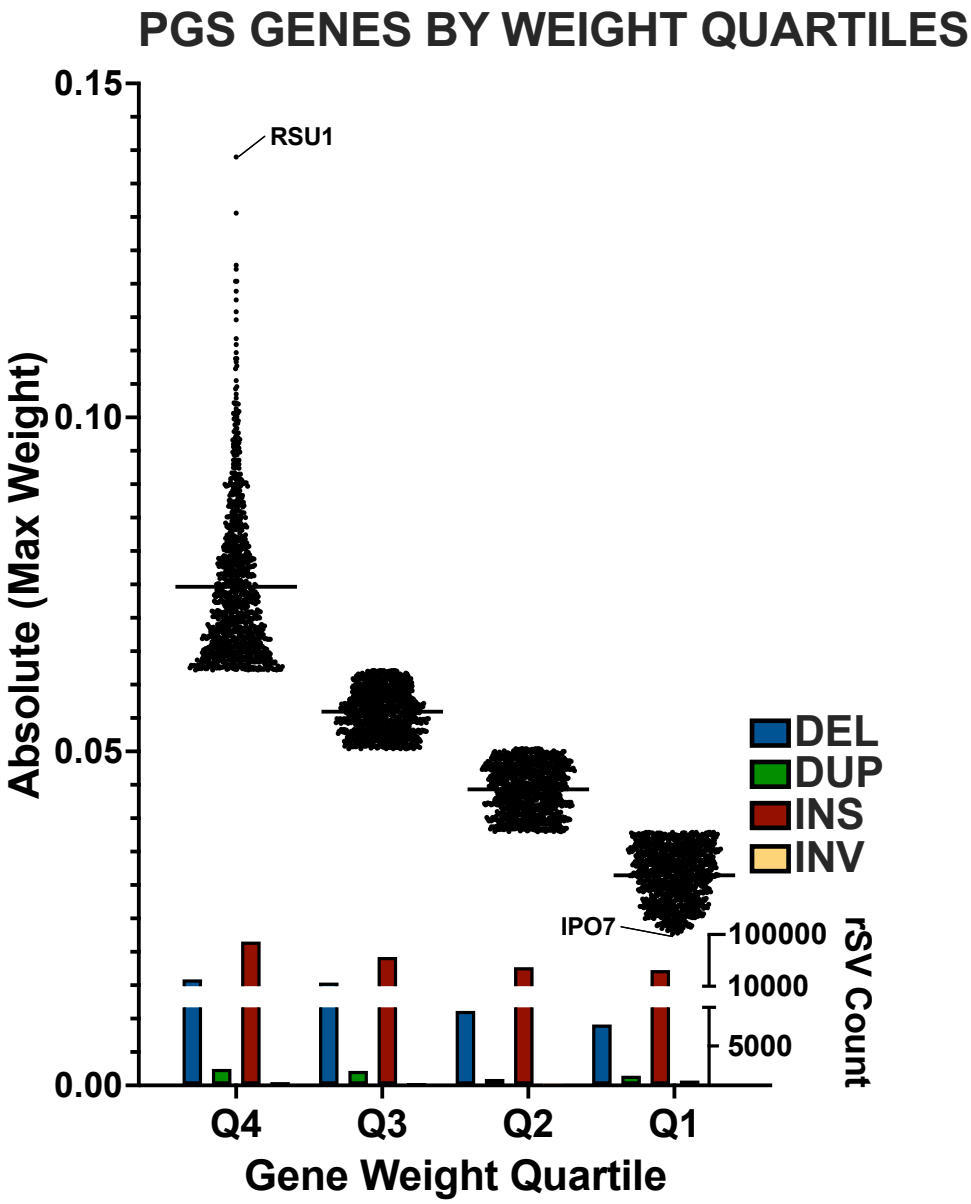


Figure 3: Top quartile PGS genes show a wide range of weights and a high count of insertions. Dot plot of genes within their quartiles based on the intersecting PGS gene weights. The highest weighted genes are in Q4 on the left. The lowest weighted genes are in Q1, to the right. The y-axis for gene weights displays the range of gene weights. The inset bar plot shows the rSV count for each quartile. A color legend is displayed above the y-axis.

We then organized these 4,271 genes into 4 quartiles according to their PGS gene weight (**Figure 3**). Each quartile contained 1063-1071 genes. All 4,271 genes were evaluated for frequency of rSVs both within the gene body and 200kb up- and down-stream of the gene body (**Supplemental**

150 **Table 2**). Focusing on the top quartile, we observed 112,100 rSV overlaps. INSs were the most
151 abundant, both in the gene body (78.3%) and in the flanking regions (upstream: 77%,
152 downstream: 79.3%). This was followed by DELs (body: 19.6%, up: 20.5%, down: 18.1%), DUPs
153 (body: 1.8%, up: 2.1%, down: 2.4%), and INVs (body: 0.2%, up: 0.3%, down: 0.2%). INS and DEL
154 frequency in the gene body paralleled a descending gene weight across all quartiles, where lower
155 gene weights correlated with a lower number of the SV (**Figure 3**). Duplications followed a similar
156 trend, but the bottom quartile had more than the third quartile. Inversions were more abundant in
157 the bottom quartile, however 431 of 514 INVs were found overlapping the gene
158 Chymotrypsinogen (CTRB1/CTRB2) – a gene region harboring many INVs as previously
159 described (Rosendahl et al. 2018). When evaluating the up- and down-stream 200kb flanking
160 regions for each gene, higher up- and down-stream rSV frequencies associated with a lower gene
161 weight quartile, where the lower gene weight had higher flanking rSV counts.
162
163 We then quantified genes in the top quartile of PGS effect with rSV frequency significantly different
164 (>5% diff, Chi-Square test, p-value < 0.05) between Autism and control groups (**Table 1;**
165 **Methods**). In doing this we move from entire cohort frequencies to an Autism-control enrichment.
166 While we define significant frequency differences between groups where genes have >5%
167 enrichment over the other group, our aim here is to resolve higher impact rSV-gene enrichments
168 between Autism and controls (>10%) in the top quartile for in-depth downstream analysis (**Table**
169 **1**). Following this approach, we observed that DELs were enriched in GRNA Binding Motif Single
170 Stranded Interacting Protein 3 (RBMS3), Contactin-associated Protein 2 (CNTNAP2), Human
171 Immunodeficiency Virus Type I Enhancer Binding Protein 3 (HIVEP3), Ribosomal Protein S6
172 Kinase A5 (RPS6KA5), Myomesin 2 (MYOM2), and Astrotactin-2 (ASTN2) (**Table 1**). Only one
173 gene, Repeat-Containing Protein 7 (WDR7), had >10% enrichment of DUPs (**Table 1**). The most
174 frequent rSV type, INSs, was enriched in Catenin Alpha-3 (CTNNA3), Neuronal Guanine
175 Nucleotide Exchange Factor (NGEF), Calmodulin-lysine N-methyltransferase (CAMKMT), Coiled-

176 Coil Domain Containing 171 (CCDC171), IQ Motif Containing GTPase Activating Protein 2
177 (IQGAP2), Myosin ID (MYO1D), Death-associated Protein Kinase 1 (DAPK1), and Neuregulin 3
178 (NRG3) (**Table 1**). No INVs were found to be > 10% differentially enriched (**Table 1**). The genes
179 RBMS3, CTNAP2, HIVEP3, ASTN2, CTNNA3, and DAPK1 have been previously reported in
180 connection to autism spectrum disorders (Pereanu et al. 2018).

		Gene Body	
Gene	**%>Control**	**SV Type**	**Locus**
*CTNNA3**	20%	INS	chr10:66007727-67358623
*WDR7**	19%	DUP	chr18:56826057-57024955
*NGEF**	17%	INS	chr2:232890042-232996364
*RBMS3**	16%	DEL	chr3:28628062-29007547
*CNTNAP2**	16%	DEL	chr7:146151646-148366160
*CAMKMT**	15%	INS	chr2:44410220-44589793
*CCDC171**	15%	INS	chr9:15568002-15883218
*HIVEP3**	14%	DEL	chr1:41531282-41606928
*RPS6KA5**	14%	DEL	chr14:91044310-91044392
*MYOM2**	13%	DEL	chr8:2081050-2130306
*IQGAP2**	12%	INS	chr5:76591340-76699893
*MYO1D**	11%	INS	chr17:32566830-32862227
*DAPK1**	10%	INS	chr9:87575220-87652891
*NRG3**	10%	INS	chr10:82132483-82534600
*ASTN2**	10%	DEL	chr9:116439185-117282935
NIPAL1	9%	DEL	chr4:47976159-47976340
AUTS2	9%	DEL	chr7:70003736-70742818
PDE11A	8%	DEL	chr2:177946020-177947205
RYR2	7%	DEL	chr1:237179371-237591190
AMPH	7%	DEL	chr7:38610296-38610398
KCNQ3	6%	INS	chr8:132128593-132187729
ATP8A2	6%	DEL	chr13:25512904-25997038
PTPRT	5%	DUP	chr20:42622625-42623766
SLC35F4	5%	DEL	chr14:57652619-57904836
EYS	5%	INV	chr6:65180037-65180262
LEPR	5%	DEL	chr1:65575256-65578219
TCAIM	5%	DUP	chr3:44378928-44385564
FHOD3	5%	DEL	chr18:36763451-36763532
TSC22D2	5%	INS	chr3:150448184-150448513

Upstream			
Gene	**%>Control**	**SV Type**	**Locus**
*CDH2**	23%	INS	chr18:27855601-27856513
*SFSWAP**	19%	DEL	chr12:131560091-131701808
*F13A1**	18%	INS	chr6:6045738-6046653
*TPO**	18%	DEL	chr2:1221441-1346236
*SH2D4A**	17%	INS	chr8:19172662-19172667
*RSPO4**	16%	INS	chr20:764442-949516
*CACNA1A**	15%	INS	chr19:13185003-13185154
*BAZ1A**	15%	INS	chr14:34557496-34704350
*TNS3**	14%	DEL	chr7:47111621-47112597
*SLC12A7**	14%	INS	chr5:853602-1045207
*C14orf159**	14%	DEL	chr14:91044310-91044392
*BCKDHB**	10%	INS	chr6:80054612-80058252
DHX30	9%	INS	chr3:47784344-47784344
KATNAL1	9%	INS	chr13:30003095-30030025
FAM19A5	9%	DEL	chr22:48306054-48415152
VRK2	9%	DEL	chr2:57755085-57755160
FBXO31	8%	DEL	chr16:87162696-87219126
BCAS1	8%	DUP	chr20:53868937-53869001
CUL1	8%	INS	chr7:148595522-148617124
ARMC4	8%	DEL	chr10:27707038-27753250
GALNT2	8%	INS	chr1:229870399-229870918
SPTBN5	7%	DUP	chr15:41656319-41656377
MSRB3	7%	INS	chr12:65083822-65194817
VWA8	7%	DEL	chr13:41399627-41400559
JARID2	7%	DEL	chr6:15052128-15194297
DPP10	7%	DEL	chr2:114371011-114372499
TNKS	7%	DUP	chr8:9486153-9496734
ARPP21	7%	DEL	chr3:35564178-35564416
RAB3C	7%	INS	chr5:58420384-58422140
MLIP	7%	INS	chr6:53744882-53801696
AKR1C1	6%	DEL	chr10:4815750-4815964
ITGA7	6%	DEL	chr12:55661570-55661646
GPR141	6%	INS	chr7:37498387-37527837
BCO1	6%	INS	chr16:81144808-81235200
DNAJC12	6%	INS	chr10:67755364-67755364
CNTN4	6%	DEL	chr3:1917906-1987559
ESR2	6%	INS	chr14:63948722-64038340
LRRTM3	6%	INS	chr10:66791760-66867117

