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

C. LeMaster[1], C. Schwendinger-Schreck[1], B. Ge[2], W. Cheung[1], J. J. Johnston[1], T. Pastinen[1], C. 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 1,827 deletions, 158 duplications, 619 insertions, and 14 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 ($P = 2x10^{-04}$). This difference was not observed in the lowest-ranked gene set ($P = 0.18$). 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 al. 2018), and autism spectrum disorder (Gaugler et al. 2014). Through integrating PGS, we can gain enhanced resolution into genes with impacts on rare diseases.

In this study, we leveraged PacBio long-read HiFi genomes available in the Genomic Answers for Kids (GA4K) study at Children's Mercy Research Institute to investigate the role of rare structural variants (rSVs) in autism spectrum disorder. The GA4K study is a large-scale, phenotypically diverse pediatric rare disease cohort of over 13,000 individuals (Cohen et al. 2022). At the time of our study, there were 497 rare disease probands with HiFi long read whole genome sequences (lrWGS) completed. Further integrating open-source PGS available from PGS Catalog (Lambert et al. 2021), we computed individual PGS liability for autism spectrum disorder and additional traits and phenotypes. Our study focused on rare deletions, duplications, insertions, and inversions overlapping trait-relevant PGS genes. We further conducted a genomic constraint analysis of rSVs, providing insights on their evolutionary tolerance and potential pathogenicity.

Overall, we found an enrichment of deletions, duplications and insertions at trait-relevant genes and a significantly higher constraint of duplications in individuals with autism. Furthermore, we trace SV-overlapped genes back to autism and implicate anterograde trans-synaptic signaling using ontology enrichment. Our study utilizes the intersection of rSVs, trait-relevant PGS, and genomic constraint to explore the structural variant landscape in a large-scale rare disease cohort, with a focus on genes associated with autism spectrum disorder.

RESULTS

Structural variant type and frequency show prevalence of deletions and insertions

We analyzed 497 probands with HiFi long read whole genome sequences (lrWGS) in the Children's Mercy GA4K database (N = 254 male; N = 243 female). We assessed ancestry using Somalier, which predicted 437 European, 58 Admixed American, and 2 Asian individuals. Structural variants (SVs) were called using PacBio SV (PBSV) after alignment with human genome GrCh38 (Pedersen et al. 2020). All individuals were assessed for total number of SVs and their inter-individual frequency (**Supplemental Table 1**). We developed a python application called *GA4K-SV-FINDER* where researchers can query these SVs by inputting user-specified genomic regions or overlapping genes (**Data Availability**).

To identify potential large-effect rare structural variants (rSVs) impacting rare disease phenotypes we removed common SVs from the dataset (5% MAF; **Methods**). This reduced a total of 11,595,521 SVs to 251,767 rSVs (**Figure 1A**). Additionally, the total number of probands was reduced from 497 to 483, due to some individuals possessing only common SVs after filtering.

Representing the lowest frequency of rSV type, inversions (INV, 1.6%) had the largest average length (27,959 bp) when compared to deletions (DEL, 65.7%, 986 bp), duplications (DUP, 7.9%, 7,917 bp), and insertions (INS, 24.9%, 580 bp) (**Figure 1A, 1B**). When comparing SV type across sexes, frequencies were similar and within 5% variation between males and females for length and count. The frequency and length of rSVs in our study closely correspond with population frequencies seen in other studies (Kosugi et al. 2019; Guo et al. 2021).

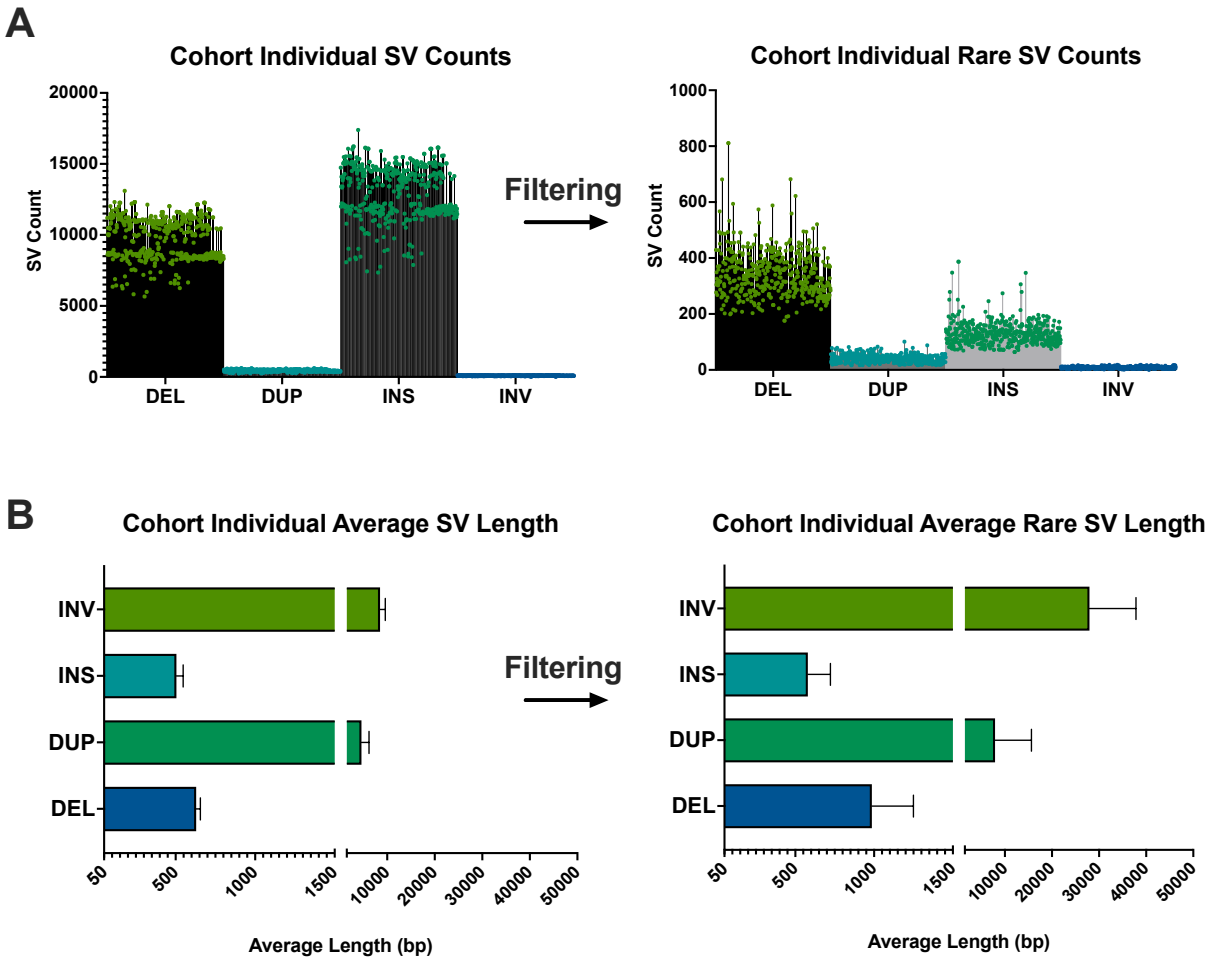


Figure 1: SV type frequency and average length in study population. (**A**) Two dot-bar graphs representing SV type (Deletions (DEL), duplications (DUP), insertions (INS), inversions (INV) counts before and after filtering for rare variants. (**B**) Two horizontal bar graphs with outlier whiskers displaying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

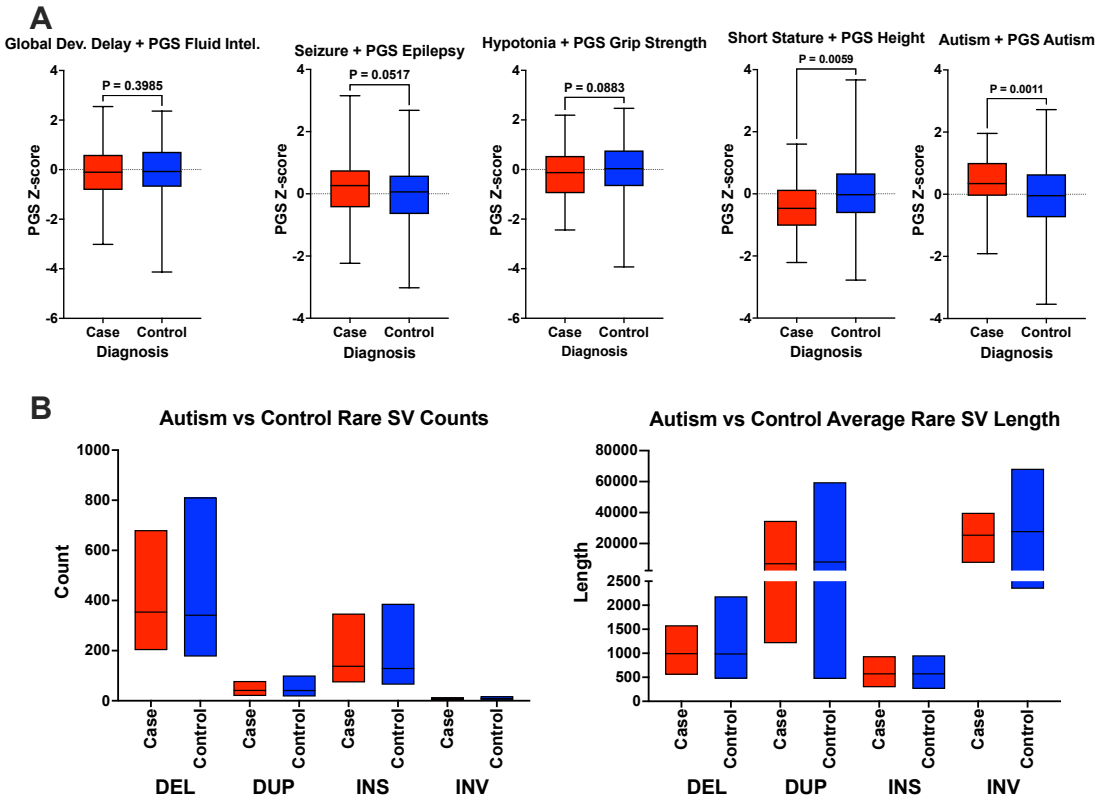


Figure 2: Box plots show significant mappings for Autism and Short Stature. (**A**) Box and whisker plots showing significance of trait-relevant PGS mapped to five rare disease phenotypes in GA4K. Red indicates case individuals. Blue indicates control individuals. PGS Z-score is indicated on the Y axis (N cohort = 497, N GDD cases = 138, N Seizure = 86, N Hypotonia = 62, N Autism = 54, N Short Stature = 45). (**B**) Box plots showing the count of rSV type (DEL, DUP, INS, INV) and rSV type length across individuals for cases (N = 52) and controls (N = 431).

