frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 03 August 2023
DOI 10.3389/fnins.2023.1106573



OPEN ACCESS

EDITED BY
Flavia Venetucci Gouveia,
University of Toronto, Canada

REVIEWED BY
Christopher Scott Ward,
Baylor College of Medicine, United States
Andersen Chang,
Baylor College of Medicine, United States

*CORRESPONDENCE
Brandon L. Pearson
✉ blp2125@cumc.columbia.edu

RECEIVED 23 November 2022
ACCEPTED 17 July 2023
PUBLISHED 03 August 2023

CITATION
Baker BH, Zhang S, Simon JM, McLarnan SM, Chung WK and Pearson BL (2023) Environmental carcinogens disproportionally mutate genes implicated in neurodevelopmental disorders. *Front. Neurosci.* 17:1106573. doi: 10.3389/fnins.2023.1106573



# Environmental carcinogens disproportionally mutate genes implicated in neurodevelopmental disorders

Brennan H. Baker[1], Shaoyi Zhang[2], Jeremy M. Simon[3], Sarah M. McLarnan[1], Wendy K. Chung[4] and Brandon L. Pearson[1]*

[1]Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, NY, United States, [2]Master of Public Health Program, Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, United States, [3]Department of Genetics and Neuroscience Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States, [4]Department of Pediatrics and Medicine, Columbia University Irving Medical Center, New York, NY, United States

**Introduction:** *De novo* mutations contribute to a large proportion of sporadic psychiatric and developmental disorders, yet the potential role of environmental carcinogens as drivers of causal *de novo* mutations in neurodevelopmental disorders is poorly studied.

**Methods:** To explore environmental mutation vulnerability of disease-associated gene sets, we analyzed publicly available whole genome sequencing datasets of mutations in human induced pluripotent stem cell clonal lines exposed to 12 classes of environmental carcinogens, and human lung cancers from individuals living in highly polluted regions. We compared observed rates of exposure-induced mutations in disease-related gene sets with the expected rates of mutations based on control genes randomly sampled from the genome using exact binomial tests. To explore the role of sequence characteristics in mutation vulnerability, we modeled the effects of sequence length, gene expression, and percent GC content on mutation rates of entire genes and gene coding sequences using multivariate Quasi-Poisson regressions.

**Results:** We demonstrate that several mutagens, including radiation and polycyclic aromatic hydrocarbons, disproportionately mutate genes related to neurodevelopmental disorders including autism spectrum disorders, schizophrenia, and attention deficit hyperactivity disorder. Other disease genes including amyotrophic lateral sclerosis, Alzheimer's disease, congenital heart disease, orofacial clefts, and coronary artery disease were generally not mutated more than expected. Longer sequence length was more strongly associated with elevated mutations in entire genes compared with mutations in coding sequences. Increased expression was associated with decreased coding sequence mutation rate, but not with the mutability of entire genes. Increased GC content was associated with increased coding sequence mutation rates but decreased mutation rates in entire genes.

**Discussion:** Our findings support the possibility that neurodevelopmental disorder genetic etiology is partially driven by a contribution of environment-induced germ line and somatic mutations.