GLIS1	5%	DEL	chr1:53408988-53468247

Downstream

Gene	%>Control	SV Type	Locus
*BANP**	17%	DEL	chr16:88123428-88237677
*SOX5**	16%	INS	chr12:24007152-24007683
*PPP2R2D**	16%	INS	chr10:132149554-132159794
*AUTS2**	16%	INS	chr7:70845576-70872295
*TENM3**	16%	INS	chr4:182831825-182834588
*CAMTA1**	16%	INS	chr1:7803433-7953145
*U2AF2**	16%	INS	chr19:55835451-55835451
*SCCPDH**	16%	INS	chr1:246783075-246785899
*EFHB**	15%	INS	chr3:19969952-19969952
*C2orf76**	12%	DEL	chr2:119444380-119552830
*CUL1**	12%	INS	chr7:148839053-148944552
*MTUS2**	12%	DEL	chr13:29558245-29559472
*ITGA7**	11%	INS	chr12:55871380-55905934
*PEMT**	10%	INS	chr17:17619242-17656596
CORIN	9%	DEL	chr4:47976159-47976340
TNS3	9%	INV	chr7:47614271-47615264
SYNE2	9%	INS	chr14:64354319-64354319
RB1CC1	8%	INS	chr8:52878480-52928679
UGDH	7%	DEL	chr4:39619159-39712200
IFT140	7%	DEL	chr16:1651054-1761705
ASIC2	7%	INS	chr17:34302244-34302244
XIRP2	7%	INV	chr2:167316081-167424243
SV2B	7%	DEL	chr15:91304256-91304539
SOX30	7%	INS	chr5:157778905-157778905
CTNND2	6%	DEL	chr5:11930715-12057490
KLB	6%	DEL	chr4:39619159-39619751
FBN2	6%	INS	chr5:128818627-128818639
MYH15	5%	DUP	chr3:108705618-108705704
MCTP1	5%	INS	chr5:95299561-95409472
ATG10	5%	DEL	chr5:82466234-82466779
DNAJC6	5%	DEL	chr1:65575256-65578219
ALMS1	5%	DEL	chr2:73624051-73686145
PRKD1	5%	INS	chr14:30385461-30385461
CFAP77	5%	DEL	chr9:132716657-132751514
SEPT14	5%	INS	chr7:55887600-55887855

181

Table 1: SV enriched genes strongly associated with Autism. Genes with significantly (Chi-Square, p-value < 0.05) higher frequency rSV enrichment (>5%) in Autism individual gene bodies (top), upstream (middle) and downstream (bottom) regions. The group frequency differential and SV type is displayed to the right of the gene. *Indicates high impact (>10%) rSV-gene differential enrichment in Autism over controls.

We also assessed enrichment in the 200kb flanking regions of these high weight PGS-linked genes, revealing significant enrichment of DELs down-stream of BTG3 Associated Nuclear Protein (BANP), MOB Kinase Activator 1A (C2orf6), and Microtubule Associated Scaffold Protein 2 (MTUS2). However, most down-stream enrichments were INSs proximal to SRY-box Transcription Factor 5 (SOX5), Protein Phosphatase 2 Regulatory Subunit B Delta (PPP2R2D), Autism Susceptibility Candidate 2 (AUTS2), Saccharopine Dehydrogenase (SCCPDH), Teneurin Transmembrane Protein 3 (TENM3), Calmodulin Binding Transcription Activator 1 (CAMTA1), U2 Small Nuclear RNA Auxiliary Factor 2 (U2AF2), EF-Hand Domain Family Member B (EFHB), Cullin 1 (CUL1), Integrin Subunit Alpha 7 (ITGA7), and Phosphatidylethanolamine N-Methyltransferase (PEMT) (**Table 1**). Furthermore, we observed an enrichment of DELs up-stream of Splicing Factor SWAP (SFSWAP), Thyroid Peroxidase (TPO), Tensin 3 (TNS3), and D-Glutamate Cyclase (C14orf159), as well as INSs enriched up-stream of genes Cadherin 2 (CDH2), Coagulation Factor XIII A Chain (F13A1), SH2 Domain Containing 4A (SH2D4A), R-Spondin 4 (RSPO4), Calcium Voltage-Gated Channel Subunit Alpha1 A (CACNA1A), Bromodomain Adjacent To Zinc Finger Domain 1A (BAZ1A), Solute Carrier Family 12 Member 7 (SLC12A7), and Branched Chain Keto Acid Dehydrogenase E1 Subunit Beta (BCKDHB). We did not observe an enrichment for DUPs or INVs within the PGS-linked gene set.

While we focused on differential enrichment between groups, it is also important to note that the two genes with the highest frequency of rSVs in Autism individuals were not found in SFARI or AutDB Autism databases (MYOM2, 100% [state what this percentage means, just to remind the reader]; NGEF, 98%) (**Table 1; Figure 4**). Moreover, the top 3 most frequent genes associated

with Autism – CTNNA3 (43%), CNTNAP2 (35%), and RBMS3 (26%), while differentially enriched in Autism individuals, were less frequent than those genes with previous associations to the disease (**Table 1; Figure 4**) (Pereanu et al. 2018). The only gene with enrichments in the gene body as well as a flanking region (downstream) was AUTS2, where DELs were found in the gene body in 13% of individuals and 42% had INSs in the downstream region (**Figure 4**). Furthermore, genes TNS3 (Upstream: 22% DEL; Downstream: 13% INV), CUL1 (Upstream: 11% INS; Downstream: 26% INS), and ITGA7 (Upstream: 7% DEL; Downstream: 24% INS) had enrichments in both up- and down-stream regions (**Figure 4**). Three of four genes with differential enrichments in more than one region of a gene were found on chromosome 7 (**Figure 4**).

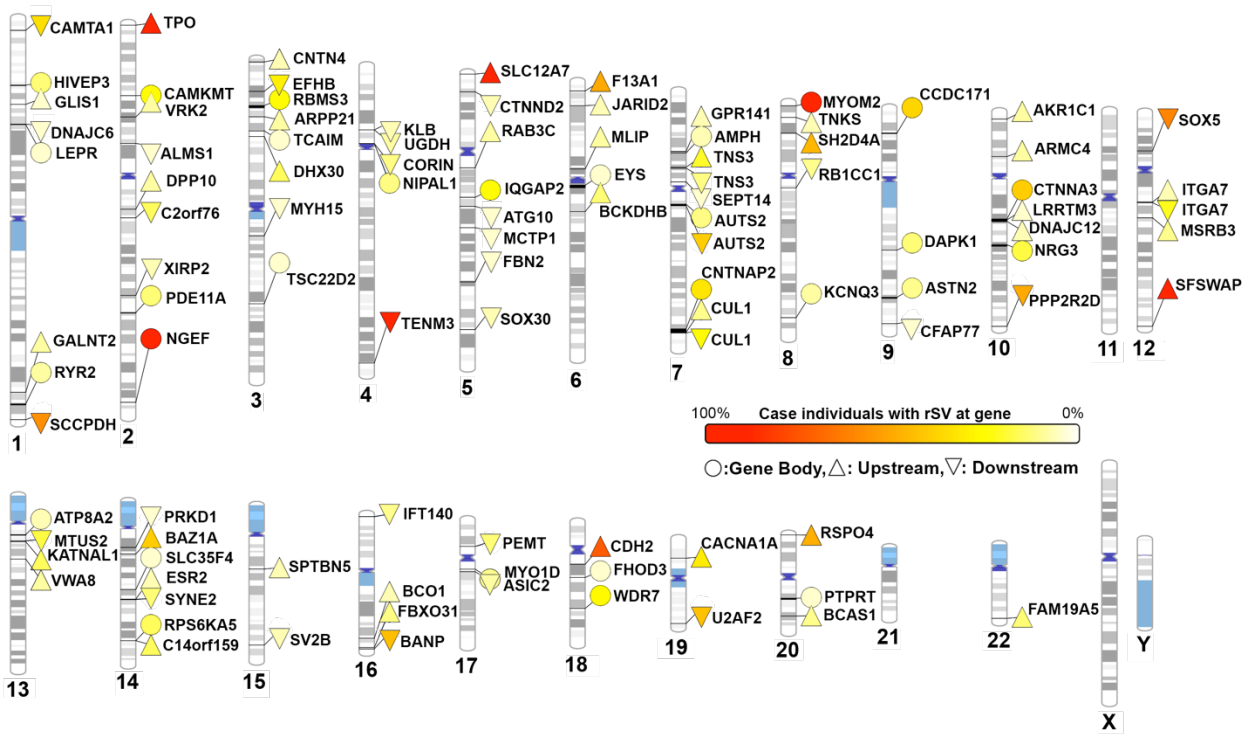


Figure 4: Genes with differential rSV enrichment tend have close chromosomal proximity to one another whereas rSV enriched genes with high frequency tend to have more distance between. PhenoGram showing the frequency of individuals with rSVs at the cut-off genes by chromosomal position (color scale from 0-100% frequency of case individuals). Cytogenic Giemsa bands show GC- (blue) and AT-rich (gray) regions. Symbols indicate whether the enrichment occurs in the gene body, the upstream, or the downstream regions of the listed gene. The color gradient of symbols indicates the degree of enrichment in the case cohort, with higher enrichment indicated in orange and red.

230 Insertions were the most common type of enrichment (49/103), followed by DELs (44/103), DUPs
231 (7/103), and INVs (3/103). Two of the 19 chromosomes with rSV enrichments at high weight PGS-
232 linked genes were exclusive to INSs. We also found that while DUPs were relatively infrequent,
233 differentially enriching the gene body of 3 of 62 total genes, DUPs at WDR7 had the second
234 highest differential percentage of an enriched gene body (19% > Controls, p value = 0.0001)
235 (**Table 1**). Though WDR7 was not found in AutDB or in HPO gene associations, it is positioned
236 down-stream of CDH2 on chromosome 18, the most frequently up-stream enriched gene (76%)
237 associated with Autism (Abrahams et al. 2013). Additionally, PTPRT and BCAS1, two more of the
238 seven genes enriched by DUPs, are both positioned on chromosome 20 and associated with
239 Autism. Moreover, enriching DUPs were at genes with a mean gene weight in the top 3% of PGS-
240 linked genes, whereas DELs, INSs, and INVs were in the top 14%. Overall, we identified genes
241 with differential SV enrichments in Autism over control individuals, specific to the rSV type and
242 locus, finding a high frequency enrichment of DELs and INSs overlapping PGS-linked genes and
243 their flanking regions.
244
245 ***SV enriched genes in the top quartile of PGS effect size significantly associate with autism***
246 ***and potentially associated pathways***
247 After deriving a list of genes with high enrichment of rSVs in Autism (>5%; N genes = 103)
248 (**Table1**), we sought to determine if these genes specifically enrich the Autism phenotype or
249 candidate biological pathways. We used the DisGeNET database with Enrichr and found a
250 significant enrichment (adj. P = $5x10^{-06}$) with Autistic Disorder using our gene list (**Table 1**).
251 Furthermore, using the GO Cellular Component database, we found that rSVs were enriched in
252 cases for the genes Formin Homology 2 Domain Containing 3 (FHOD3), MYO1D, Katanin
253 Catalytic Subunit A1 Like 1 (KATNAL1), AUTS2, DAPK1, MTUS2, Amphiphysin (AMPH),
254 IQGAP2, CORIN, and Spectrin Repeat Containing Nuclear Envelope Protein 2 (SYNE2) all of
255 which are involved in the cytoskeleton (GO:0015629; GO:0005856) (adj. p-value = 0.02) (**Table**