Focusing on the larger case number of the two significant phenotypes, we then compared rSVs in the Autism group (N cases = 52) group with controls. We observed that probands with Autism had slightly higher median counts, aside from duplications, and shorter average lengths, except for DELs, compared with controls (**Figure 2B**).

High weight PGS genes reveal rSV enrichment in autism

Due to the complex genetic nature of Autistic disorders, we decided to further focus on individuals in our cohort with Autism. We evaluated genes harboring variants included in a PGS for Autism (PGS Catalog ID: PGS000327). Each variant is weighted according to its relative association with a phenotype. We intersected the variant coordinates with the coordinates of genes. This revealed 5,830 genes intersected by at least one overlapping PGS variant. After intersecting cohort rSVs with PGS genes, we found at least one overlap in 1,893

PGS genes (**Supplemental Table 2**). Of those genes, Ras Suppressor Protein 1 (RSU1) had the highest weight (gene weight = 0.139). The lowest weighted gene was Importin 7 (IPO7) (gene weight = 0.023) (**Supplemental Table 2; Figure 3**). Additionally, there are 257 genes associated with the Human Phenotype Ontology (HPO) ID for autism spectrum disorder (HP:0000717) (Kohler et al. 2021); of those genes, 42 were found to have overlap with rSVs (average gene weight = 0.06).

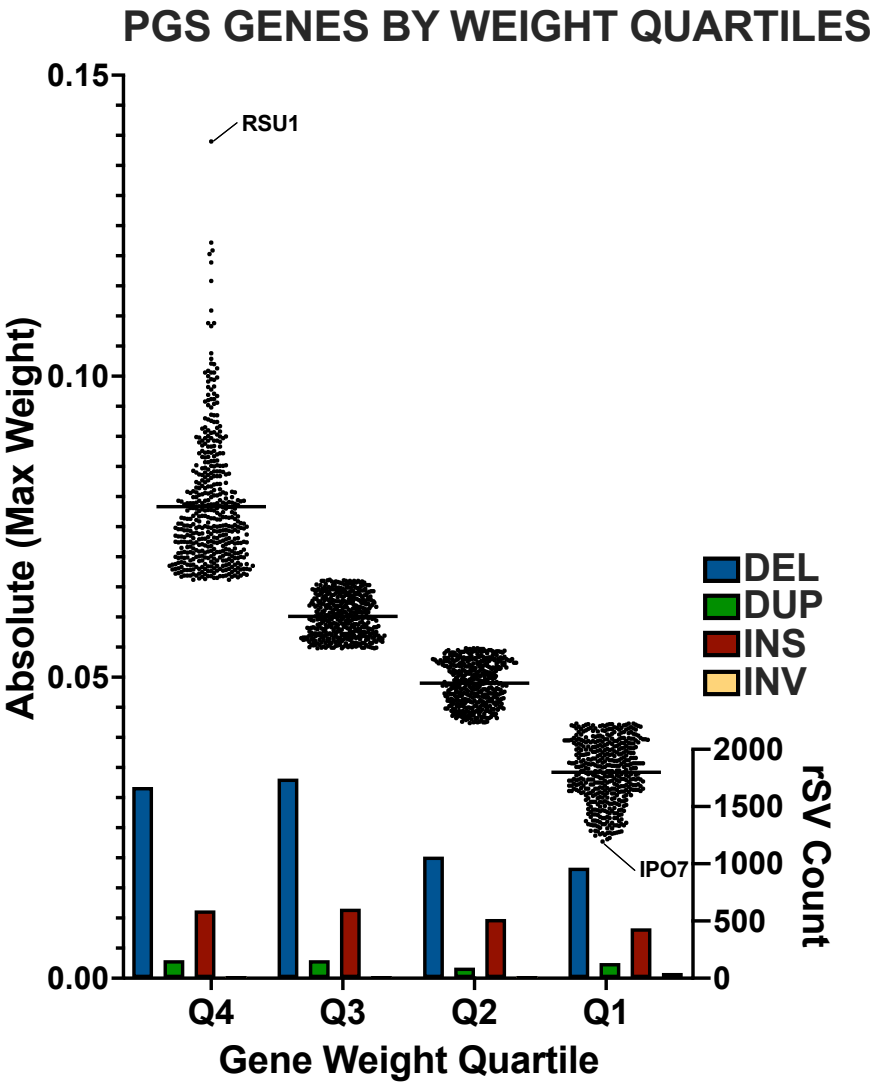


Figure 3: Top quartile PGS genes show a wide range of weights and an enrichment of deletions. Dot plot of genes within their quartiles based on the intersecting PGS gene weights. The highest weighted quartile is Q4 on the left. The lowest weighted genes are in Q1, to the right. The y-axis for gene weights 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.

130 We then organized these 1,893 genes into 4 quartiles according to their PGS gene weight (**Figure 3**). Each
131 quartile contained 471-474 genes (**Figure 3**). All 1,893 genes were evaluated for frequency of rSVs both within
132 the gene body and 200kb up- and down-stream of the gene body (**Supplemental Table 2**). Focusing on the top
133 quartile, we observed 4,648 rSV overlaps. Deletions were the most abundant (69%), both in the gene body (39%)
134 and in the flanking regions (upstream: 16%, downstream: 14%). They were followed by INS (24%; body: 13%,
135 up: 6%, down: 5%), DUP (6%; body: 3%, up: 1%, down: 1%), and INV (0.7%; body: 0.3%, up: 0.3%, down:
136 0.1%). DEL frequency of the gene body paralleled a descending gene weight across all quartiles, where lower
137 gene weights correlated with a lower number of DELs (**Figure 3**). A similar trend was reflected across the other
138 SV types. When evaluating the up- and down-stream 200kb flanking regions for each gene, a higher down-
139 stream rSV frequency was associated with a lower gene body frequency, where more rSVs in a quartile gene
140 body reflected a lower number in the down-stream region. This pattern was not seen in the up-stream region.
141
142 We then provide a cut-off for genes in the top quartile with significant frequency difference (>5% diff, Chi-Square
143 test, p-value < 0.05) between case-control groups (**Table 1; Methods**). In doing this we move from entire cohort
144 frequencies to a case-control enrichment. While we define significant frequency differences between groups
145 where genes have >5% enrichment over the other group, we wanted to focus on resolving higher impact rSV-
146 gene enrichments between groups (>10%) in the top quartile for an in-depth analysis (**Table 1**). Thus, Genes
147 Solute Carrier Family 12 Member 7 (SLC12A7), WD Repeat-Containing Protein 7 (WDR7), RNA Binding Motif
148 Single Stranded Interacting Protein 3 (RBMS3), KN motif and ankyrin repeat domain-containing protein 1
149 (KANK1), Amphiphysin (AMPH), Astrotactin-2 (ASTN2), and Catenin alpha-2 (CTNNA2) were more enriched in
150 cases. While DELs were the largest contributor to enrichment in SLC12A7 (diff: 10%), GBRB3 (diff: 10%), and
151 RBMS3 (diff: 10%), duplications were enriched in WDR7 (diff: 22%), and INS were mostly present in Cadherin
152 4 (CDH, diff: 16%) and Diacylglycerol Kinase Beta (DGKB, diff: 14%) (**Table 1**). No INVs were found to be
153 differentially enriched (case > 10% different). The genes KANK1, ASTN2, CTNNA2, RBMS3, and CDH4 have
154 been previously reported in connection Autism (Bauleo et al. 2021; Nassir et al. 2021; Mitra et al. 2016; Yin et
155 al. 2020; Grissa et al. 2022). When assessing rSVs in the control group, Aryl Hydrocarbon Receptor Repressor
156 (AHRR, diff: 18%), Genetic Suppressor Element 1 (GSE1, diff: 16%), SH3 And Multiple Ankyrin Repeat Domains
157 2 (SHANK2, diff: 11%), Teneurin Transmembrane Protein 2 (TENM2, diff: 10%), Sidekick Cell Adhesion

158 Molecule 1 (SDK1, diff: 10%), and Roundabout Guidance Receptor 2 (ROBO2, diff: 10%) showed high
159 enrichment. DELs were enriched in controls at ATPase Phospholipid Transporting 10A (ATP10A, diff: 16%) and
160 Catenin Delta 2 (CTNND2, diff: 12%), where INSs had enrichment at DLGAP2 (diff: 18%).

Upstream	Differential Freq. (>Controls)	Gene Body	Differential Freq. (>Controls)	Downstream	Differential Freq. (>Controls)
RFTN1	8%	SLC12A7	18%*	AHRR	23%*
VRK2	8%	WDR7	14%*	POR	9%
MCC	7%	RBMS3	12%*	ZFPM2	9%
DLG2	5%	KANK1	11%*	ADAMTS19	7%
RSU1	5%	AMPH	11%*	PCDH7	6%
NKAIN2	5%	ASTN2	10%*	CNTNAP5	5%
MCPH1	5%	CTNNA2	10%*	CSMD2	5%
PHLPP1	5%	GABRB3	9%		
PKIB	5%	PTPRM	8%		
		SPON1	8%		
		NRCAM	7%		
		ACOT11	7%		
		MTFR1	5%		

161

162 **Table 1: SV enriched genes strongly associated with Autism.** Genes with significantly (Chi-Square, p-value
163 < 0.05) higher frequency rSV enrichment (>5%) in case individual gene bodies are displayed for each region.
164 The group frequency differential is displayed to the right of the gene. *Indicates high impact (>10%) rSV-gene
165 enrichment in cases.
166

167 We also assessed for differential enrichments in the 200kb flanking regions of genes. This revealed an overall
168 enrichment down-stream of AHRR (diff: 23%) in cases (**Table 1**). A high enrichment in the control group was
169 seen upstream and downstream of Intraflagellar Transport Protein 140 (IFT140, diff: 16%) and E1A Binding
170 Protein P400 (EP400, diff: 17%), respectively. While we focused on differential enrichment between groups, it is
171 also important to note that rSVs in the gene body of SLC12A7 were found in 92% of case individuals (**Figure 4**).
172 For comparison, autism-associated gene ASTN2 had rSV overlap in 13% of case individuals (**Figure 4)**.