KEYWORDS

**somatic mutation, mutagenesis, *de novo* mutation, carcinogen, neurodevelopmental disorders, autism**

# Introduction

While cancer epidemiologic studies have a long history of integrating genetic and environmental factors into disease causation (Shields and Harris, 2000), researchers, with small exception (Kinney et al., 2010; Pugsley et al., 2021), have not readily implicated environmentally-induced mutations as etiological drivers of neurodevelopmental disorders (NDD) and other diseases. *De novo* mutations contribute to a large proportion of sporadic cases of ASD, schizophrenia, and intellectual disability (De Ligt et al., 2012; Xu et al., 2012; Fromer et al., 2014; Iossifov et al., 2014), yet the underlying mutational processes have not been interrogated, or have been attributed to intrinsic mutational processes (e.g., random replication error) rather than environmental carcinogens. Similarly, environmental exposures may be responsible for a large proportion of NDD (Landrigan, 2010; Bellinger, 2012; Rauh and Margolis, 2016). While potential underlying molecular mechanisms such as epigenetics have been explored in great detail (Perera and Herbstman, 2011; Tran and Miyake, 2017; Emberti Gialloreti et al., 2019), environmentally induced mutation remains a strong yet generally untested candidate mechanism that may link environmental exposures to neurodevelopment (Kinney et al., 2010; Pugsley et al., 2021). For instance, PAHs—a class of chemicals found in tobacco smoke and air pollution—form metabolites in the body that bind with DNA and promote mutation (Whyatt et al., 1998). Consequently, PAHs are well known causes of cancer (Boffetta et al., 1997; Kriek et al., 1998; Kim et al., 2013). Epidemiologic studies have linked prenatal PAH exposure to cognitive developmental delays, reduced intelligence, and ASD (Perera et al., 2006; Edwards et al., 2010; von Ehrenstein et al., 2014; Jedrychowski et al., 2015). However, no studies have examined whether mutations in NDD genes induced by PAHs and other environmental exposures contribute to these epidemiologic associations despite evidence that NDD genes are generally longer (King et al., 2013; Sugino et al., 2014; Gabel et al., 2015) and show considerable overlap with cancer driver genes (Crawley et al., 2016; Qi et al., 2016). To test the hypothesis that NDD genes are more susceptible to mutagens than non-NDD genes, we analyzed a whole genome sequencing (WGS) dataset containing nearly 200,000 single nucleotide substitution mutations in human induced pluripotent stem cell (iPSC) clonal lines exposed to 12 classes of environmental carcinogens (Kucab et al., 2019). We assessed the susceptibility to environmental mutation of genes and disease-associated gene sets by (1) evaluating gene ontology for top mutated genes; (2) developing an online tool for assessing the propensity of 12 mutagen classes to cause mutations in gene sets associated with specific human diseases; (3) investigating gene length, expression, and GC content as potential drivers of elevated mutability using Quasi-Poisson models; and (4) testing whether specific disease-related genes are enriched for bulky DNA adduct repair.

# Materials and methods

## Environmental mutation vulnerability of disease genes

Analyses were performed using R (Team, 2018). We analyzed the substitution mutations from 324 iPSC subclones dosed with 79 environmental carcinogens (Kucab et al., 2019). From whole-genome-sequencing data at ~ 30-fold depth, Kucab et al. (2019) called mutations in subclones subtracting on the primary iPSC parental clone. We compared the observed rates of exposure-induced mutations in disease-related gene sets with the expected rates of mutations based on control genes randomly sampled from the genome. Disease gene sets contained 91 ASD (Abrahams et al., 2013), 104 schizophrenia (Wang et al., 2019), 25 ADHD (Demontis et al., 2019), 33 Alzheimer's (Giri et al., 2016), 18 ALS (Association T.A., 2019), 81 type 2 diabetes (Mahajan et al., 2014), 80 coronary artery disease (Nikpay et al., 2015), 96 obesity (Locke et al., 2015), 253 congenital heart disease (Jin et al., 2017), and 31 orofacial cleft genes (Beaty et al., 2016; Supplementary Table S1). Gene sets were either curated (i.e., published in review articles or curated by scientific organizations) or based on genes with significant disease-associated loci from genome wide association studies (GWAS). We included adult onset, congenital, heritable, and life-style-associated diseases to determine if our hypothesized NDD enrichment was specific. Since our analyses were restricted to just a handful of disease gene sets and results could depend on the methods of gene set curation, we created an online tool where custom gene lists can be queried using the algorithm we generated.[1] Using this tool, users may input more up-to-date gene lists. For example, our ASD list included all genes labeled as high confidence by the Simons Foundation Autism Research Initiative (SFARI) at the time of the analysis, but SFARI is constantly updating this gene list as our understanding of the genetic basis of ASD evolves.