256 **1**). Our results here provide a framework for linking increased rSV burden in Autism to potential
257 downstream impacts on gene pathways relevant to Autism pathogenesis.
258
259 ***Duplications are enriched at regions of high constraint in top PGS genes***
260 We next used 1kb gnomAD constraint windows to evaluate potential pathogenicity amongst rSVs
261 (**Methods**). In short, higher constraint indicates an evolutionary intolerance to variation within that
262 locus. Though constraint scores were not available across every 1kb region of the genome, we
263 were able to intersect constraint windows for 7.9% of rSVs in the gene body and 6.7% in the
264 flanking regions. We graphed the density of rSV constraint scores across different SV types
265 (**Figure 5**). This analysis revealed that rare DELs and DUPs in the Autism group tended to overlap
266 higher constraint windows compared with controls (DEL: $p < 0.0001$, Autism z mean = 0.14,
267 control z mean = -0.08; DUP: $p < 0.0001$, Autism z mean = 0.48, control z mean = 0.03).
268 Significant differences were not observed for INSs or INVs (INS: $p = 0.66$, Autism z mean = 0.2,
269 control z mean = 0.18; INV: $p = 0.14$, Autism z mean = -2.77, control z mean = -2.6). Furthermore,
270 Autism DUPs tended to overlap higher constraint windows within 200kb up-stream of PGS-linked
271 genes (DUP: $p < 0.0001$, Autism z mean = 0.66, control z mean = 0.14). We observed a similar
272 pattern for DUPs in Autism individuals for rSVs within the down-stream region (DUP: $p < 0.0001$,
273 Autism z mean = 0.42, control z mean = 0.11). Thus, DUPs were the only SV type in Autism with
274 significantly higher constraint than controls in all three regions.

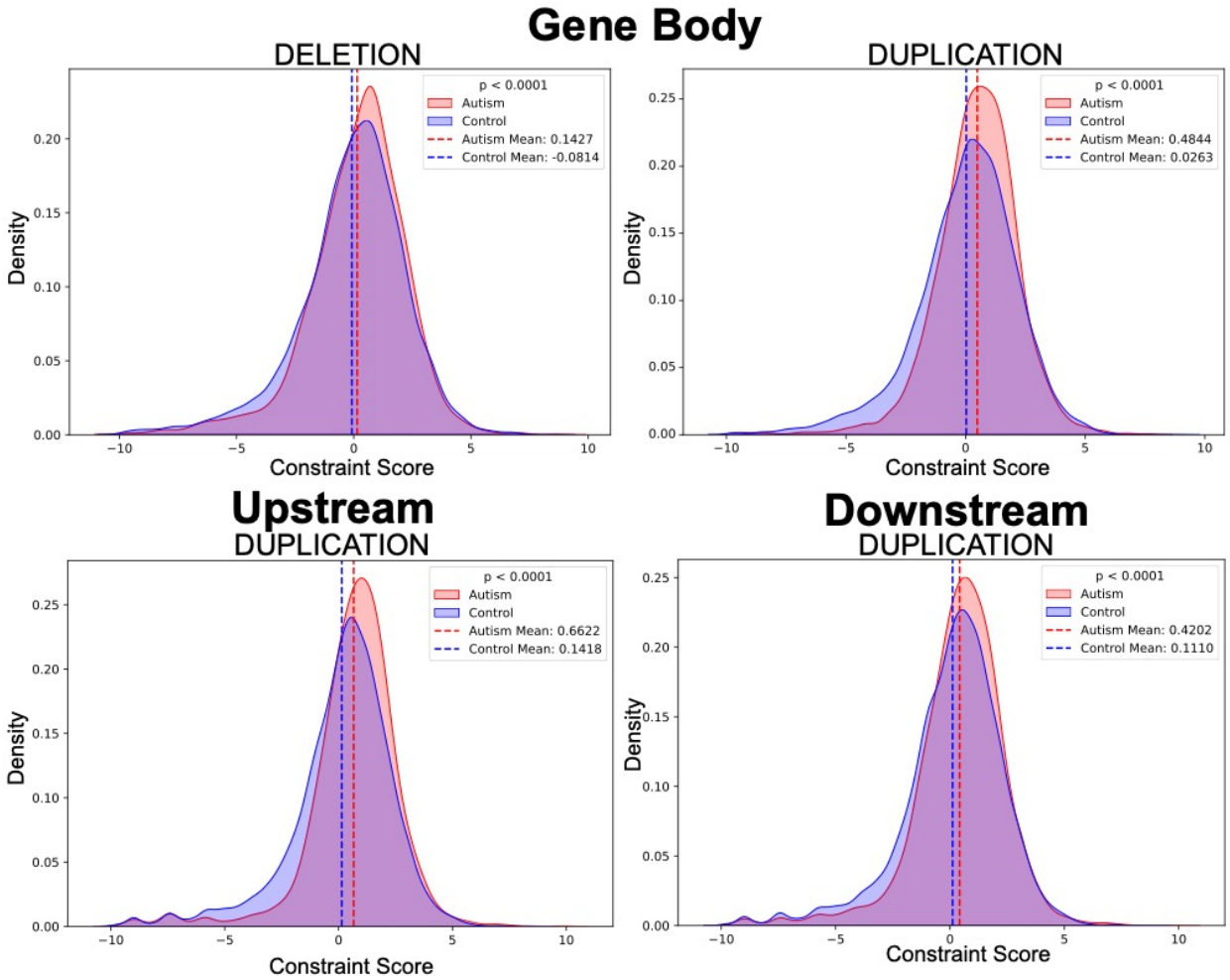


Figure 5: Duplication constraints are significantly different between Autism cases and controls inside and outside of the gene body. Four density graphs representing two SV types with significantly different constraint scores across Autism-control groups and gene regions. The top two graphs show significantly different constraints across groups for the gene body. The bottom two show up-stream and down-stream differences.

We then focused on DUPs in the top quartile of PGS gene weight and observed a significant enrichment in rSVs overlapping higher constraint regions in Autism compared to controls ($p = 0.001$, Autism z mean = 0.40, control z mean = -0.66) (**Figure 6**). This trend was not observed for genes in the bottom quartile of PGS effect ($p = 0.14$, case z mean = 1.0, control z mean = 0.54) (**Figure 6**) or across other SV types. In summary, DUPs were found to overlap higher constraint windows more frequently in the Autism group compared to controls for SVs overlapping, or proximal to, large- effect autism PGS genes.

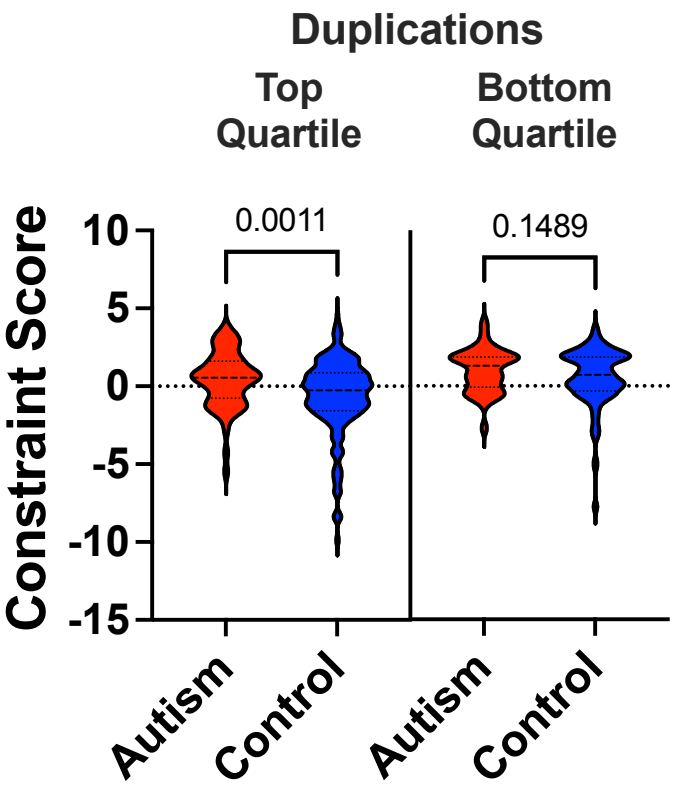


Figure 6: Constraint scores for duplications in top PGS quartile genes are significantly different across Autism cases and controls. Violin plot of Autism and control duplication constraint scores within top and bottom quartile genes.

DISCUSSION

Our study highlights the significance of rare structural variants (rSVs) in complex rare diseases such as Autism. By utilizing long-read PacBio HiFi sequencing, we have enhanced the detection and characterization of rSVs in a rare disease cohort, particularly deletions (DELs), duplications (DUPs), and insertions (INSs) observed within or proximal to autism genes. Our analysis indicates a higher incidence of INSs and DELs in genes linked to Autism, with DUPs occurring frequently in large-effect PGS genes with high genomic constraint, suggesting a potential pathogenic influence. The integration of PGS has provided further insight into the genetic predisposition of various phenotypes within the cohort by using correlations between PGS and the clinical presentation of Autism.

304

305 Large-scale lrWGS assays in a rare disease cohort such as GA4K enables insights into structural
306 variation not previously explored. This analysis identified specific genes with significant rSV
307 intersections in the context of Autism, such as INSs in the gene body of CTNNA3, the upstream
308 region of CDH2, and downstream of SOX5. Interestingly, all three genes were not only associated
309 with the PGS for Autism, but they were also found in autism gene databases (Banerjee-Basu and
310 Packer 2010; Pereanu et al. 2018). Their regional frequencies were complex, where the flanking
311 enrichments of CDH2 and SOX5 were seen in over 63% of the Autism group, and only 42% of
312 the group was enriched in the body of CTNNA3. This suggests that flanking enrichments of
313 previously correlated Autism genes may provide important insights into the pathogenic influence
314 of these genes on associated disease phenotypes. Additionally evaluating rSVs in genes with
315 investigative association to Autism, we further resolved rSV enriched genes in Autism individuals
316 as autistic-disorder- and cytoskeleton-specific through gene ontology enrichment. Furthermore,
317 the online database of Autism spectrum disorder genes by the Simons Foundation Autism
318 Research Initiative (SFARI) lists 1,176 genes related to the disorder through SNVs (Abrahams et
319 al. 2013). In our study, we implicated over 5,000 genes using PGS associations and resolved
320 4,271 of those genes with rSV overlaps in our cohort. Of those 1,176 SFARI genes, 503 were in
321 our list of intersected PGS-genes. That leaves 3,768 genes absent from SFARI but with PGS and
322 rSV Autism overlaps identified in our cohort. Because long-read mapping-based methods for SV
323 calling are more accurate than short-read approaches, due to their ability to provide more
324 accurate alignments by capturing entire SV alleles in a single fragment – additionally considering
325 that HiFi lrWGS is relatively new – our study underscores the importance in adopting lrWGS in
326 rare disease where no single nucleotide variant is implicated.