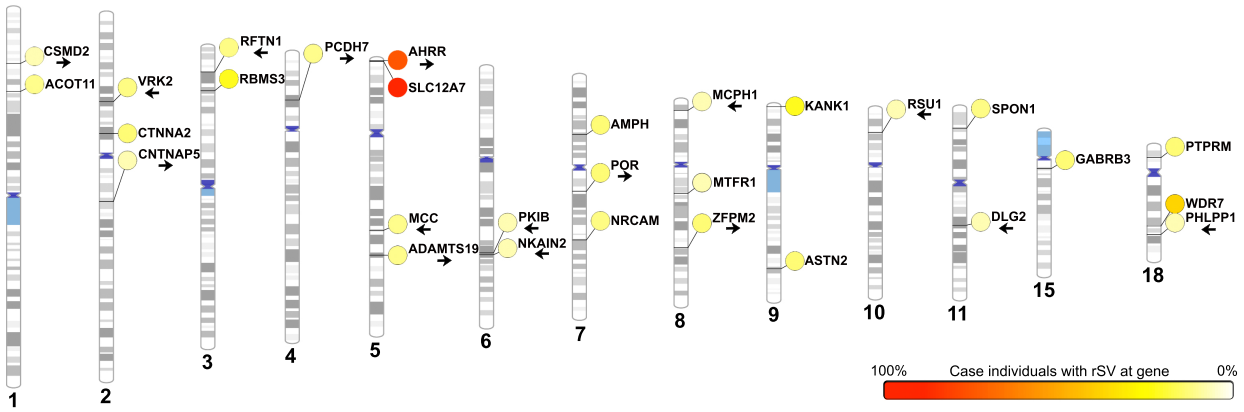


Figure 4: SLC12A7 shows rSV enrichment in the gene body and upstream of the gene body, also present in a large majority of cases. PhenoGram showing the frequency of individuals with rSVs at the cut-off genes by chromosomal position (color scale from 0-100% frequency of case individuals). Cytogenic Giemsa bands show GC- (blue) and AT-rich (gray) regions.

We found that there were a high number of rSVs in and around top quartile genes strongly associated with Autism. Conversely, there was also an enrichment of DELs at SLC12A7 and duplications at WDR7, two genes previously unassociated with Autism (Abrahams et al. 2013).

Some genes offered a more complex consideration for rSV enrichment, with only controls possessing DELs among autism-associated gene CTNND2 (Turner et al. 2015). Furthermore, the case DEL-enriched gene body of SLC12A7 (CHR5:1061326-1096876, avg length= 1,286 bp) is positioned on the same chromosome but downstream of AHRR (CHR5:304176-438291), which was also enriched in the 200kb downstream region by cases (DEL: CHR5-554357-633297, avg length = 1,255bp) but by controls in the gene body itself (CHR5:307170-429323, avg length = 466 bp) (**Figure 4**). As such, while significant differences in frequency were seen across groups, there was a high intra-group frequency of rSVs in the gene body of SLC12A7 and downstream AHRR, two genes positioned relatively close to one another on chromosome 5, but without overlapping coordinates (**Figure 4**). This suggests that rSVs may co-localize around trait-relevant genes.

SV enriched genes in the top quartile significantly associate with autism and potentially associated pathways

Ideally, all genes implicated by PGS variants for autism would have an association with the phenotype, but many genes do not. Furthermore, we only chose genes that were implicated by rSVs and, more specifically, had an

198 rSV enrichment profile different from controls. Thus, after deriving a list of genes with high enrichment of rSVs
199 in cases (N genes = 29) (**Table1**), we wanted to unbiasedly determine if these genes specifically enrich the
200 autism phenotype. We used the DisGeNET database with Enrichr and found a significant enrichment (adj. p-
201 value = 0.01) with Autistic Disorder using our gene list (**Table 1**). Furthermore, using the GO Biological Processes
202 database, we found that rSV enriched genes GABRB3, DLG2, AMPH, and SLC12A7 are involved in anterograde
203 trans-synaptic signaling (GO:0098916) (adj. p-value = 0.03) (**Table 1**). Our results here provide a framework for
204 linking increased rSV burden in cases to potential downstream impacts on gene pathways relevant to Autism
205 pathogenesis.
206
207 ***Duplications are enriched at regions of high constraint in top PGS genes***
208 We next used 1kb gnomAD constraint windows to evaluate potential pathogenicity amongst rSVs (**Methods**). In
209 short, higher constraint indicates an evolutionary intolerance to variation within that loci. Though constraint
210 scores were not available across every 1kb region of the genome, we were able to intersect 23% of rSVs. We
211 then graphed the density of rSV constraint scores across these rSVs and performed a Mann-Whitney test to
212 determine significance across groups (**Figure 5**). This revealed that rare DELs and DUPs in the case group
213 tended to overlap higher constraint windows compared with controls (DEL: $p < 0.0001$, case z mean = 0.16,
214 control z mean = -0.19; DUP: $p < 0.0001$, case z mean = 0.62, control z mean = 0.06). These differences were
215 not observed for INSs and INVs. Furthermore, case DUPs tended to overlap higher constraint windows within
216 200kb down-stream of PGS genes ($p = 0.0014$, case z mean = 0.23, control z mean = 0.14). Conversely, DUPs
217 in controls within 200kb up-stream of PGS genes tended to overlap higher constraint windows compared with
218 cases ($p = 0.02$, case z mean = -0.08, control z mean = -0.08), but with low constraint overlap across both groups
219 on average.

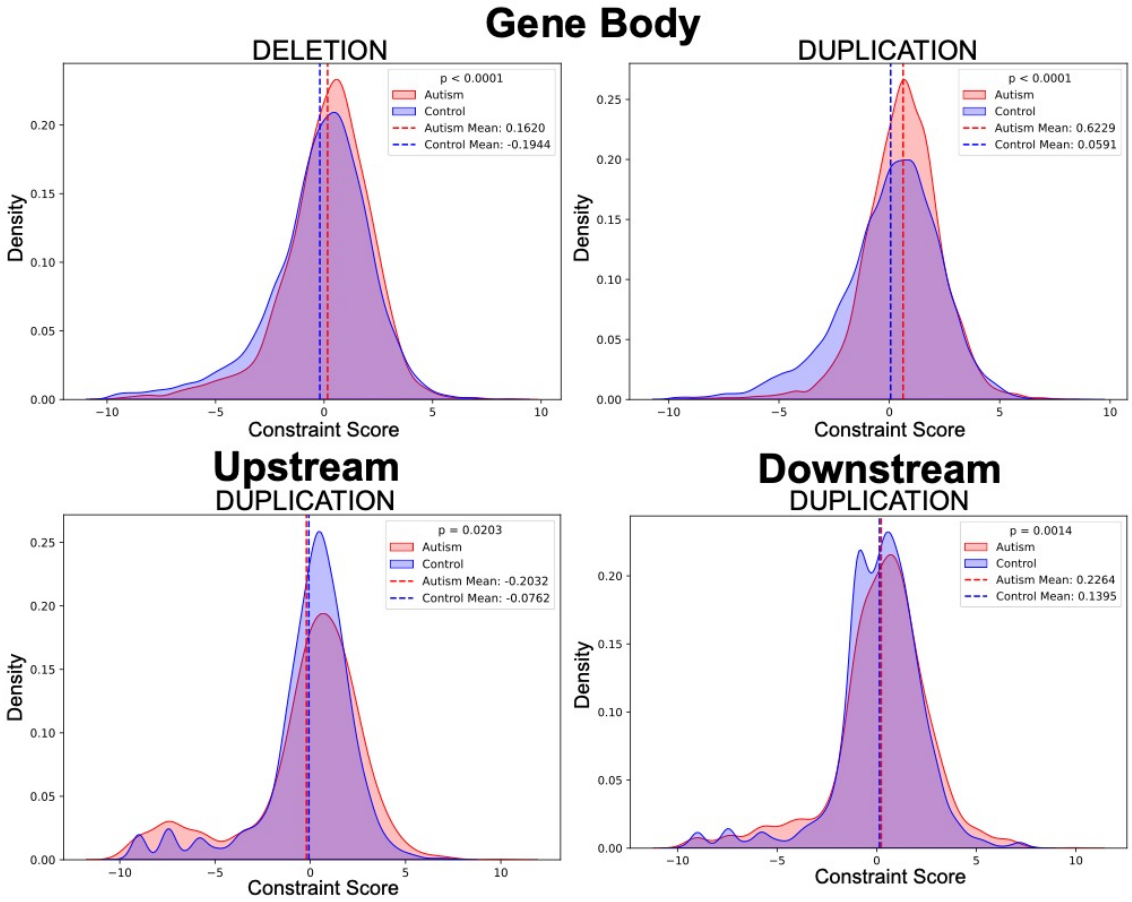


Figure 5: Duplication constraints are significantly different between cases and controls inside and outside of the gene body. Four density graphs representing SV types with significantly different constraint scores across case-control groups as well as gene regions. The top two graphs show significantly different constraints across groups for the gene body. The bottom two for 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 cases compared to controls (p = 0.0002, case z mean = 0.49, control z mean = -0.53) (**Figure 6**). This was not reflected in the bottom PGS gene quartile (p = 0.18 case z mean = 0.92, control z mean = 0.38) (**Figure 6**) or across other SV types. Additionally, 67% of cases had a DUP with a constraint score in a top quartile gene and 31% had a constraint score greater than 1; in comparison; 59% of controls had a DUP with a constraint score and only 16% had a score greater than 1.

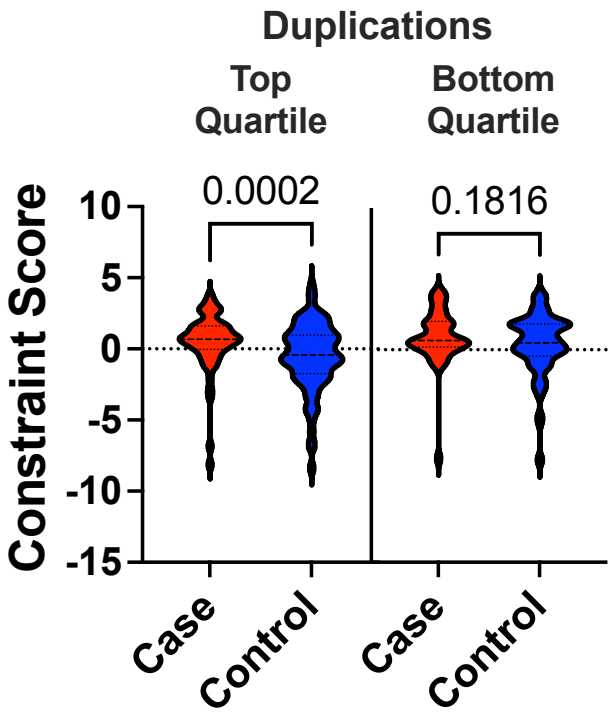


Figure 6: Constraint scores for duplications for top PGS quartile genes is significantly different across cases and controls. Violin plot of case and control duplication constraint scores within top and bottom quartile genes.