To determine expected mutation rates, we randomly sampled 1,000 sets of 300 genes from the human genome and used the iPSC mutation dataset (Kucab et al., 2019) to calculate average rates of mutation per-gene-per-treated iPSC subclone within each exposure class. Our unit of analysis was mutations per gene, so it was not necessary to match the number of randomly sampled genes with the number of genes in each disease set. To check this assumption, we plotted the relationship between the size of randomly sampled gene sets, varying from 10 to 300 genes, with the number of mutations per subclone treated with the radiation class of chemicals. To characterize the degree to which certain disease gene sets were mutated more than expected, we compared these hypothesized expected mutation rates to the mutation rates for each disease gene set within each environmental exposure in clonal iPSC cultures (Kucab et al., 2019) using two-sided exact binomial tests. For a given chemical exposure and disease gene set, the exact binomial test null hypothesis was that the disease gene set had the same per gene mutation rate as the per gene mutation rate of the 1,000 sets of 300 genes described above. Rejection of the null hypothesis indicated that the disease gene set was mutated more or less than the mutation rate of randomly sampled genes. Single genes were allowed to contribute multiple mutations to the mutation rate numerators. Significance was assessed at alpha level 0.05 with table wide Bonferroni corrections. This analysis was repeated for mutations in entire genes as well as coding sequence (CDS) mutations determined using the Ensembl variant effect predictor (McLaren et al., 2016). Although entire genes contain introns and other non-coding sequence, a large proportion of GWAS signals map to non-coding regions (Zhang and Lupski, 2015), so variants in these loci may still contribute to disease.

For mutations in entire genes in PAH-treated iPSCs, we conducted a sensitivity analysis by calculating *p*-values from empirical null distributions rather than from exact binomial tests. Monte Carlo null distributions for each disease gene set were obtained by randomly sampling 1,000 sets of genes from the human genome equal to the number of genes in a given disease gene set. The total number of

1 http://environmentalmutation.com

mutations in each randomly sampled gene set was determined. Two-tailed p-values were calculated as the proportion of randomly sampled gene sets mutated more or less than the comparison disease gene set, whichever was smallest, multiplied by two.

To externally-validate this approach, we repeated the gene mutation analysis in an independent dataset of human WGS data from 14 lung cancers from individuals living in highly polluted regions (Yu et al., 2015). Because PAHs are a major component of pollution, we hypothesized that mutational patterns would be similar between these samples and the PAH-treated iPSCs.

We conducted a sensitivity analysis to explore the role of gene length in environmental mutagen vulnerability. In this analysis, a selected NDD gene list was created by combining all ASD, ADHD, and schizophrenia genes from the lists described above. We then divided the list into four separate lists based on gene length quartiles, and repeated the above analysis for mutations in entire gene bodies.

In an additional sensitivity analysis, we explored the mutational susceptibility of cancer driver genes, and genes with overlap between cancer and NDD. We utilized a list of 233 high confidence cancer genes with confidence scores $\geq 1.5$ based on a scoring system developed by (Bailey et al., 2018), and a list of 14 genes that overlap between this cancer gene list and the selected NDD gene list described above.

## Gene ontology

We performed gene ontology (GO) analysis on all genes which contained coding sequence (CDS) variants in PAH-treated iPSCs. GO analysis was performed using FUMA with ensembl version 92, protein coding genes set as the background, and a Bonferroni correction (Watanabe et al., 2017). In an additional sensitivity analysis, we included all genes with CDS mutations in iPSCs exposed to all environmental mutagens rather than just PAH-treated iPSCs.

## Sequence characteristics and mutation vulnerability

Autism spectrum disorder-implicated NDD genes tend to be longer than other genes (King et al., 2013; Zylka et al., 2015). To visually examine if vulnerability of neurodevelopmental genes or CDS to mutagens is attributable to gene length, we plotted the distributions of entire gene and CDS lengths for our disease gene sets, along with the distributions of lengths for entire genes and CDS mutated entirely at random. Random mutations were modeled by randomly sampling (i.e., mutating) 100,000 nucleotides from all genes or all CDS in the human genome, so the probability of a sequence being mutated was entirely governed by its length.