327

328 In future studies, there should be consideration into the complexities of rare disease phenotypes
329 and any potential comorbidities. Controlling for phenotype, age, gender, and family transmission

330 might strengthen gene-rSV associations. Our findings underscore the complexity of the genetic
331 factors impacting Autism risk and reinforce the importance of comprehensive analysis required in
332 uncovering the nuances of rare disease variants. These insights lay the groundwork for more
333 precise genetic diagnostics and the development of targeted interventions involving rSVs.
334

335 **METHODS**

336 **Study Cohort**

337 The study cohort described includes 497 affected probands across 419 families with 49% females
338 and 51% males. Subjects were enrolled in the Genomics Answers for Kids (GA4K) program.
339 Probands at age of enrollment ranged from 0 to 57 years (median = 8) (older individuals were
340 typically ascertained as follow-up from an affected family member). Subjects were eligible for the
341 study if they had a suspected genetic diagnosis based on clinical presentation and/or an existing
342 molecular or cytogenic findings. The study complies with all relevant ethical regulations as
343 approved by the Children's Mercy Institutional Review Board (IRB) (Study #11120514). Informed
344 or written consent was obtained from all participants prior to study inclusion. Participants were
345 not compensated for participation.
346

347 **SV Calling**

348 PacBio HiFi Sequencing results from GA4K proband subjects were processed through PacBio's
349 structural variant (PBSV, Version: 2.6.2) calling workflows after alignment to human reference
350 GRCh38. Variants achieved a PASS score if they had at least 1 read per strand, were at least
351 1000 bp from the end of a contig, and away (< 1000 bp) from gaps (run of >= 50 Ns) in the
352 reference assembly. Additionally, for PBSV, we used the following parameters: -ccs (Circular
353 Consensus Sequence/HiFi), -m 20 (SVs larger than 20 bp were used), -A 3 (3 supporting reads
354 were required to call a variant as homozygous alternate), -O 3 (3 supporting reads were required
355 to call a variant as heterozygous), -P 20 (minimum phred) were used for structural variant calling.

356 Due to known false-positives, CNVs were not included in variant analysis. Additionally, only SVs
357 greater than 50 bp in length were used for each variant type. rSVs were defined as having a minor
358 allele frequency (MAF) < 0.5% or allele count =< 4 in the study cohort. Somalier was used to
359 predict ancestry (> 0.5 probability).
360
361 **GA4K-SV-FINDER**
362 We created a python-based application for querying allele frequencies and genes associated with
363 SVs in our cohort. This application can be found, along with a readme and the raw data file, in a
364 GitHub repository available at https://github.com/smail-lab-cmh/ga4k-sv-finder.
365
366 **Polygenic Risk Scores**
367 Polygenic risk scores were calculated for each specified polygenic score using open-source PGS
368 available in the PGS catalog (www.pgscatalog.org). Associated PGS variant coordinates were
369 downloaded from their respective catalog entries. PGS variants were converted to hg38
370 coordinates using dbSNP26 (version 155) where necessary. PGS variants were restricted to
371 autosomes only and variants mapping to the HLA region were removed. Individual-level PGS
372 were calculated using PLINK (version 1.9) (Purcell et al. 2007) ("sum" flag) on all variants
373 available in the GA4K imputed genotype callset with R2>=0.8 ("exclude-if-info" flag). PGS scores
374 were converted to Z-scores within each PGS in the full GA4K EUR ancestry cohort (proband and
375 other available family members, N = 7,436). Individuals with an extreme outlier PGS Z-score
376 (abs(PGS Z-score)>=10) in one or more PGS were removed from the final PGS dataset.
377
378 **Intersect**
379 To create compatible chromosomal coordinate files for structural and PGS variants, pseudo-BED
380 files were first created for each PGS SNP coordinate file, containing the chromosome,
381 coordinates, base change, and associated weight within the PGS as a z-score. This file was first

intersected, using the bedtools intersect command (v2.31.0), with a gene position BED file (gencode v26 hg38) to determine the gene the SNP was associated with. After determining the genes associated with the PGS, they were then intersected with pseudo-BED files generated from each pbsv VCF from the cohort. This created intersect files that provide the chromosomal coordinates for the SV and the genes, which included sv type, length, weight, and overlap length.

Flanking Regions

Both up- and down-stream intersects were performed by first using BEDtools Slop command on gene start and end coordinates separately. For example, start coordinates are transformed into a 200kb upstream value, expanding the gene start coordinate within the boundaries of its respective chromosome. When intersecting these expanded genes with SVs, any SV found to enrich the gene body that is also matching an SV found in the new flanking region is removed from the flanking region analysis, as it was already intersecting the gene body. In doing this, we separated SVs found up- or down-stream of the gene from the SVs overlapping the gene body itself.

Gene Weighting

Gene coordinates were retrieved from the gencode v26 hg38 basic annotation file (gencodegenes.org). Gene coordinates were intersected by weighted PGS variant coordinates. For each gene, PGS variant weights were transformed to absolute values before taking the maximum value across all intersecting variants. In turn, this provided an absolute maximum weight per gene. Following gene weighting, gene coordinates were intersected by rSVs.

Constraint Scoring

Each rSV was intersected with the gnomAD v3 QCed genomic constraint by 1kb regions file (file name: constraint_z_genome_1kb.qc.download.txt.gz). This file scores 1kb regions throughout the genome with a constraint z-score value. Each of the proband rSVs were intersected with these

408 1kb constraint windows. For window density assessment of each rSV type, all intersected
409 constraint windows are considered. For top and bottom quartile assessment, the scores are
410 averaged across windows for each SV, excluding non-coding regions or any undefined region
411 lacking a constraint value. This produces a single score for each SV intersecting a gene.
412

413 **Statistical and Enrichment Analysis**

414 Graphing and statistics were generated using Graph Pad Prism 9 and Python (SciPy-Stats &
415 Pandas). For enrichment, SV frequencies between Autism and control groups were determined
416 and Chi-Square tests were used to determine significant differences. When SV counts were low
417 for a gene (<5), a Fisher's Exact test was used. These tests were performed for body, upstream,
418 and downstream of genes separately.
419

420 **Gene Analysis**

421 Gene analysis was conducted using the DISEASES database, AutDB, GeneCards Version 5.19,
422 and Enrichr (Grissa et al. 2022; Stelzer et al. 2016; Chen et al. 2013; Pereanu et al. 2018).
423

424 **ADDITIONAL INFORMATION**

425 Further information and requests for resources should be directed to and will be fulfilled by the
426 corresponding author, Cas LeMaster (clemaster@cmh.edu).
427

428 **DATA AVAILABILITY**

429 GA4K study data can be found at the ANVIL host at
430 https://anvilproject.org/data/studies/phs002206/workspaces. Code used, as well as the
431 application for querying SVs (GA4K-SV-FINDER) can be found in a git-hub repository at
432 https://github.com/smail-lab-cmh/ga4k-sv-finder.
433

ACKNOWLEDGEMENTS

C.S. is supported by NIH grant R35GM146966. We thank all the individuals that participated in making the GA4K study possible. This work was funded through internal institutional funds from Children's Mercy Research Institute and Children's Mercy Kansas City.

AUTHOR CONTRIBUTIONS

Conceptualization: C.S., T.P, C.L.; data analysis: C.L., C.S., C.S-S., B.G., W.C., J.J.; writing: C.L., C.S.; R.M.; editing: all authors; funding: C.S.

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

Abrahams, B. S., D. E. Arking, D. B. Campbell, H. C. Mefford, E. M. Morrow, L. A. Weiss, I. Menashe, T. Wadkins, S. Banerjee-Basu, and A. Packer. 2013. 'SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs)', *Mol Autism*, 4: 36.

Banerjee-Basu, S., and A. Packer. 2010. 'SFARI Gene: an evolving database for the autism research community', *Dis Model Mech*, 3: 133-5.

Chen, E. Y., C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, and A. Ma'ayan. 2013. 'Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool', *BMC Bioinformatics*, 14: 128.

Cohen, A. S. A., E. G. Farrow, A. T. Abdelmoity, J. T. Alaimo, S. M. Amudhavalli, J. T. Anderson, L. Bansal, L. Bartik, P. Baybayan, B. Belden, C. D. Berrios, R. L. Biswell, P. Buczkowicz, O. Buske, S. Chakraborty, W. A. Cheung, K. A. Coffman, A. M. Cooper, L. A. Cross, T. Curran, T. T. T. Dang, M. M. Elfrink, K. L. Engleman, E. D. Fecske, C. Fieser, K. Fitzgerald, E. A. Fleming, R. N. Gadea, J. L. Gannon, R. N. Gelineau-Morel, M. Gibson, J. Goldstein, E. Grundberg, K. Halpin, B. S. Harvey, B. A. Heese, W. Hein, S. M. Herd, S. S. Hughes, M. Ilyas, J. Jacobson, J. L. Jenkins, S. Jiang, J. J. Johnston, K. Keeler, J. Korlach, J. Kussmann, C. Lambert, C. Lawson, J. B. Le Pichon, J. S. Leeder, V. C. Little, D. A. Louiselle, M. Lypka, B. D. McDonald, N. Miller, A. Modrcin, A. Nair, S. H. Neal, C. M. Oermann, D. M. Pacicca, K. Pawar, N. L. Posey, N. Price, L. M. B. Puckett, J. F. Quezada, N. Raje, W. J. Rowell, E. T. Rush, V. Sampath, C. J. Saunders, C. Schwager, R. M. Schwend, E. Shaffer, C. Smail, S. Soden, M. E. Strenk, B. R. Sullivan, B. R. Sweeney, J. B. Tam-Williams, A. M. Walter, H. Welsh, A. M. Wenger, L. K. Willig, Y. Yan, S. T. Younger, D. Zhou, T. N. Zion, I. Thiffault, and T. Pastinen. 2022. 'Genomic answers for children: Dynamic analyses of >1000 pediatric rare disease genomes', *Genet Med*, 24: 1336-48.

Davies, R. W., A. M. Fiksinski, E. J. Breetvelt, N. M. Williams, S. R. Hooper, T. Monfeuga, A. S. Bassett, M. J. Owen, R. E. Gur, B. E. Morrow, D. M. McDonald-McGinn, A. Swillen, E. W. C. Chow, M. van den Bree, B. S. Emanuel, J. R. Vermeesch, T. van Amelsvoort, C. Arango, M. Armando, L. E. Campbell, J. F. Cubells, S. Eliez, S. Garcia-Minaur, D. Gothelf, W. R. Kates, K. C. Murphy, C. M. Murphy, D. G. Murphy, N. Philip, G. M. Repetto, V. Shashi, T. J. Simon, D. H. Suner, S. Vicari, S. W. Scherer, Brain International 22q, Consortium Behavior, C. E. Bearden, and J. A. S. Vorstman. 2020. 'Using common genetic variation to examine phenotypic expression and risk prediction in 22q11.2 deletion syndrome', *Nat Med*, 26: 1912-18.

Gaugler, T., L. Klei, S. J. Sanders, C. A. Bodea, A. P. Goldberg, A. B. Lee, M. Mahajan, D. Manaa, Y. Pawitan, J. Reichert, S. Ripke, S. Sandin, P. Sklar, O. Svantesson, A. Reichenberg, C. M. Hultman, B. Devlin, K. Roeder, and J. D. Buxbaum. 2014. 'Most genetic risk for autism resides with common variation', *Nat Genet*, 46: 881-5.