CONCLUSION

Our study highlights the significance of rare structural variants (rSVs) in complex rare diseases such as Autism. By utilizing long-read HiFi sequencing, we have enhanced the detection and characterization of rSVs in the GA4K cohort, particularly deletions and duplications associated with autism-related genes. Our analysis indicates a higher incidence of deletions in genes linked to Autism, with duplications occurring frequently in regions of 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.

A large-scale lrWGS dataset such as GA4K enables insights into structural variation not previously explored. This analysis identified specific genes with significant rSV intersections not previously explored in the context of Autism, such as SLC12A7 and WDR7, differentiating between case and control frequencies. Considering that variants in SLC genes are strongly associated with Autism phenotypes, there may be a correlation between deletions at SLC12A7 and phenotype presentation (Mir et al. 2022; Hu et al. 2020). Additionally evaluating rSVs

251 in genes with investigative association to Autism, we further resolved rSV enriched genes in case individuals as
252 autistic-disorder- and anterograde-trans-synaptic-signaling-specific. The online database of Autism spectrum
253 disorder genes by the Simons Foundation Autism Research Initiative (SFARI) lists 1,162 genes related to the
254 disorder through SNVs (Abrahams et al. 2013). In our study, we implicated over 5,000 genes using PGS
255 associations and resolved 1,893 of those genes with rSV overlaps in our cohort. Of those 1,162 SFARI genes,
256 295 were in our list. Sixty of the 295 were considered syndromic by SFARI metrics. That leaves 1,598 genes
257 absent from SFARI but with PGS and rSV case overlaps identified in our cohort. Because long-read mapping-
258 based methods for SV calling are more accurate than short-read approaches, due to their ability to provide more
259 accurate alignments by capturing entire SV alleles in a single fragment – additionally considering that HiFi lrWGS
260 is relatively new – our study underscores the importance in adopting lrWGS in rare disease where no single
261 variant is implicated.
262
263 In future studies, there should be consideration into the complexities of rare disease phenotypes and any
264 potential comorbidities. Controlling for phenotype, age, gender, and family might strengthen gene-rSV
265 associations. While we would prefer a true control group of healthy non-diseased individuals, we hope that any
266 significant determinations we've made are further strengthened by these considerations. Our findings underscore
267 the complexity of the genetic factors at play in Autism and reinforce the importance of comprehensive analysis
268 required in uncovering the nuances of rare disease variants. These insights lay the groundwork for more precise
269 genetic diagnostics and the development of targeted interventions involving rSVs.
270
271 **METHODS**
272 **Study Cohort**
273 The study cohort described includes 497 affected probands across 419 families with 49% females and 51%
274 males. Subjects were enrolled in the Genomics Answers for Kids (GA4K) program. Probands at age of enrollment
275 ranged from 0 to 57 years (median = 8) (older individuals were typically ascertained as follow-up from an affected
276 family member). Subjects were eligible for the study if they had a suspected genetic diagnosis based on clinical
277 presentation and/or an existing molecular or cytogenic findings. Providers introduce the study and ask if the
278 family is interested in the study. The study complies with all relevant ethical regulations as approved by the

Children's Mercy Institutional Review Board (IRB) (Study #11120514). Informed or written consent was obtained from all participants prior to study inclusion. Participants were not compensated for participation.

SV Calling

PacBio HiFi Sequencing results from GA4K proband subjects were processed through PacBio's structural variant (PBSV, Version: 2.6.2) calling workflows after alignment to human reference GRCh38. Variants achieved a PASS score if they had at least 1 read per strand, were at least 1000 bp from the end of a contig, and away (< 1000 bp) from gaps (run of >= 50 Ns) in the reference assembly. Additionally, for PBSV, we used the following parameters: -ccs (Circular Consensus Sequence/HiFi), -m 20 (SVs larger than 20 bp were used), -A 3 (3 supporting reads were required to call a variant as homozygous alternate), -O 3 (3 supporting reads were required to call a variant as heterozygous), -P 20 (minimum phred) were used for structural variant calling. Due to known false-positives, CNVs were not included in variant analysis. Additionally, only reads greater than 50 bp in length were used for each variant type. ChrX and ChrY variants were excluded from analysis and, for all analyses of rare SVs, SVs analyzed had a minor allele frequency (MAF) < 5% or count =< 4 in the study cohort. Somalier was used to predict ancestry (> 0.5 probability).

GA4K-SV-FINDER

We created a python-based application for querying allele frequencies and genes associated with SVs in our cohort. This application can be found, along with a readme and the raw data file, in our GitHub repository available at https://github.com/smail-lab-cmh/ga4k-sv-finder.

Polygenic Risk Scores

Polygenic risk scores were calculated for each specified polygenic score using open-source PGS available in the PGS catalog (www.pgscatalog.org). Associated PGS variant coordinates were downloaded from their respective catalog entries. PGS variants were converted to hg38 coordinates using dbSNP26 (version 155) where necessary. PGS variants were restricted to autosomes only and variants mapping to the HLA region were removed. Individual-level PGS were calculated using PLINK (version 1.9) (Purcell et al. 2007) ("sum" flag) on all variants available in the GA4K imputed genotype callset with R2>=0.8 ("exclude-if-info" flag). PGS scores were

307 converted to Z-scores within each PGS in the full GA4K EUR ancestry cohort (proband and other available family
308 members, N = 7,436). Individuals with an extreme outlier PGS Z-score (abs(PGS Z-score)>=10) in one or more
309 PGS were removed from the final PGS dataset.
310
311 **Intersect**
312 To create compatible chromosomal coordinate files for structural and PGS variants, pseudo-BED files were first
313 created for each PGS SNP coordinate file, containing the chromosome, coordinates, base change, and
314 associated weight within the PGS as a z-score. This file was first intersected, using the bedtools intersect
315 command (v2.31.0), with a gene position BED file (gencode v26 hg38) to determine the gene the SNP was
316 associated with. After determining the genes associated with the PGS, they were then intersected with pseudo-
317 BED files generated from each pbsv VCF from the cohort. This created intersect files that provide the
318 chromosomal coordinates for the SV and the genes, which included sv type, length, weight, and overlap length.
319
320 **Flanking Regions**
321 Both up- and down-stream intersects were performed by first using BEDtools Slop command on gene start and
322 end coordinates separately. For example, start coordinates are transformed into a 200kb upstream value,
323 expanding the gene start coordinate within the boundaries of its respective chromosome. When intersecting
324 these expanded genes with SVs, any SV found to enrich the gene body that is also matching an SV found in the
325 new flanking region is removed from the flanking region analysis, as it was already intersecting the gene body.
326 In doing this, we separated SVs found up- or down-stream of the gene from the SVs overlapping the gene body
327 itself.
328
329 **Gene Weighting**
330 Gene coordinates were retrieved from the gencode v26 hg38 basic annotation file (gencodegenes.org). Gene
331 coordinates were intersected by weighted PGS variant coordinates. For each gene, PGS variant weights were
332 transformed to absolute values before taking the maximum value across all intersecting variants. In turn, this
333 provided an absolute maximum weight per gene. Following gene weighting, gene coordinates were intersected
334 by rSVs.

335

336 **Constraint Scoring**

337 Each rSV was intersected with the gnomAD v3 QCed genomic constraint by 1kb regions file. This file scores 1kb
338 regions throughout the genome with a constraint z-score value. Each of the proband rSVs were intersected with
339 these 1kb constraint windows and the scores are averaged across windows for each SV, excluding non-coding
340 regions or any undefined region lacking a constraint value. This produces a single score for each SV intersecting
341 a gene. The same method applies to up- and down-stream SV constraint evaluations.

342

343 **Statistical and Enrichment Analysis**

344 Graphing and statistics were generated using Graph Pad Prism 9 and Python (SciPy-Stats & Pandas). For
345 enrichment, SV frequencies between case and control groups were determined and Chi-Square tests were used
346 to determine (p-value < 0.05) significant differences. When SV counts were low for a gene (<5), a Fisher's Exact
347 test was used. These tests were performed for body, upstream, and downstream of genes separately.

348

349 **Gene Analysis**

350 Gene analysis was conducted using the DISEASES database, AutDB, GeneCards Version 5.19, and Enrichr
351 (Grissa et al. 2022; Basu, Kollu, and Banerjee-Basu 2009; Stelzer et al. 2016; Chen et al. 2013).

352

353 **ADDITIONAL INFORMATION**

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

356

357 **DATA AVAILABILITY**

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

361

362

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.; 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.

Basu, S. N., R. Kollu, and S. Banerjee-Basu. 2009. 'AutDB: a gene reference resource for autism research', *Nucleic Acids Res*, 37: D832-6.

Bauleo, A., A. Montesanto, V. Pace, R. Brando, L. De Stefano, D. Puntorieri, L. Cento, S. Loddo, C. Calacci, A. Novelli, and E. Falcone. 2021. 'Rare copy number variants in ASTN2 gene in patients with neurodevelopmental disorders', *Psychiatr Genet*, 31: 239-45.

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.

Hon, T., K. Mars, G. Young, Y. C. Tsai, J. W. Karalius, J. M. Landolin, N. Maurer, D. Kudrna, M. A. Hardigan, C. C. Steiner, S. J. Knapp, D. Ware, B. Shapiro, P. Peluso, and D. R. Rank. 2020. 'Highly accurate long-read HiFi sequencing data for five complex genomes', *Sci Data*, 7: 399.

Hu, C., L. Tao, X. Cao, and L. Chen. 2020. 'The solute carrier transporters and the brain: Physiological and pharmacological implications', *Asian J Pharm Sci*, 15: 131-44.