To further explore associations of length with mutability, we modeled the effects of sequence length, expression, and percent GC content on mutation rate using multivariate Quasi-Poisson regressions, with separate models for mutations in CDS and entire genes. Gene and CDS start and stop positions were obtained from GENCODE Release 38[2] and used in conjunction with the "BSgenome.Hsapiens.UCSC.hg38" R package to calculate genomic sequence lengths (end minus start position) and GC content (proportion of sequence positions with either a G or C nucleotide). When modeling associations of sequence properties with CDS mutations, CDS lengths and GC content were computed per gene: all CDS segments within a single gene were summed as the total coding sequence length, and CDS GC content was calculated per gene rather than per individual CDS segment. Gene expression data were reads per kilobase of transcript per million mapped reads (RPKM) obtained from RNA-seq of iPSCs generated using the Sendai virus method (Churko et al., 2017), the same method used to create the iPSCs used by Kucab et al. (2019). After excluding genes with missing length, expression, or GC content data, Quasi-Poisson models included 75,756 gene and 1,852 CDS mutations. Coefficients from these models were multiplied by the interquartile range (IQR) for each variable and then exponentiated into rate ratios per IQR increase.

## Local sequence and mutation vulnerability

To explore the role of local sequence context on mutability, we aligned 7-mers centered on each gene or CDS mutation identified by Kucab et al. (2019), along with 50,000 7-mers randomly sampled from the human genome. We performed this analysis for all mutations, and stratified by chemical exposure class. We also examined the role of local sequence context by generating COSMIC signatures (Tate et al., 2019) for *de novo* mutations in individuals with neuropsychiatric diseases, including 42,607 ASD cases (Feliciano et al., 2018; Zhou et al., 2021), 617 schizophrenia cases (Fromer et al., 2014), and 145 individuals with severe intellectual disability (De Ligt et al., 2012; Rauch et al., 2012).

## Polycyclic aromatic hydrocarbon adduct repair associated mutation

To determine if NDD genes are linked to PAH adduct repair, we analyzed an existing genome-wide PAH adduct repair assay dataset (Li et al., 2017) to see if adducts are preferentially located in specific disease-related gene sets. Genome-wide PAH adduct repair data come from translesion excision repair-sequencing (tXR-seq) of GM12878 cells, which were grown to ~ 80% confluence before treatment with 2 μM benzo[a]pyrene diol epoxide-deoxyguanosine for 1 h at 37°C in a 5% $CO_2$ humidified chamber. tXR-seq captures all DNA damage, regardless of whether or not it is repaired (Li et al., 2017). We computed the average DNA damage enrichment across each gene or CDS in all 10 disease-related gene sets by inputting bigWig files from Li et al. (2017) into the deepTools2 'computeMatrix' function (Ramírez et al., 2016). By default, the 'computeMatrix' function scales input sequences to the same length. Output enrichment values from the 'computeMatrix' function are based on the units of the input bigWig files, which, in this case, were tXR-seq counts normalized for total sequencing depth on each chromosome (Li et al., 2017). We used one-way ANOVA to compare levels of DNA damage in genes, which were normally distributed, between disease gene sets, and performed pairwise contrasts with a false discovery rate correction. Kruskal–Wallis and Dunn's Test were employed for CDS DNA damage data, which were not normally distributed.