Grissa, D., A. Junge, T. I. Oprea, and L. J. Jensen. 2022. 'Diseases 2.0: a weekly updated database of disease-gene associations from text mining and data integration', *Database (Oxford)*, 2022.

Groza, C., C. Schwendinger-Schreck, W. A. Cheung, E. G. Farrow, I. Thiffault, J. Lake, W. B. Rizzo, G. Evrony, T. Curran, G. Bourque, and T. Pastinen. 2024. 'Pangenome graphs improve the analysis of structural variants in rare genetic diseases', *Nat Commun*, 15: 657.

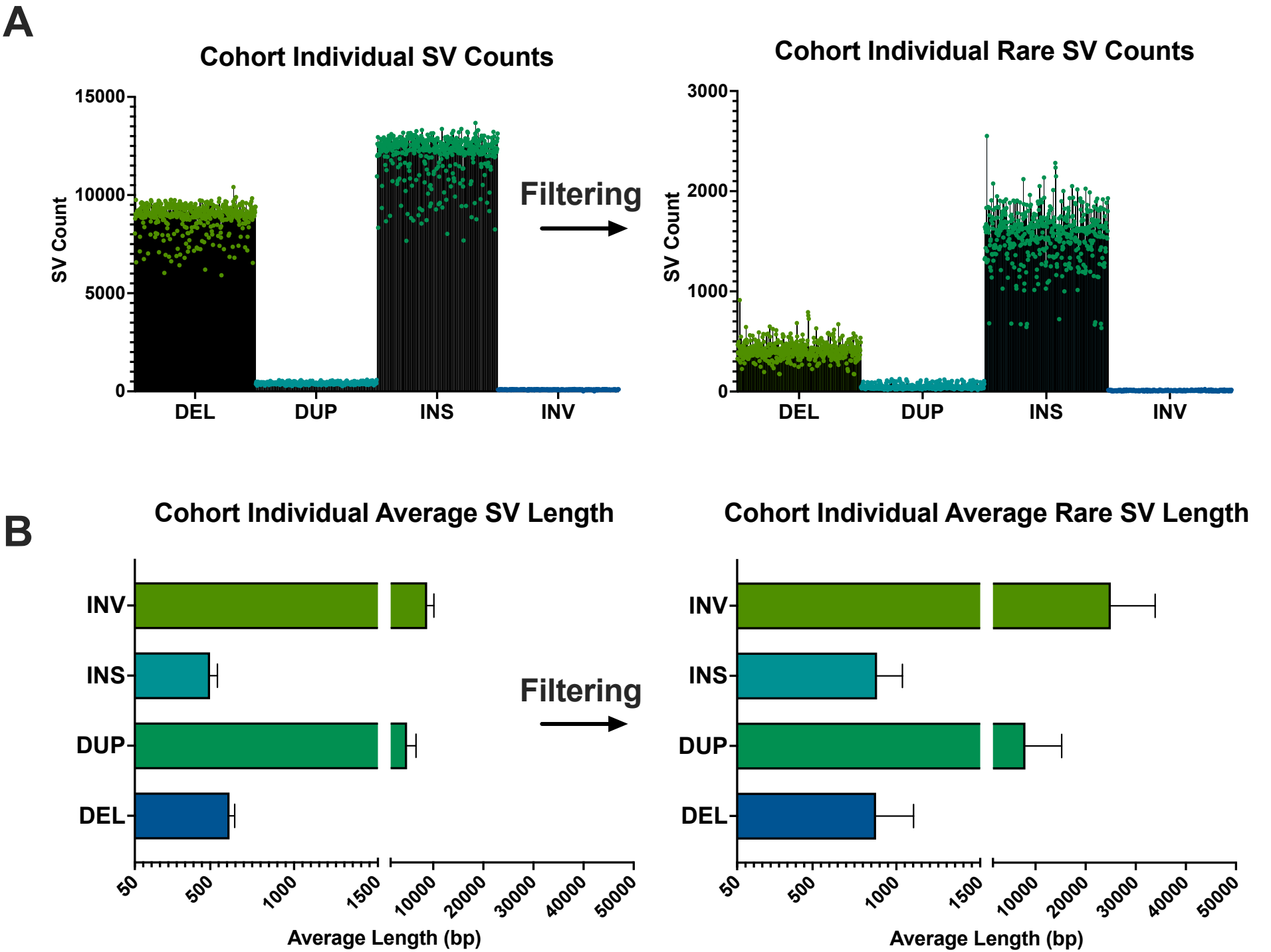
Guo, M., S. Li, Y. Zhou, M. Li, and Z. Wen. 2021. 'Comparative Analysis for the Performance of Long-Read-Based Structural Variation Detection Pipelines in Tandem Repeat Regions', *Front Pharmacol*, 12: 658072.

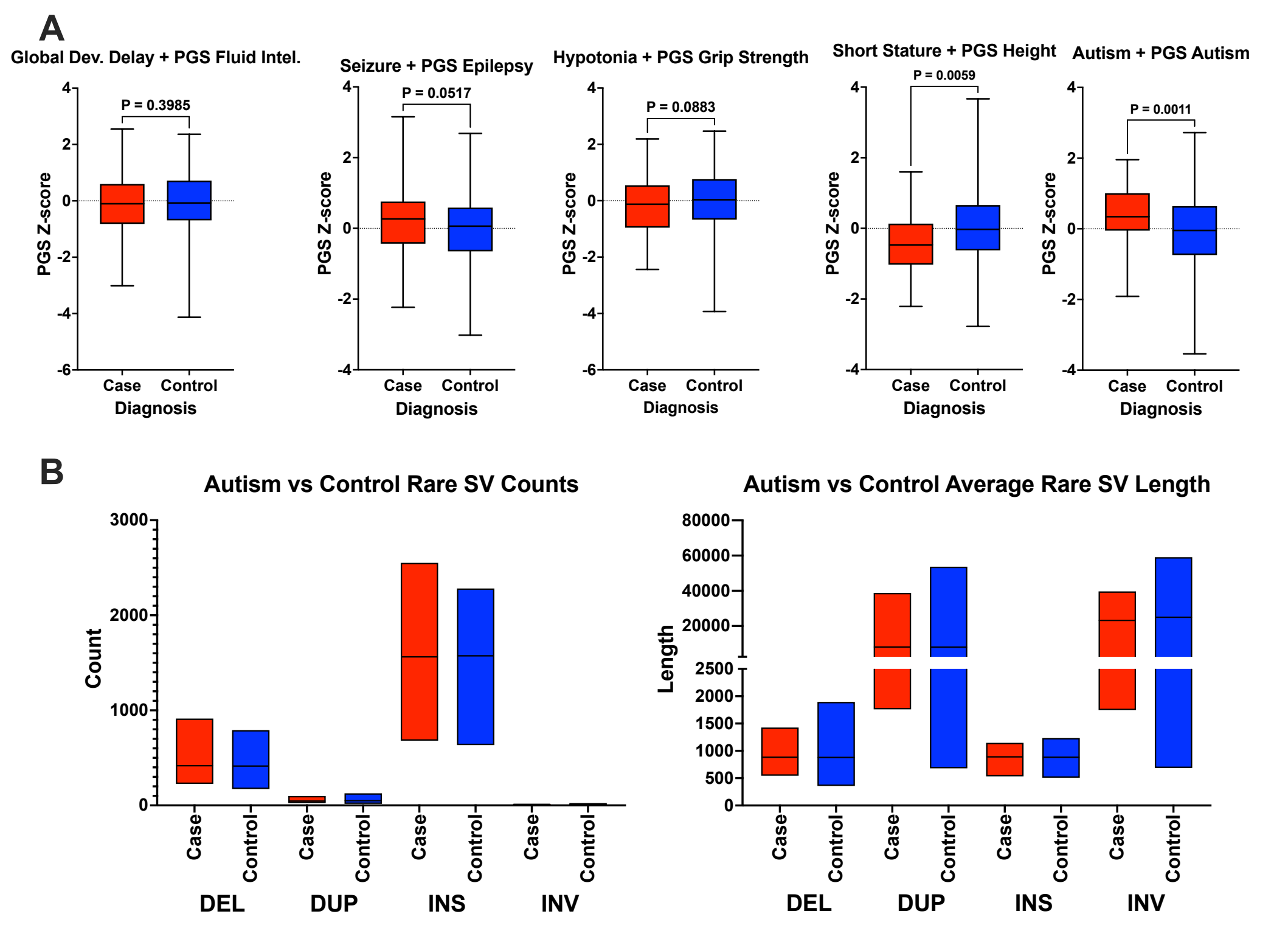
498 Hon, T., K. Mars, G. Young, Y. C. Tsai, J. W. Karalius, J. M. Landolin, N. Maurer, D. Kudrna, M.
499 A. Hardigan, C. C. Steiner, S. J. Knapp, D. Ware, B. Shapiro, P. Peluso, and D. R. Rank.
500 2020. 'Highly accurate long-read HiFi sequencing data for five complex genomes', *Sci*
501 *Data*, 7: 399.
502 Kainer, D., A. R. Templeton, E. T. Prates, D. Jacboson, E. R. O. Allan, S. Climer, and M. R.
503 Garvin. 2023. 'Structural variants identified using non-Mendelian inheritance patterns
504 advance the mechanistic understanding of autism spectrum disorder', *HGG Adv*, 4:
505 100150.
506 Kohler, S., M. Gargano, N. Matentzoglu, L. C. Carmody, D. Lewis-Smith, N. A. Vasilevsky, D.
507 Danis, G. Balagura, G. Baynam, A. M. Brower, T. J. Callahan, C. G. Chute, J. L. Est, P.
508 D. Galer, S. Ganesan, M. Griese, M. Haimel, J. Pazmandi, M. Hanauer, N. L. Harris, M.
509 J. Hartnett, M. Hastreiter, F. Hauck, Y. He, T. Jeske, H. Kearney, G. Kindle, C. Klein, K.
510 Knoflach, R. Krause, D. Lagorce, J. A. McMurry, J. A. Miller, M. C. Munoz-Torres, R. L.
511 Peters, C. K. Rapp, A. M. Rath, S. A. Rind, A. Z. Rosenberg, M. M. Segal, M. G. Seidel,
512 D. Smedley, T. Talmy, Y. Thomas, S. A. Wiafe, J. Xian, Z. Yuksel, I. Helbig, C. J.
513 Mungall, M. A. Haendel, and P. N. Robinson. 2021. 'The Human Phenotype Ontology in
514 2021', *Nucleic Acids Res*, 49: D1207-D17.
515 Kosugi, S., Y. Momozawa, X. Liu, C. Terao, M. Kubo, and Y. Kamatani. 2019. 'Comprehensive
516 evaluation of structural variation detection algorithms for whole genome sequencing',
517 *Genome Biol*, 20: 117.
518 Kucuk, E., Bpgh van der Sanden, L. O'Gorman, M. Kwint, R. Derks, A. M. Wenger, C. Lambert,
519 S. Chakraborty, P. Baybayan, W. J. Rowell, H. G. Brunner, Lelm Vissers, A. Hoischen,
520 and C. Gilissen. 2023. 'Comprehensive de novo mutation discovery with HiFi long-read
521 sequencing', *Genome Med*, 15: 34.
522 Lambert, S. A., L. Gil, S. Jupp, S. C. Ritchie, Y. Xu, A. Buniello, A. McMahon, G. Abraham, M.
523 Chapman, H. Parkinson, J. Danesh, J. A. L. MacArthur, and M. Inouye. 2021. 'The
524 Polygenic Score Catalog as an open database for reproducibility and systematic
525 evaluation', *Nat Genet*, 53: 420-25.
526 Marx, V. 2023. 'Method of the year: long-read sequencing', *Nat Methods*, 20: 6-11.
527 Merker, J. D., A. M. Wenger, T. Sneddon, M. Grove, Z. Zappala, L. Fresard, D. Waggott, S.
528 Utiramerur, Y. Hou, K. S. Smith, S. B. Montgomery, M. Wheeler, J. G. Buchan, C. C.
529 Lambert, K. S. Eng, L. Hickey, J. Korlach, J. Ford, and E. A. Ashley. 2018. 'Long-read
530 genome sequencing identifies causal structural variation in a Mendelian disease', *Genet*
531 *Med*, 20: 159-63.
532 Niemi, M. E. K., H. C. Martin, D. L. Rice, G. Gallone, S. Gordon, M. Kelemen, K. McAloney, J.
533 McRae, E. J. Radford, S. Yu, J. Gecz, N. G. Martin, C. F. Wright, D. R. Fitzpatrick, H. V.
534 Firth, M. E. Hurles, and J. C. Barrett. 2018. 'Common genetic variants contribute to risk
535 of rare severe neurodevelopmental disorders', *Nature*, 562: 268-71.
536 Pedersen, B. S., P. J. Bhetariya, J. Brown, S. N. Kravitz, G. Marth, R. L. Jensen, M. P. Bronner,
537 H. R. Underhill, and A. R. Quinlan. 2020. 'Somalier: rapid relatedness estimation for
538 cancer and germline studies using efficient genome sketches', *Genome Med*, 12: 62.
539 Pereanu, W., E. C. Larsen, I. Das, M. A. Estevez, A. A. Sarkar, S. Spring-Pearson, R. Kollu, S.
540 N. Basu, and S. Banerjee-Basu. 2018. 'AutDB: a platform to decode the genetic
541 architecture of autism', *Nucleic Acids Res*, 46: D1049-D54.
542 Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar,
543 P. I. de Bakker, M. J. Daly, and P. C. Sham. 2007. 'PLINK: a tool set for whole-genome
544 association and population-based linkage analyses', *Am J Hum Genet*, 81: 559-75.
545 Rosendahl, J., H. Kirsten, E. Hegyi, P. Kovacs, F. U. Weiss, H. Laumen, P. Lichtner, C. Ruffert,
546 J. M. Chen, E. Masson, S. Beer, C. Zimmer, K. Seltsam, H. Algul, F. Buhler, M. J. Bruno,
547 P. Bugert, R. Burkhardt, G. M. Cavestro, H. Cichoz-Lach, A. Farre, J. Frank, G.
548 Gambaro, S. Gimpfl, H. Grallert, H. Griesmann, R. Grutzmann, C. Hellerbrand, P. Hegyi,