Kainer, D., A. R. Templeton, E. T. Prates, D. Jacboson, E. R. O. Allan, S. Climer, and M. R. Garvin. 2023. 'Structural variants identified using non-Mendelian inheritance patterns advance the mechanistic understanding of autism spectrum disorder', *HGG Adv*, 4: 100150.

Kohler, S., M. Gargano, N. Matentzoglu, L. C. Carmody, D. Lewis-Smith, N. A. Vasilevsky, D. Danis, G. Balagura, G. Baynam, A. M. Brower, T. J. Callahan, C. G. Chute, J. L. Est, P. D. Galer, S. Ganesan, M. Griese, M. Haimel, J. Pazmandi, M. Hanauer, N. L. Harris, M. J. Hartnett, M. Hastreiter, F. Hauck, Y. He, T. Jeske, H. Kearney, G. Kindle, C. Klein, K. Knoflach, R. Krause, D. Lagorce, J. A. McMurry, J. A. Miller, M. C. Munoz-Torres, R. L. Peters, C. K. Rapp, A. M. Rath, S. A. Rind, A. Z. Rosenberg, M. M. Segal, M. G. Seidel, D. Smedley, T. Talmy, Y. Thomas, S. A. Wiafe, J. Xian, Z. Yuksel, I. Helbig, C. J. Mungall, M. A. Haendel, and P. N. Robinson. 2021. 'The Human Phenotype Ontology in 2021', *Nucleic Acids Res*, 49: D1207-D17.

Kosugi, S., Y. Momozawa, X. Liu, C. Terao, M. Kubo, and Y. Kamatani. 2019. 'Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing', *Genome Biol*, 20: 117.

Kucuk, E., Bpgh van der Sanden, L. O'Gorman, M. Kwint, R. Derks, A. M. Wenger, C. Lambert, S. Chakraborty, P. Baybayan, W. J. Rowell, H. G. Brunner, Lelm Vissers, A. Hoischen, and C. Gilissen. 2023. 'Comprehensive de novo mutation discovery with HiFi long-read sequencing', *Genome Med*, 15: 34.

Lambert, S. A., L. Gil, S. Jupp, S. C. Ritchie, Y. Xu, A. Buniello, A. McMahon, G. Abraham, M. Chapman, H. Parkinson, J. Danesh, J. A. L. MacArthur, and M. Inouye. 2021. 'The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation', *Nat Genet*, 53: 420-25.

Marx, V. 2023. 'Method of the year: long-read sequencing', *Nat Methods*, 20: 6-11.

Merker, J. D., A. M. Wenger, T. Sneddon, M. Grove, Z. Zappala, L. Fresard, D. Waggott, S. Utiramerur, Y. Hou, K. S. Smith, S. B. Montgomery, M. Wheeler, J. G. Buchan, C. C. Lambert, K. S. Eng, L. Hickey, J. Korlach, J. Ford, and E. A. Ashley. 2018. 'Long-read genome sequencing identifies causal structural variation in a Mendelian disease', *Genet Med*, 20: 159-63.

Mir, A., M. Almudhry, F. Alghamdi, R. Albaradie, M. Ibrahim, F. Aldurayhim, A. Alhedaithy, M. Alamr, M. Bawazir, S. Mohammad, S. Abdelhay, S. Bashir, and Y. Housawi. 2022. 'SLC gene mutations and pediatric neurological disorders: diverse clinical phenotypes in a Saudi Arabian population', *Hum Genet*, 141: 81-99.

Mitra, I., K. Tsang, C. Ladd-Acosta, L. A. Croen, K. A. Aldinger, R. L. Hendren, M. Traglia, A. Lavillaureix, N. Zaitlen, M. C. Oldham, P. Levitt, S. Nelson, D. G. Amaral, I. Hertz-Picciotto, M. D. Fallin, and L. A. Weiss. 2016. 'Pleiotropic Mechanisms Indicated for Sex Differences in Autism', *PLoS Genet*, 12: e1006425.

Nassir, N., A. Bankapur, B. Samara, A. Ali, A. Ahmed, I. M. Inuwa, M. Zarrei, S. A. Safizadeh Shabestari, A. AlBanna, J. L. Howe, B. K. Berdiev, S. W. Scherer, M. Woodbury-Smith, and M. Uddin. 2021. 'Single-cell transcriptome identifies molecular subtype of autism spectrum disorder impacted by de novo loss-of-function variants regulating glial cells', *Hum Genomics*, 15: 68.

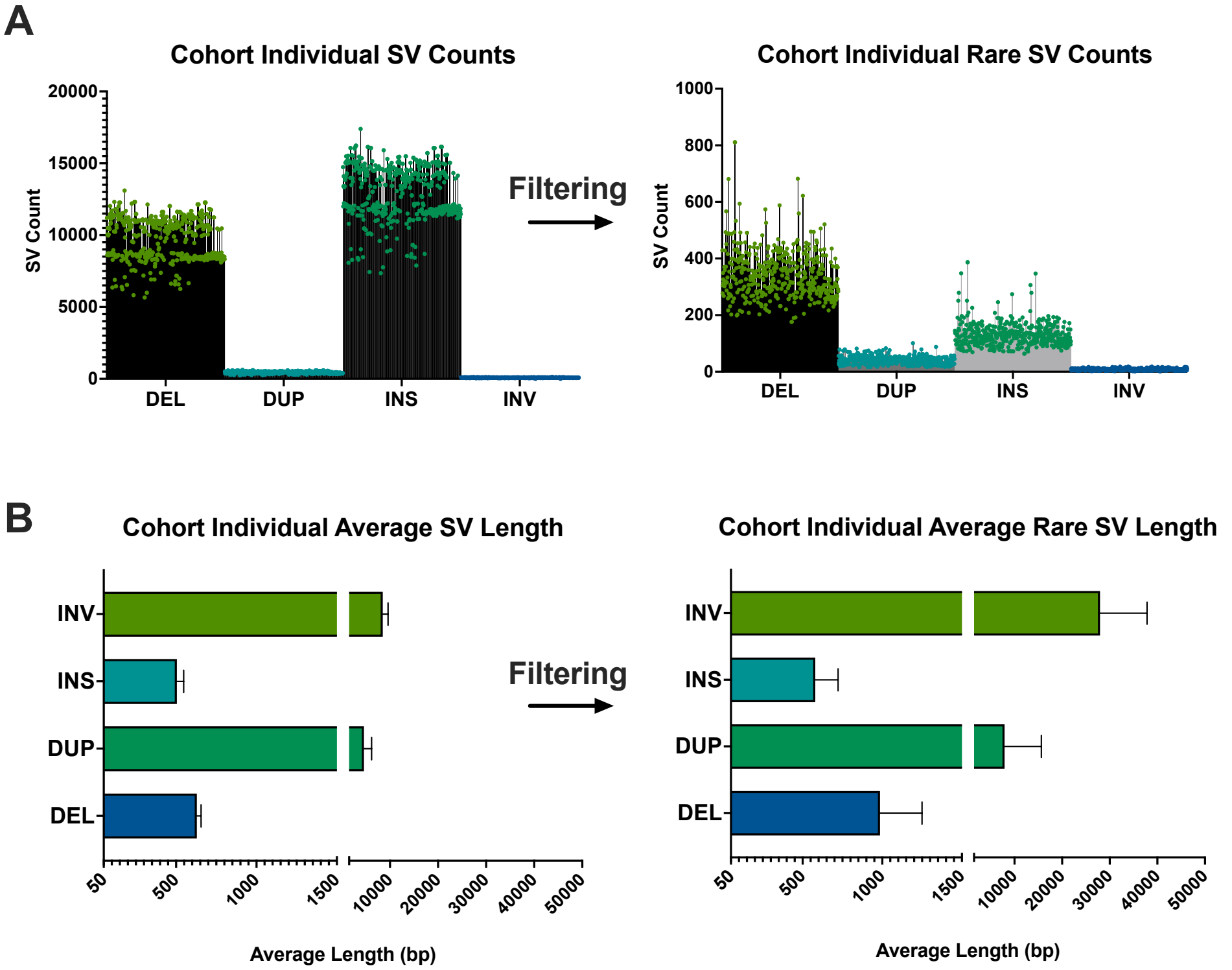
Niemi, M. E. K., H. C. Martin, D. L. Rice, G. Gallone, S. Gordon, M. Kelemen, K. McAloney, J. McRae, E. J. Radford, S. Yu, J. Gecz, N. G. Martin, C. F. Wright, D. R. Fitzpatrick, H. V. Firth, M. E. Hurles, and J. C. Barrett. 2018. 'Common genetic variants contribute to risk of rare severe neurodevelopmental disorders', *Nature*, 562: 268-71.