2 https://www.gencodegenes.org/human/

# Results

## Environmental mutation vulnerability of disease genes

There were 769 unique genes across all ten disease gene sets, and overlap between sets was minimal (Figure 1A). Among the 183,133 substitution mutations identified by Kucab et al. (2019), 92,204 occurred in known genes. Across all chemical treatments, all disease gene set genes had a combined 7,587 total mutations. Per-nucleotide mutation rates for each chemical by disease gene set combination ranged from 0 to $5.77 \times 10^{-7}$. We plotted the distributions of entire gene and CDS lengths for our disease gene sets, along with the distributions of lengths for the 1,000 sets of 300 randomly sampled gene and CDS sequences (Figures 1B,C). NDD and metabolic disease gene sets such as ASD, schizophrenia, ADHD, obesity, and type-2 diabetes contained the longest genes, while ALS genes and randomly sampled genes were the shortest (Figure 1B). On the other hand, average CDS lengths for most disease gene sets were comparable (Figure 1C).

We compared per gene mutation rates in our disease gene sets with genes randomly sampled from the human genome (1,000 sets of 300 genes). Mutations per subclone increased linearly with the number of randomly sampled genes while the number of mutations per gene per subclone remained flat (Figure 2), confirming that it was not necessary to match the number of randomly sampled genes with the number of genes in each disease set.

The most mutagenic exposures were radiation and PAHs, which induced an average of 0.066 and 0.058 substitution mutations per-gene-per-treated iPSC subclone, respectively across our 10 disease related gene sets (Figure 3A; Supplementary Table S2). ASD, ADHD, schizophrenia, obesity, and type-2 diabetes genes were mutated significantly more than expected by almost every chemical class. Congenital heart disease, oral cleft, and coronary artery disease were rarely mutated more than expected; while Alzheimer's disease genes were never mutated more than expected (Figure 3A; Supplementary Table S3). There was some evidence for ALS genes being mutated less than the expected per-gene mutation rate by aromatic amines and nitro-PAHs, although these differences were not statistically significant after Bonferroni correction. Results were similar in a sensitivity analysis calculating *p*-values from Monte Carlo empirical null distributions for PAH-treatment mutations (Supplementary Figure S1). As in the main analysis, ASD, ADHD, schizophrenia, obesity, and type-2 diabetes genes were mutated significantly more than expected.

Among the 2,061 identified coding sequence variants, these overarching patterns were not observed (Figure 3B). While there was evidence for exposure causing more coding sequence mutations than