M. Hollenbach, S. Iordache, G. Jurkowska, V. Keim, F. Kiefer, S. Krug, O. Landt, M. D. Leo, M. M. Lerch, P. Levy, M. Loffler, M. Lohr, M. Ludwig, M. Macek, N. Malats, E. Malecka-Panas, G. Malerba, K. Mann, J. Mayerle, S. Mohr, R. H. M. Te Morsche, M. Motyka, S. Mueller, T. Muller, M. M. Nothen, S. Pedrazzoli, S. P. Pereira, A. Peters, R. Pfutzer, F. X. Real, V. Rebours, M. Ridinger, M. Rietschel, E. Rosmann, A. Saftoiu, A. Schneider, H. U. Schulz, N. Soranzo, M. Soyka, P. Simon, J. Skipworth, F. Stickel, K. Strauch, M. Stumvoll, P. A. Testoni, A. Tonjes, L. Werner, J. Werner, N. Wodarz, M. Ziegler, A. Masamune, J. Mossner, C. Ferec, P. Michl, P. H. Drenth J, H. Witt, M. Scholz, M. Sahin-Toth, and A. C. P. all members of the PanEuropean Working group on. 2018. 'Genome-wide association study identifies inversion in the CTRB1-CTRB2 locus to modify risk for alcoholic and non-alcoholic chronic pancreatitis', *Gut*, 67: 1855-63.

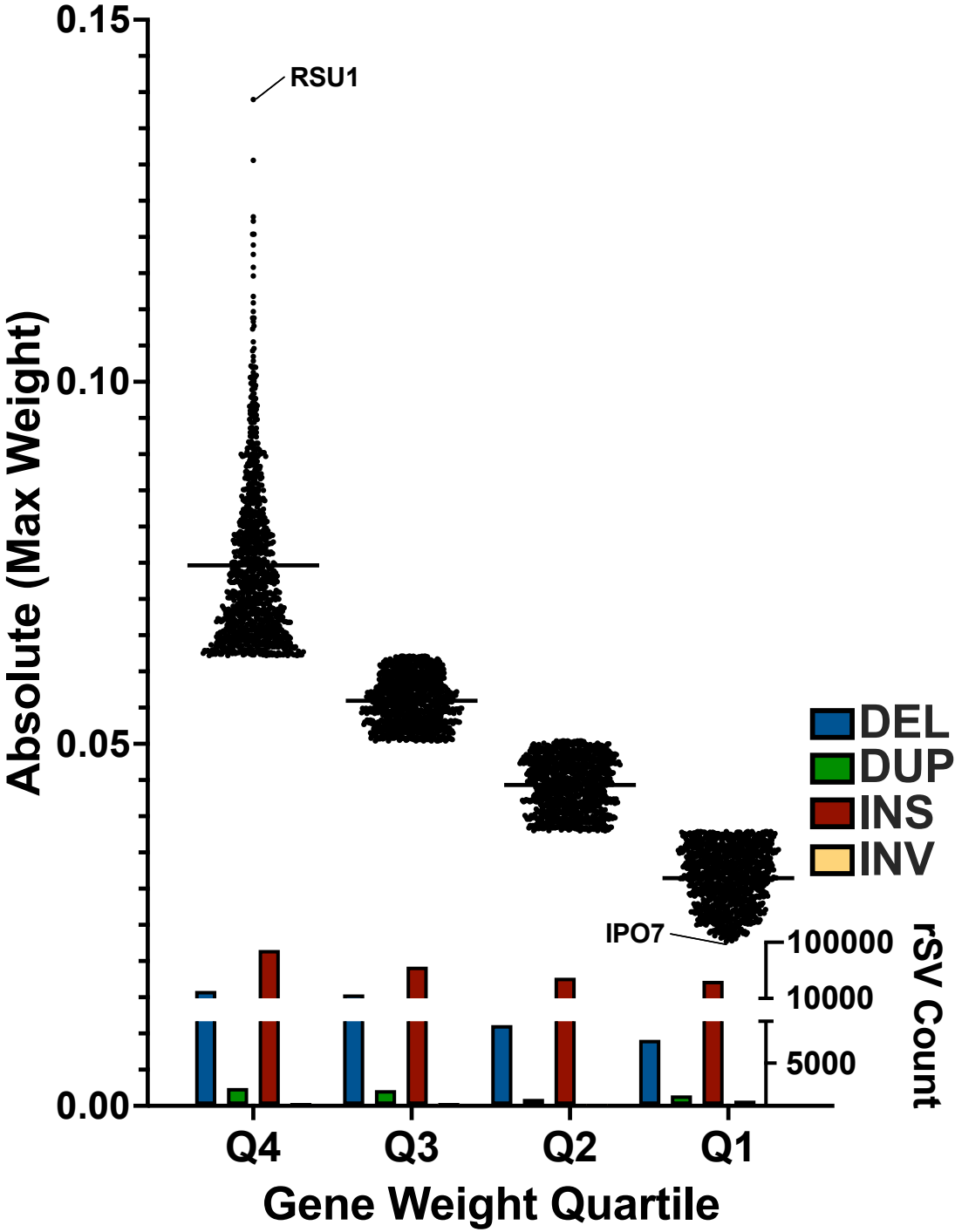
Simona, A., W. Song, D. W. Bates, and C. F. Samer. 2023. 'Polygenic risk scores in pharmacogenomics: opportunities and challenges-a mini review', *Front Genet*, 14: 1217049.

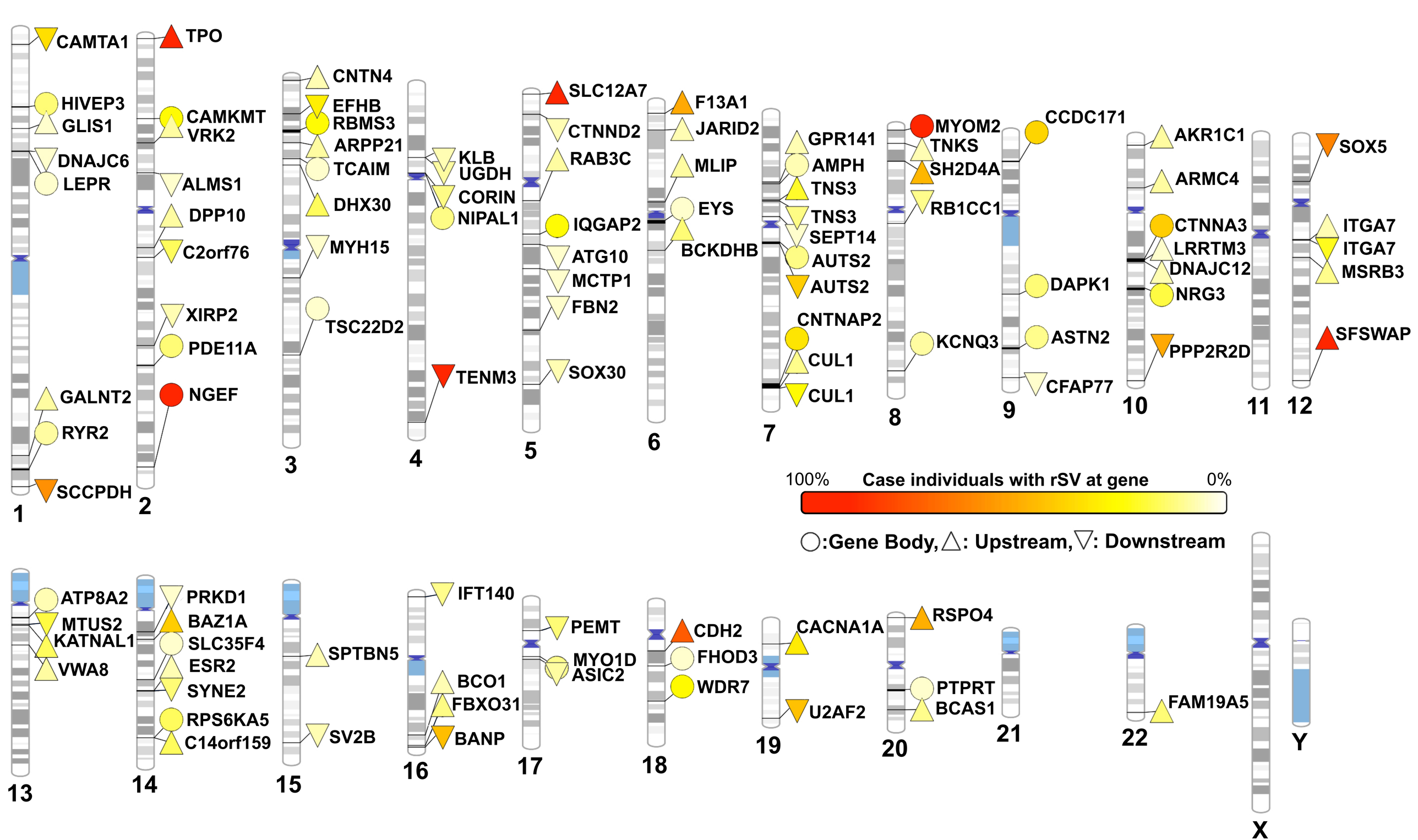
Stelzer, G., N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, S. Kaplan, D. Dahary, D. Warshawsky, Y. Guan-Golan, A. Kohn, N. Rappaport, M. Safran, and D. Lancet. 2016. 'The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses', *Curr Protoc Bioinformatics*, 54: 1 30 1-1 30 33.