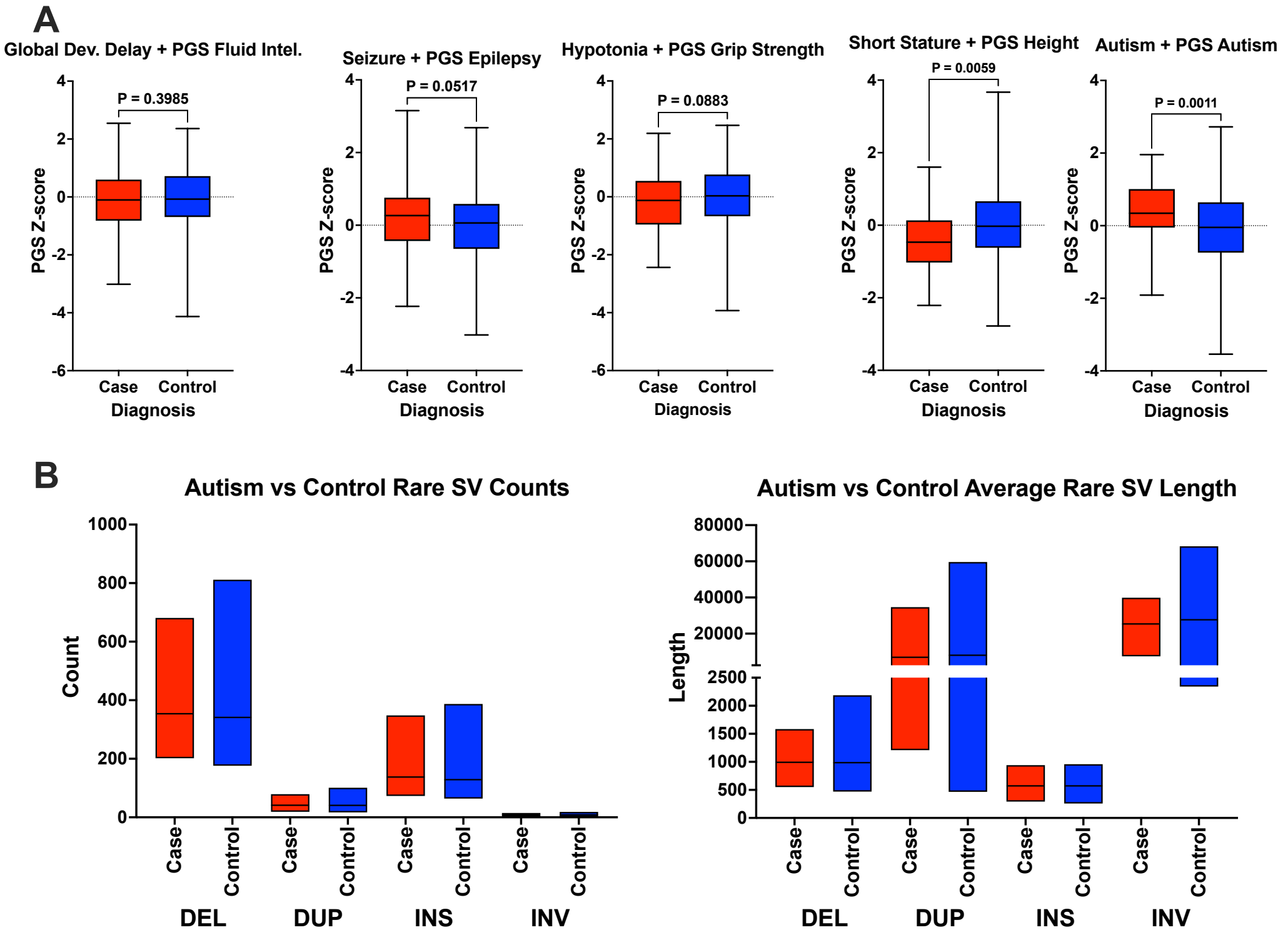
Pedersen, B. S., P. J. Bhetariya, J. Brown, S. N. Kravitz, G. Marth, R. L. Jensen, M. P. Bronner, H. R. Underhill, and A. R. Quinlan. 2020. 'Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches', *Genome Med*, 12: 62.

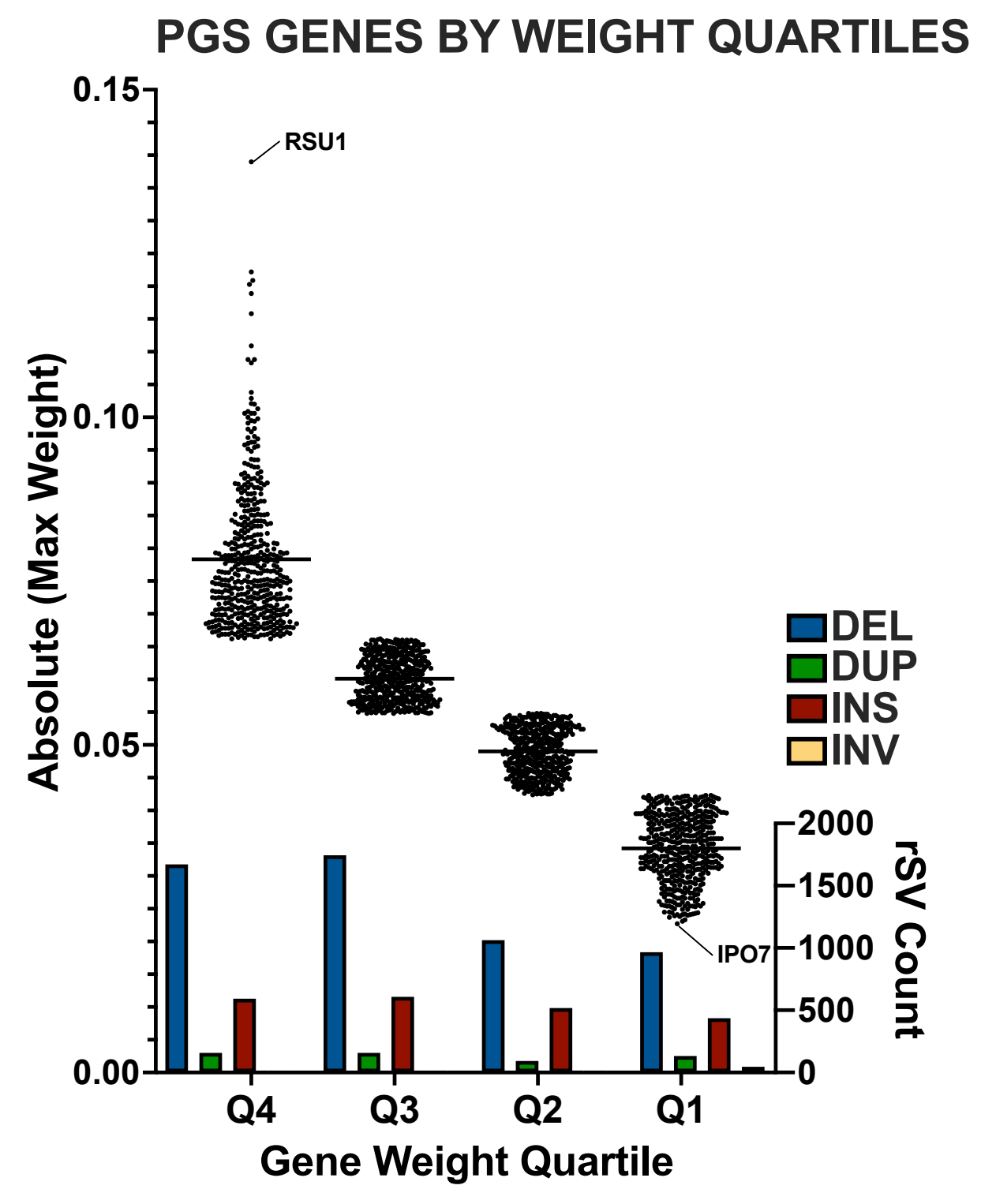
Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. 2007. 'PLINK: a tool set for whole-genome association and population-based linkage analyses', *Am J Hum Genet*, 81: 559-75.

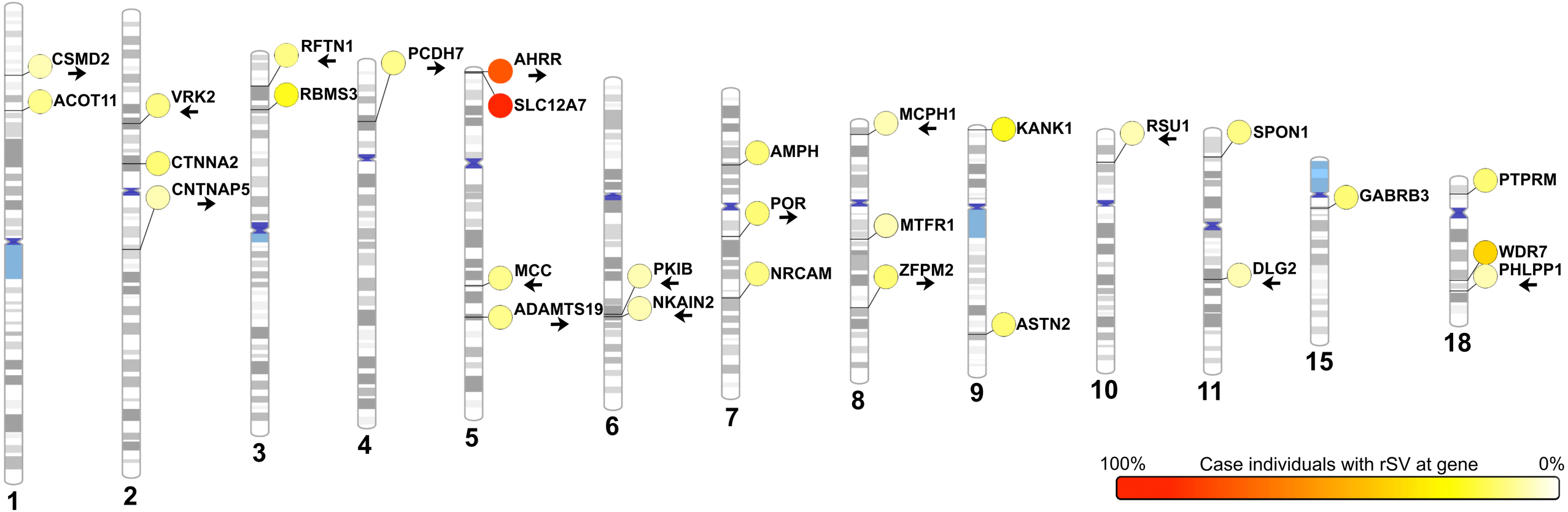
Simona, A., W. Song, D. W. Bates, and C. F. Samer. 2023. 'Polygenic risk scores in pharmacogenomics: opportunities and challenges-a mini review', *Front Genet*, 14: 1217049.