A
Cohort Individual SV Counts
SV Count
15000
10000
5000
0
DEL
DUP
INS
INV
Filtering
Cohort Individual Rare SV Counts
SV Count
3000
2000
1000
0
DEL
DUP
INS
INV
B
Cohort Individual Average SV Length
INV
INS
DUP
DEL
50
500
1000
1500
10000
20000
30000
40000
50000
Average Length (bp)
Filtering
Cohort Individual Average Rare SV Length
INV
INS
DUP
DEL
50
500
1000
1500
10000
20000
30000
40000
50000
Average Length (bp)

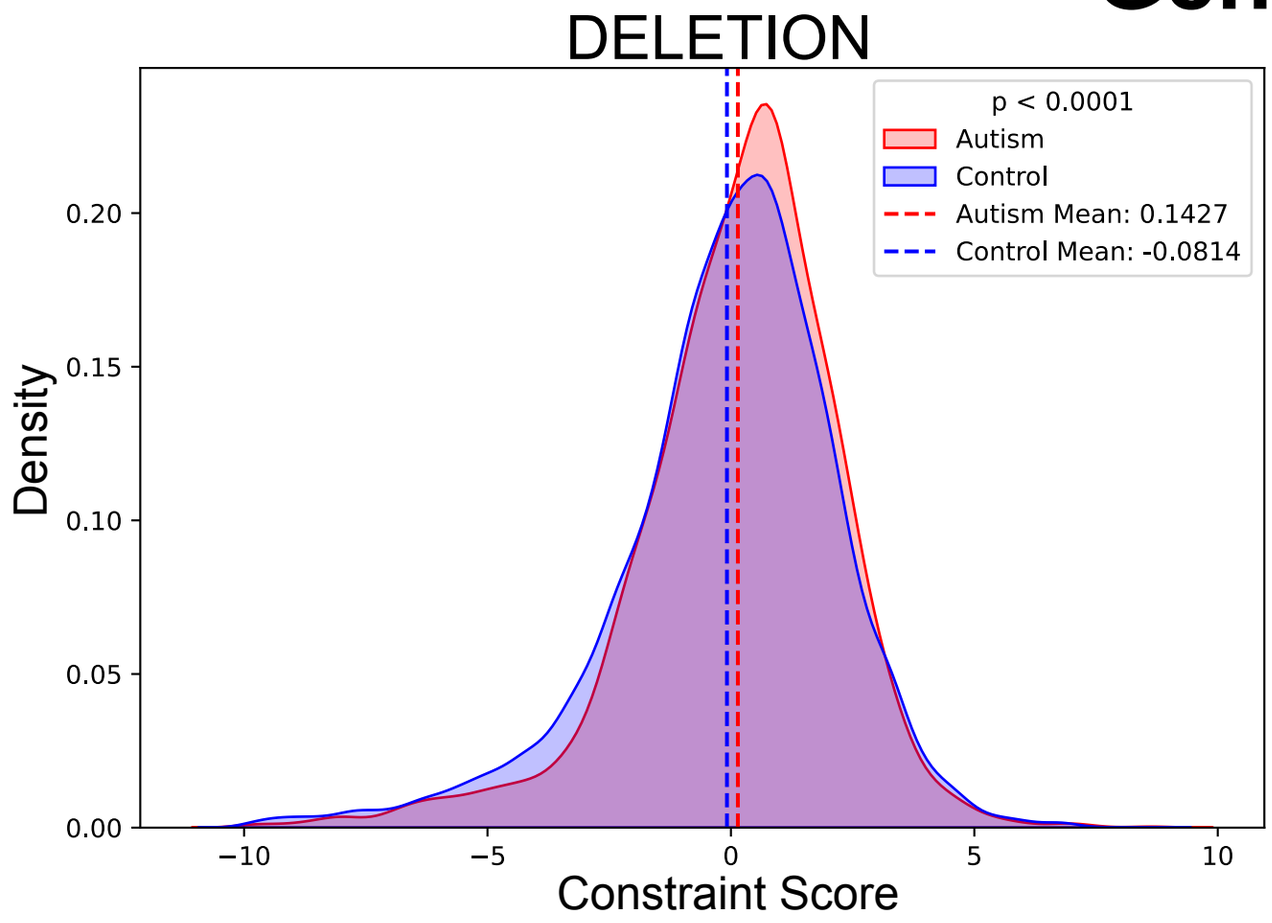

A
Global Dev. Delay + PGS Fluid Intel.
P = 0.3985
Seizure + PGS Epilepsy
P = 0.0517
Hypotonia + PGS Grip Strength
P = 0.0883
Short Stature + PGS Height
P = 0.0059
Autism + PGS Autism
P = 0.0011
PGS Z-score
4
2
0
-2
-4
-6
Case
Control
Diagnosis
B
Autism vs Control Rare SV Counts
Count
3000
2000
1000
0
Case
Control
DEL
DUP
INS
INV
Autism vs Control Average Rare SV Length
Length
80000
60000
40000
20000
2500
2000
1500
1000
500
0