472 Stelzer, G., N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y.
473 Mazor, S. Kaplan, D. Dahary, D. Warshawsky, Y. Guan-Golan, A. Kohn, N. Rappaport, M. Safran, and D.
474 Lancet. 2016. 'The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses',
475 *Curr Protoc Bioinformatics*, 54: 1 30 1-1 30 33.
476 Turner, T. N., K. Sharma, E. C. Oh, Y. P. Liu, R. L. Collins, M. X. Sosa, D. R. Auer, H. Brand, S. J. Sanders, D.
477 Moreno-De-Luca, V. Pihur, T. Plona, K. Pike, D. R. Soppet, M. W. Smith, S. W. Cheung, C. L. Martin, M.
478 W. State, M. E. Talkowski, E. Cook, R. Huganir, N. Katsanis, and A. Chakravarti. 2015. 'Loss of delta-
479 catenin function in severe autism', *Nature*, 520: 51-6.
480 Yin, J., C. A. Chun, N. N. Zavadenko, N. L. Pechatnikova, O. Y. Naumova, H. V. Doddapaneni, J. Hu, D. M. Muzny,
481 C. P. Schaaf, and E. L. Grigorenko. 2020. 'Next Generation Sequencing of 134 Children with Autism
482 Spectrum Disorder and Regression', *Genes (Basel)*, 11.
483

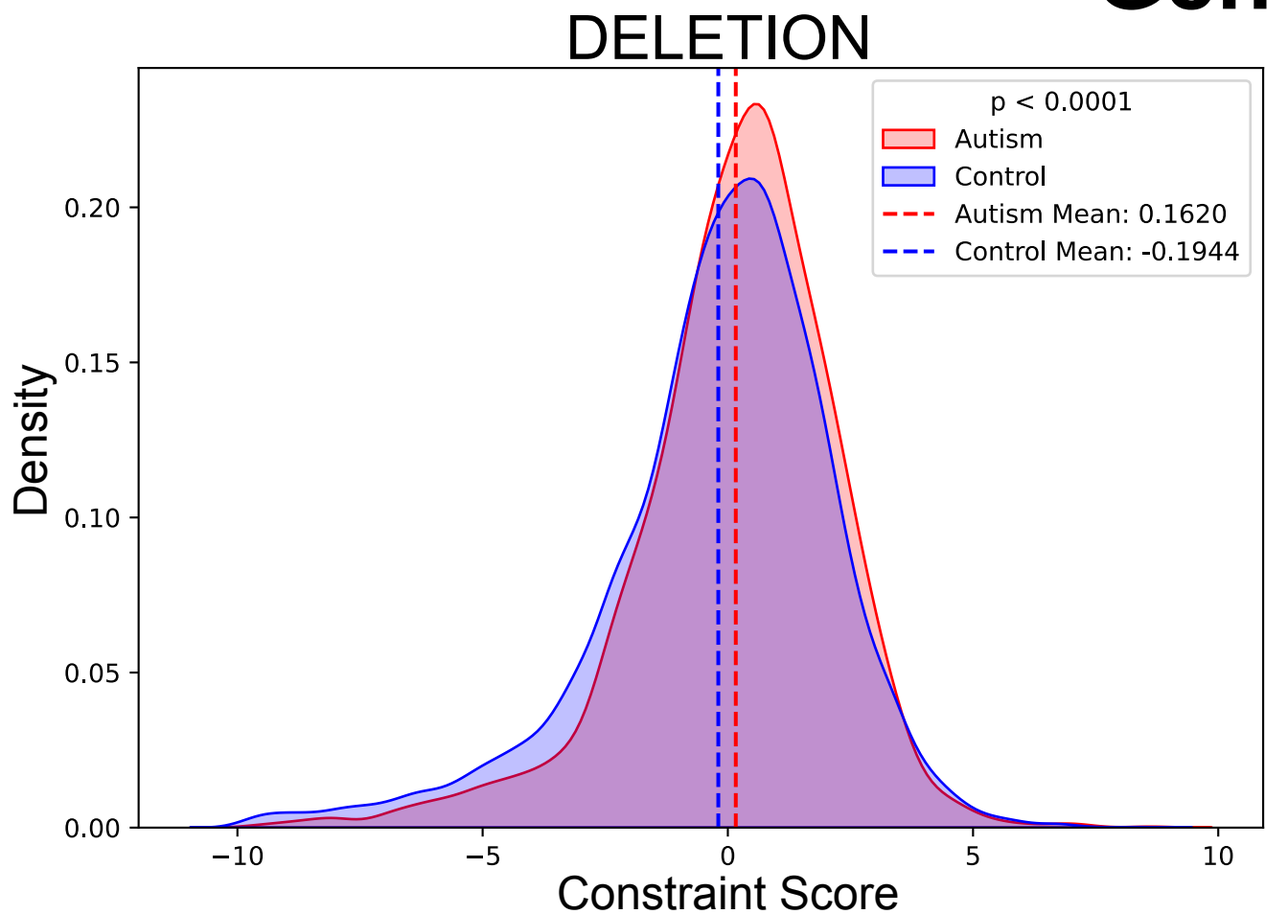

A
Cohort Individual SV Counts
SV Count
20000
15000
10000
5000
0
DEL
DUP
INS
INV
Filtering
Cohort Individual Rare SV Counts
SV Count
1000
800
600
400
200
0
DEL
DUP
INS
INV
B
Cohort Individual Average SV Length
INV
INS
DUP
DEL
50
500
1000
1500
10000
20000
30000
40000
50000
Average Length (bp)
Filtering
Cohort Individual Average Rare SV Length
INV
INS
DUP
DEL
50
500
1000
1500
10000
20000
30000
40000
50000
Average Length (bp)

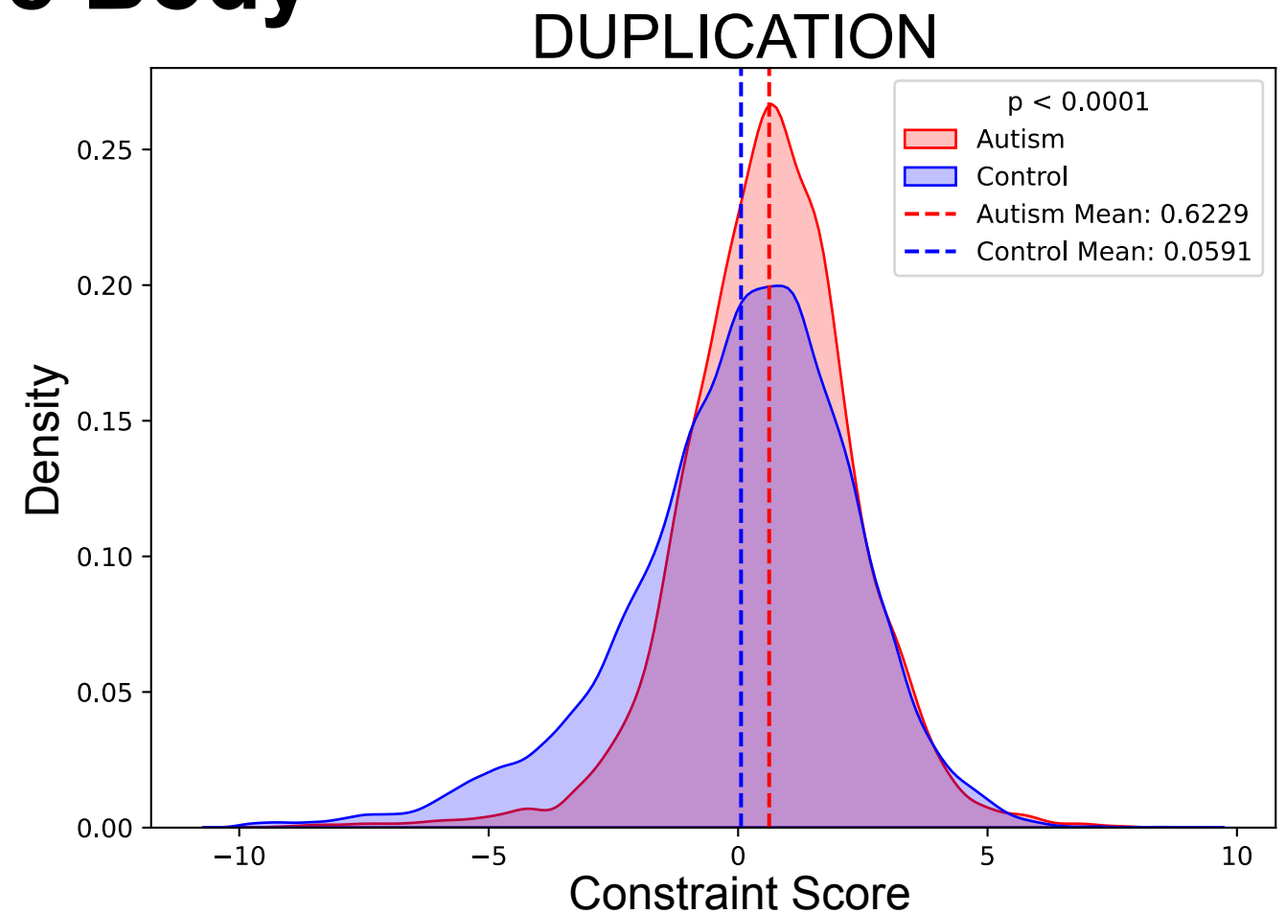

A
Global Dev. Delay + PGS Fluid Intel.
P = 0.3985
Seizure + PGS Epilepsy
P = 0.0517
Hypotonia + PGS Grip Strength
P = 0.0883
Short Stature + PGS Height
P = 0.0059
Autism + PGS Autism
P = 0.0011
PGS Z-score
4
2
0
-2
-4
-6
Case
Control
Diagnosis
B
Autism vs Control Rare SV Counts
Count
1000
800
600
400
200
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
Absolute (Max Weight)
0.15
0.10
0.05
0.00
RSU1
IPO7
DEL
DUP
INS
INV
2000
1500
1000
500
0
rSV Count
Q4
Q3
Q2
Q1
Gene Weight Quartile

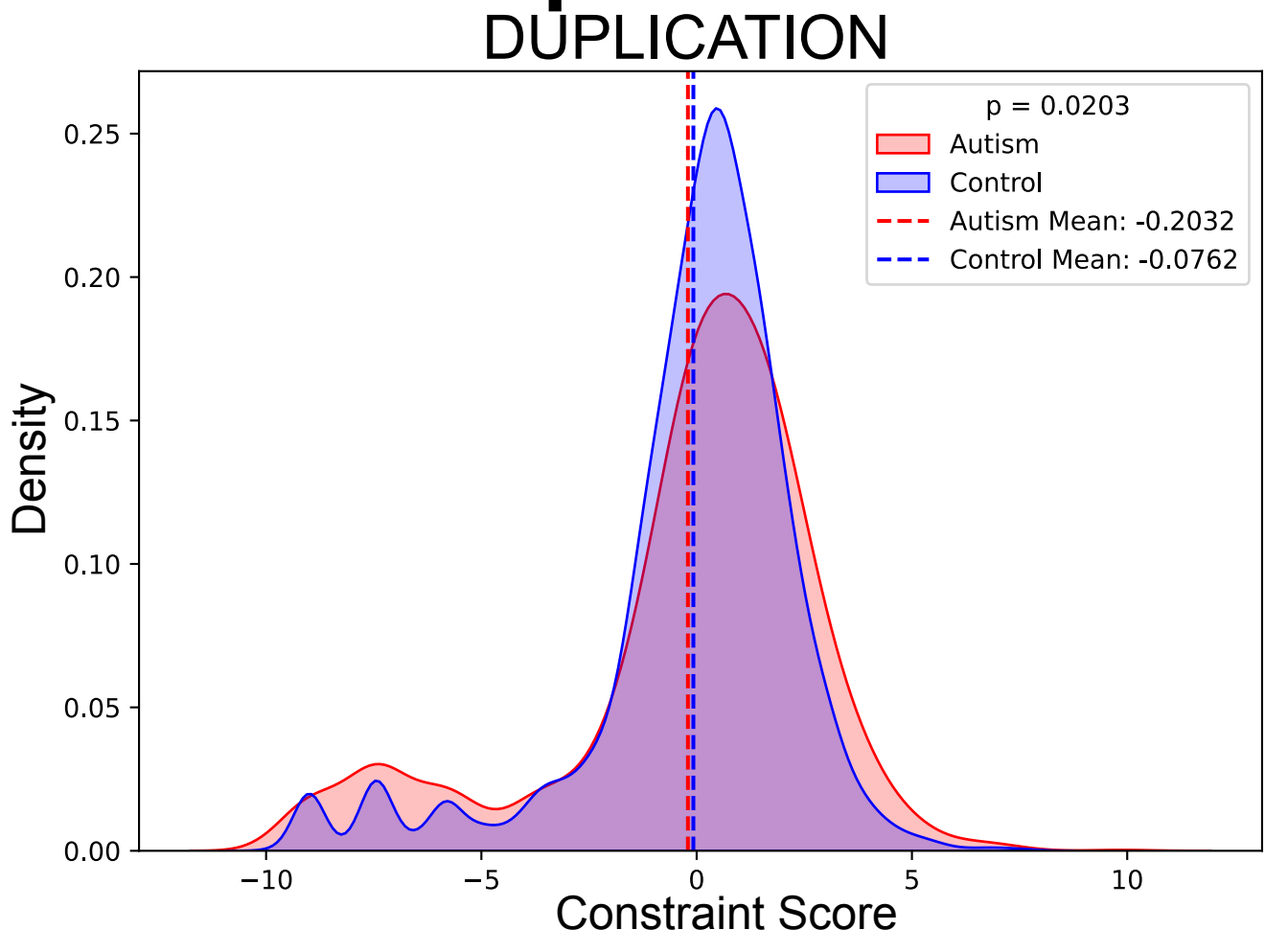

CSMD2
ACOT11
VRK2
CTNNA2
CNTNAP5
RFTN1
RBMS3
PCDH7
AHRR
SLC12A7
MCC
ADAMTS19
PKIB
NKAIN2
AMPH
POR
NRCAM
MCPH1
MTFR1
ZFPM2
KANK1
ASTN2
RSU1
SPON1
DLG2
GABRB3
PTPRM
WDR7
PHLPP1
1
2
3
4
5
6
7
8
9
10
11
15
18
100%
Case individuals with rSV at gene
0%

Gene Body

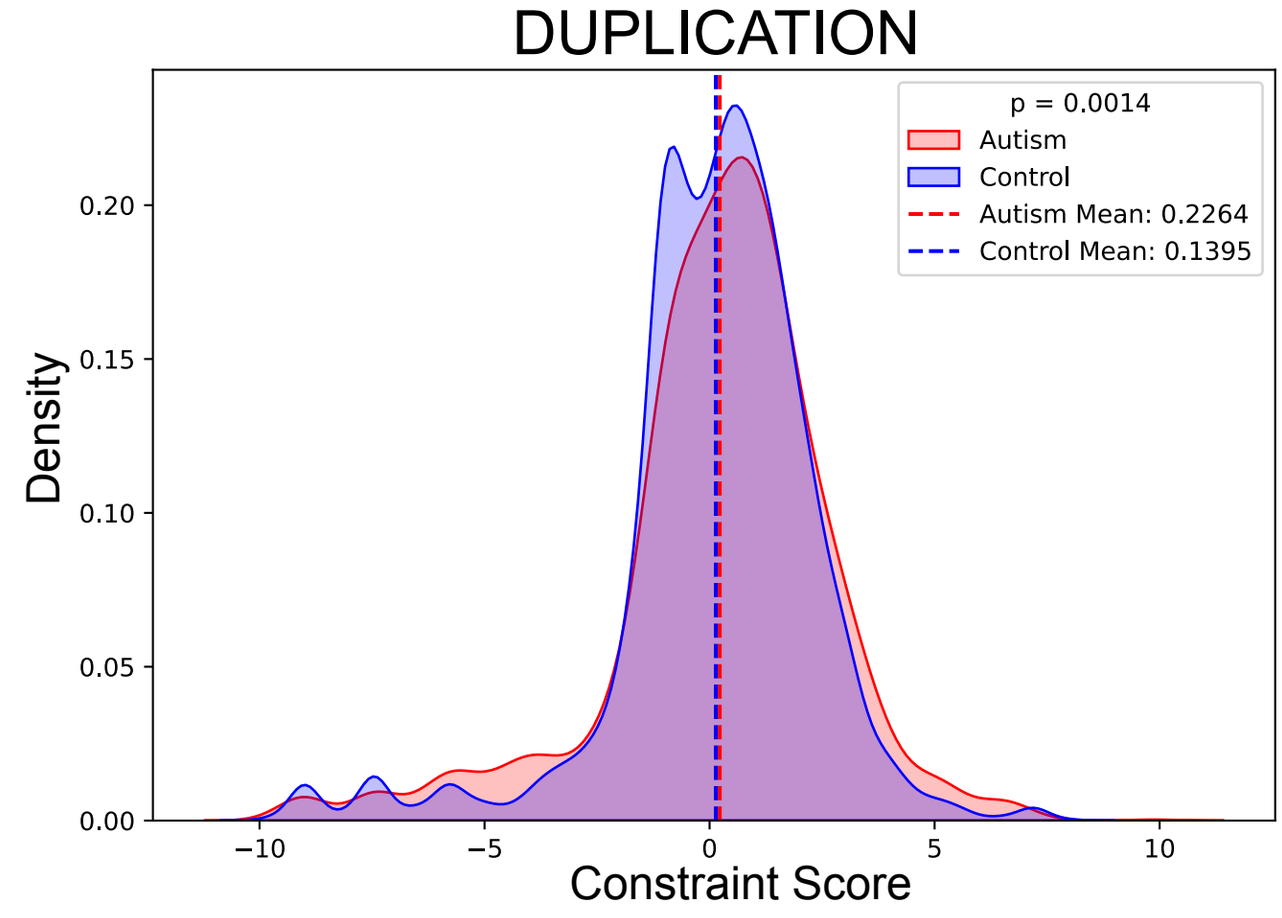


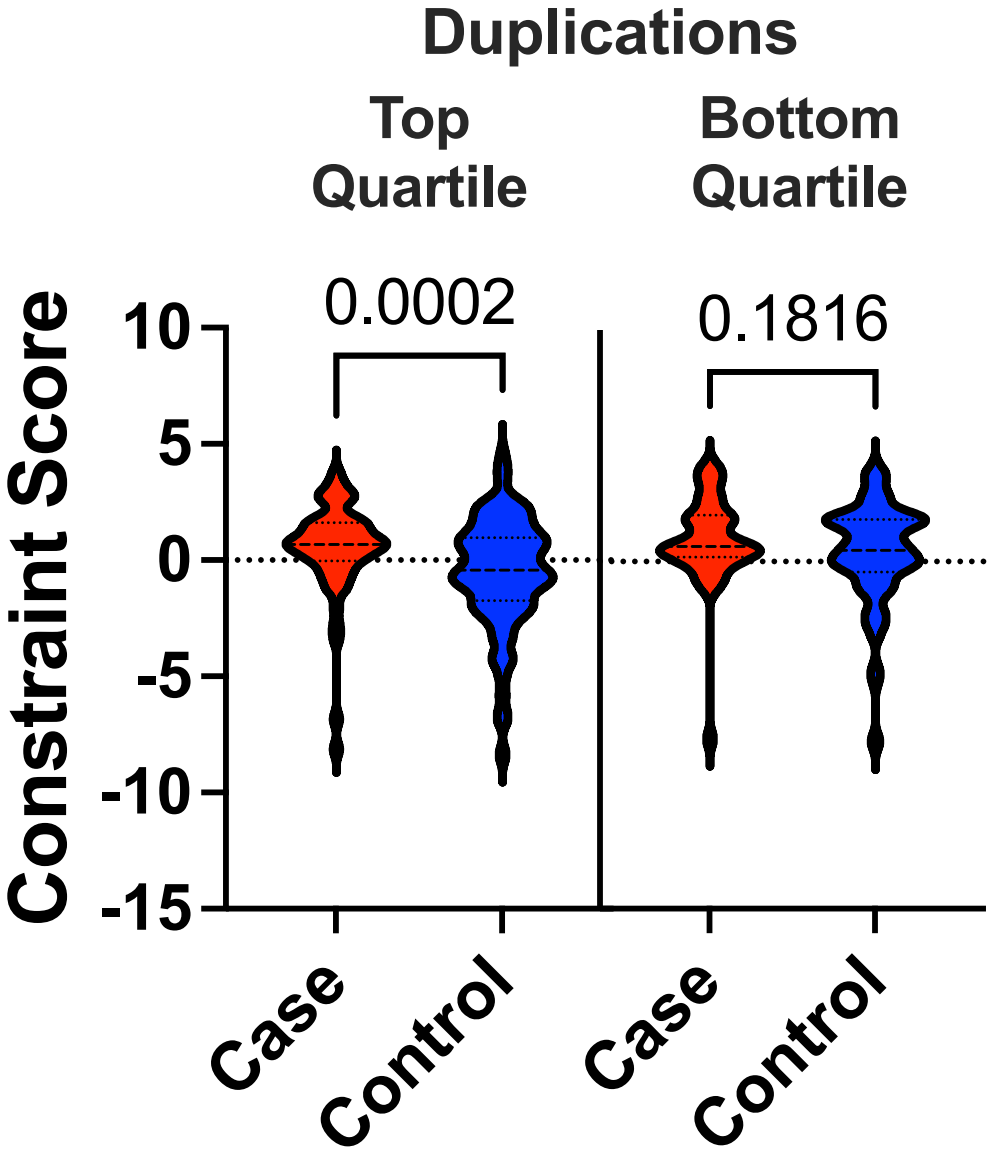


Upstream



Downstream




Duplications
Top Quartile
Bottom Quartile
0.0002
0.1816
Constraint Score
10
5
0
-5
-10
-15
Case
Control
Case
Control