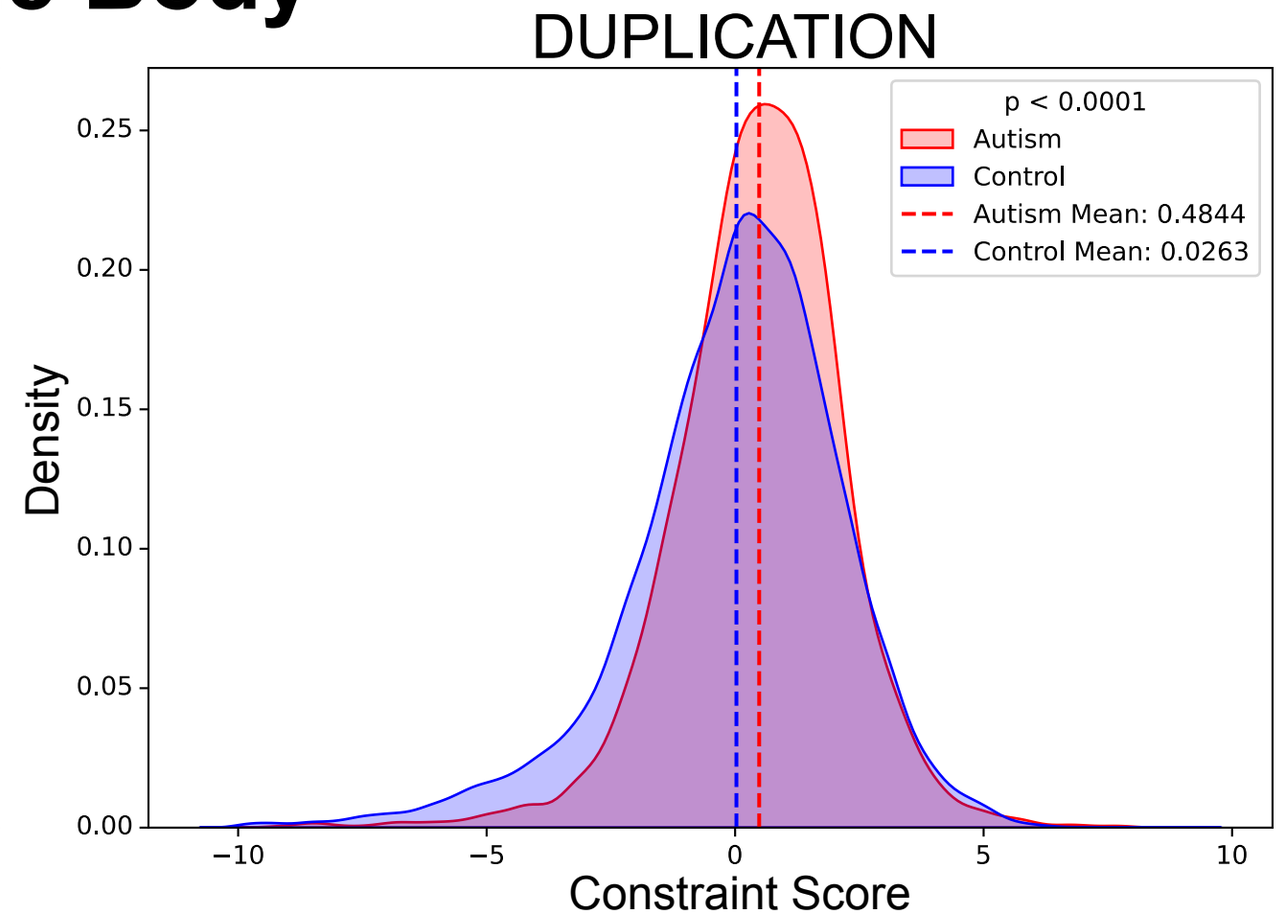
PGS GENES BY WEIGHT QUARTILES



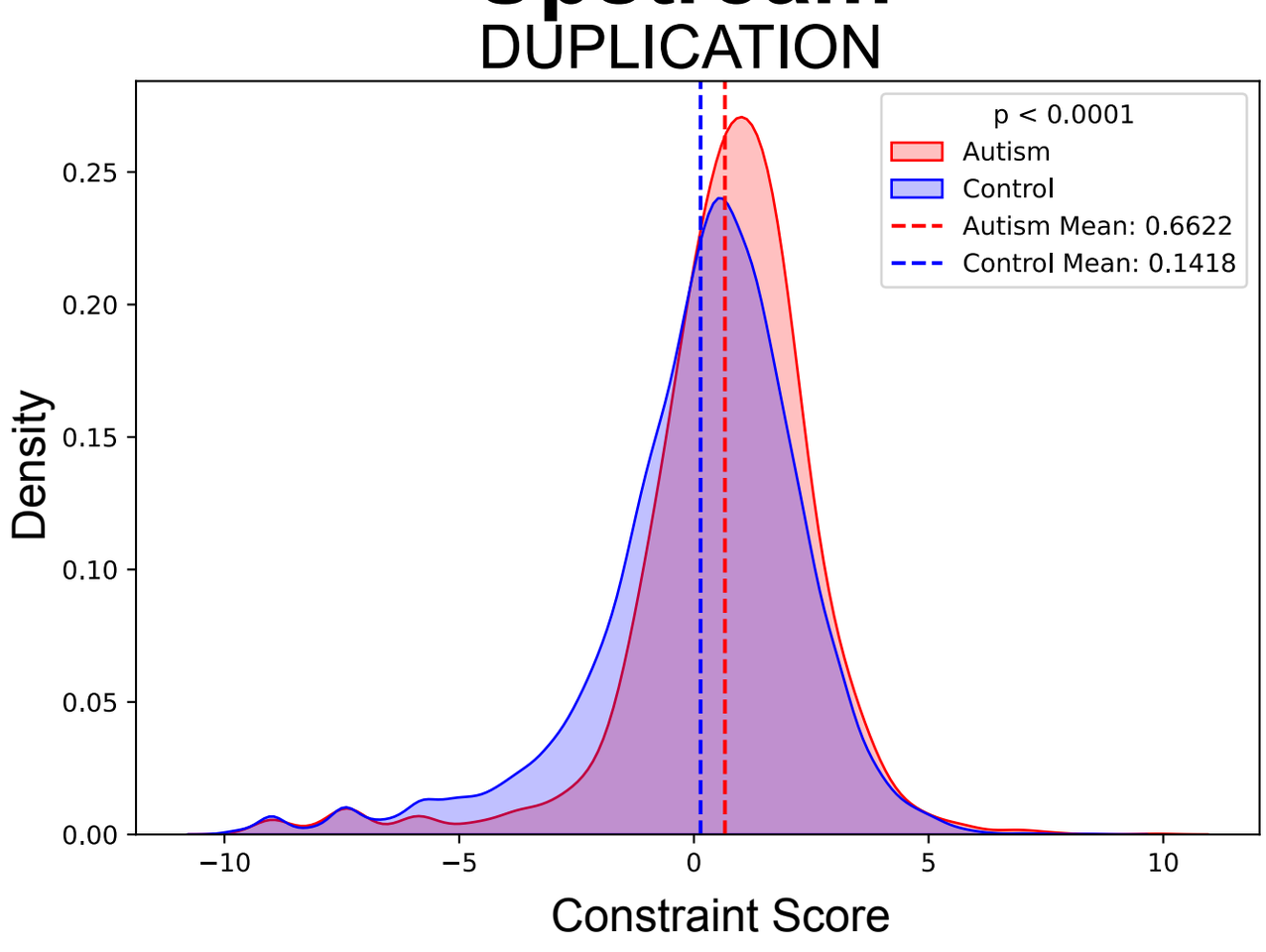

CAMTA1
HIVEP3
GLIS1
DNAJC6
LEPR
GALNT2
RYR2
SCCPDH
1
TPO
CAMKMT
VRK2
ALMS1
DPP10
C2orf76
XIRP2
PDE11A
NGEF
2
CNTN4
EFHB
RBMS3
ARPP21
TCAIM
DHX30
MYH15
TSC22D2
3
KLB
UGDH
CORIN
NIPAL1
TENM3
4
SLC12A7
CTNND2
RAB3C
IQGAP2
ATG10
MCTP1
FBN2
SOX30
5
F13A1
JARID2
MLIP
EYS
BCKDHB
6
GPR141
AMPH
TNS3
TNS3
SEPT14
AUTS2
AUTS2
CNTNAP2
CUL1
CUL1
7
MYOM2
TNKS
SH2D4A
RB1CC1
KCNQ3
8
CCDC171
DAPK1
ASTN2
CFAP77
9
AKR1C1
ARMC4
CTNNA3
LRRTM3
DNAJC12
NRG3
PPP2R2D
10
11
SOX5
ITGA7
ITGA7
MSRB3
SFSWAP
12
100%
Case individuals with rSV at gene
0%
○:Gene Body, △: Upstream, ▽: Downstream
ATP8A2
MTUS2
KATNAL1
VWA8
13
PRKD1
BAZ1A
SLC35F4
ESR2
SYNE2
RPS6KA5
C14orf159
14
SPTBN5
SV2B
15
IFT140
BCO1
FBXO31
BANP
16
PEMT
MYO1D
ASIC2
17
CDH2
FHOD3
WDR7
18
CACNA1A
U2AF2
19
RSPO4
PTPRT
BCAS1
20
21
22
FAM19A5
X
Y

Gene Body

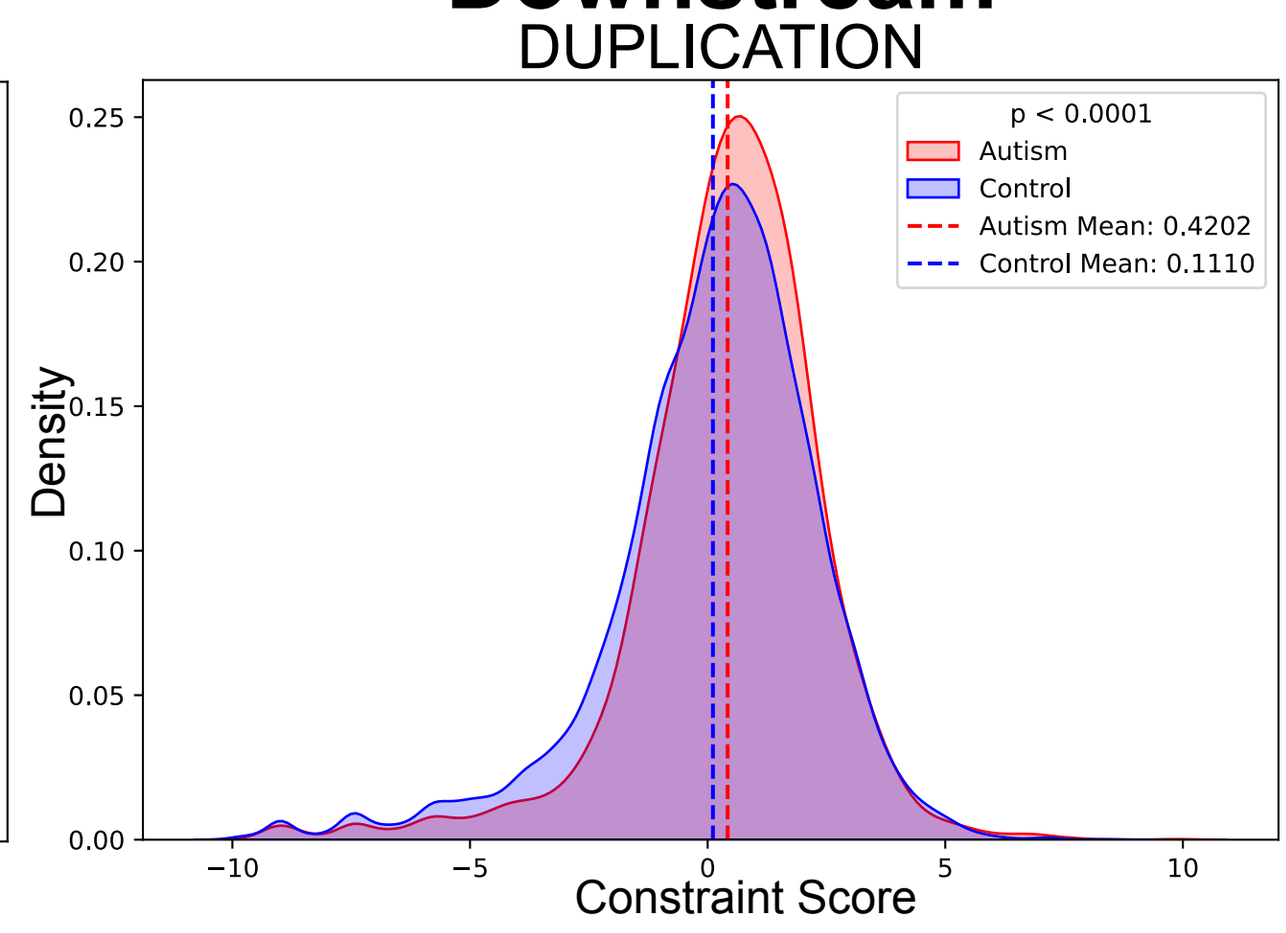


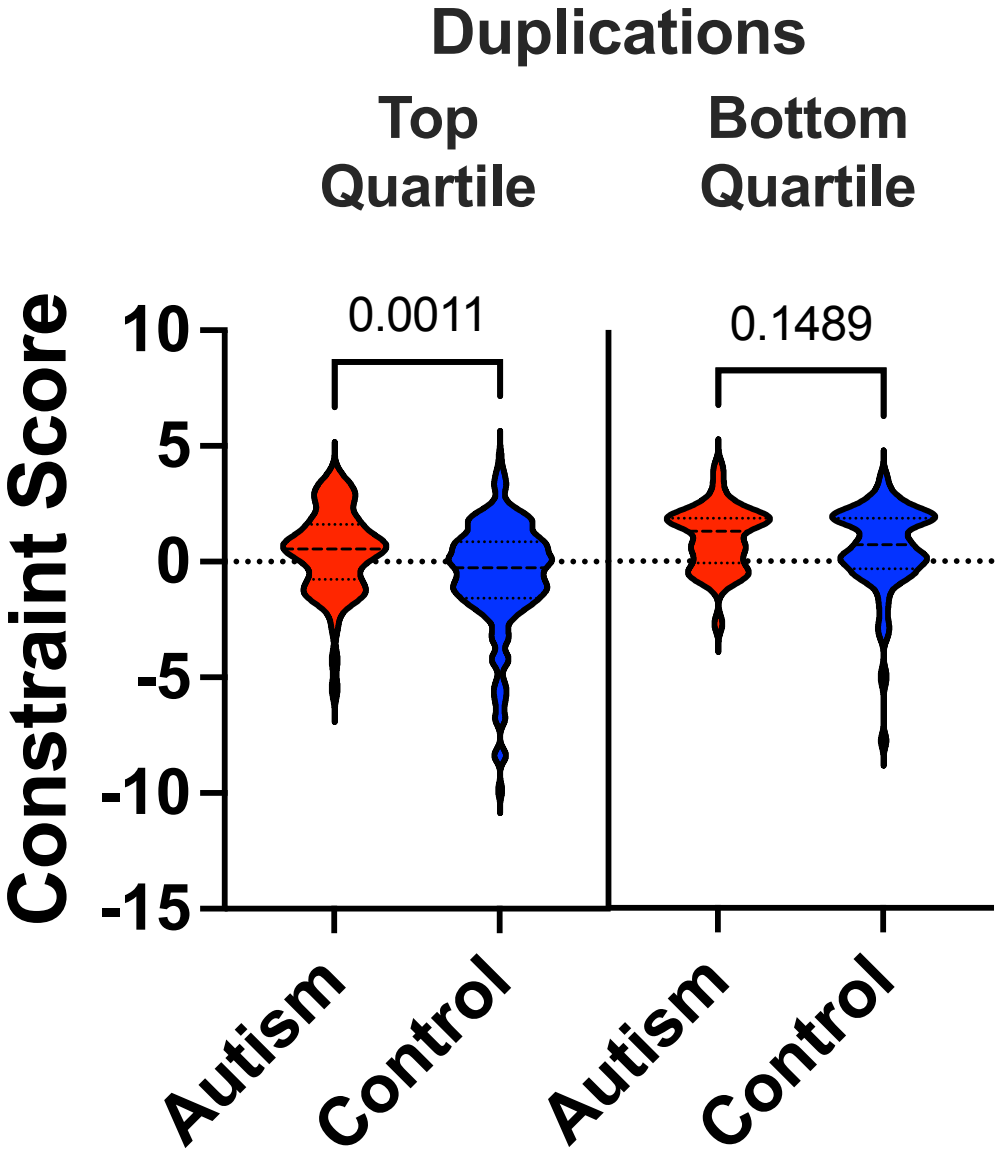


Upstream



Downstream




Duplications
Top Quartile
Bottom Quartile
Constraint Score
10
5
0
-5
-10
-15
0.0011
0.1489
Autism
Control
Autism
Control