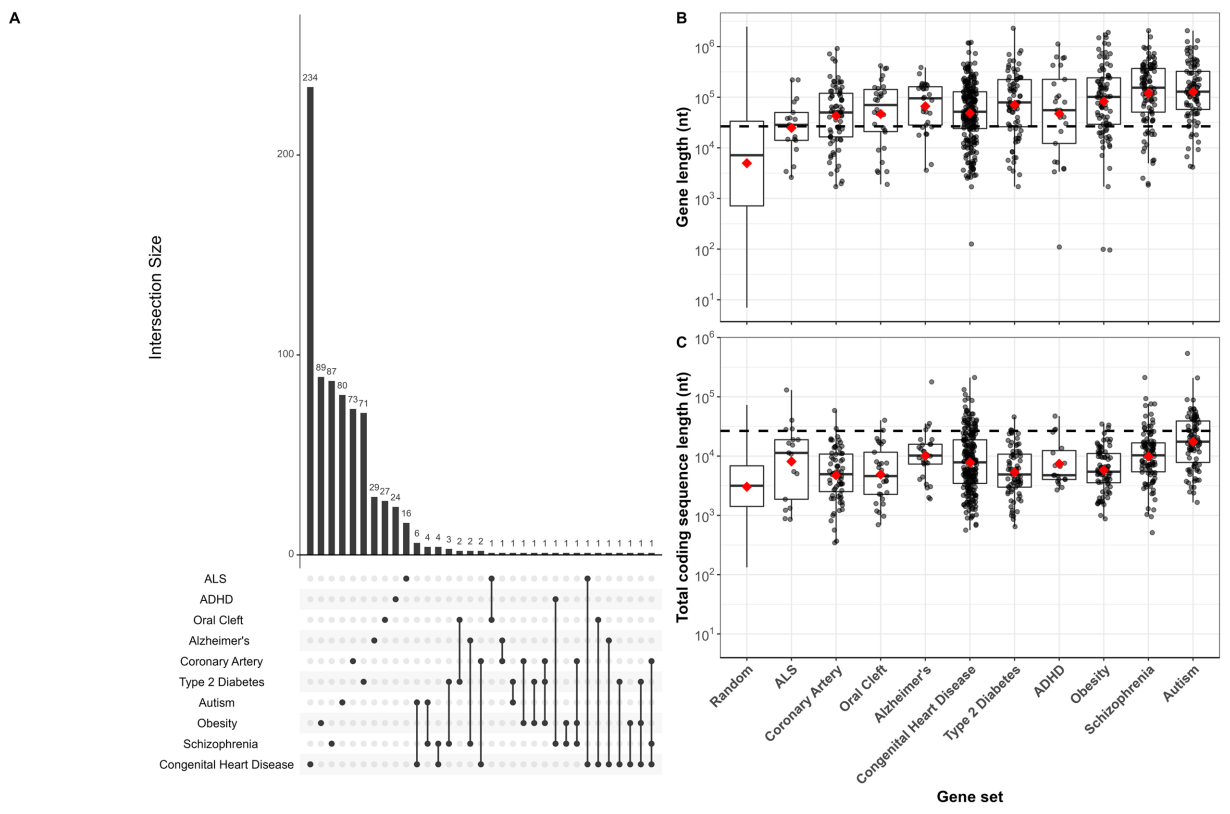


FIGURE 1
**(A)** Gene set intersection plot. Horizontal bars depict gene set sizes. Gene sets included in each intersection are indicated by a single or connected dot(s), and sizes of the intersections are indicated by vertical bars. For single, unconnected dots, vertical bars indicate the number of genes exclusive to that set. For example, among 253 Congenital Heart Disease genes, 6 overlapped with Autism genes, while 234 had no overlap with other gene sets. Intersections with zero overlap not shown. **(B,C)** Sequence length boxplots with median and 1.5 interquartile range (IQR) whiskers shown for each disease implicated gene set for entire genes **(B)** and coding sequence **(C)**. Red diamonds indicate mean sequence length for each set indicated on x-axis, and dashed black line across entire plot indicates mean gene length of randomly mutated genes (modeled by randomly sampling (i.e., mutating) 100,000 nucleotides from all genes or all CDS in the human genome). Gene and CDS boxplots also shown for all 300,000 randomly sampled sequences (1,000 sets of 300 genes or CDS).

expected for 9 specific exposure/disease combinations, none of them remained statistically significant following Bonferroni correction (Figure 3B).

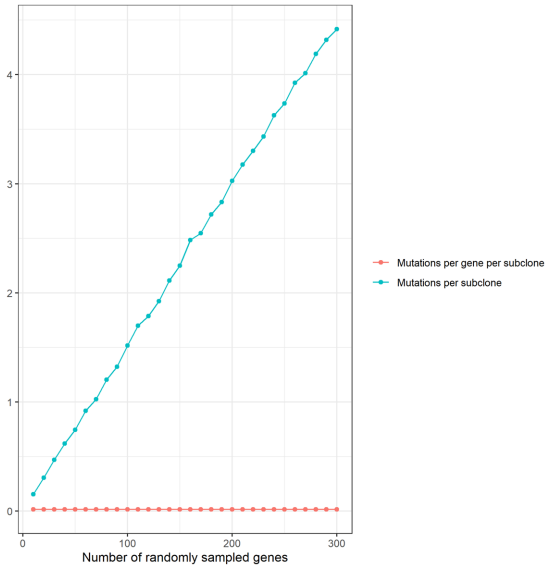


FIGURE 2
Thousand sets of genes were randomly sampled for each gene set size ranging from 10 to 300 in intervals of 10. For each random gene set size, the number of mutations in subclones treated with the radiation class of chemicals was determined. Mean mutations per subclone and mutations per gene per subclone were plotted against the random set size.

We repeated this analysis in an independent dataset of human WGS data from 14 lung cancers from individuals living in highly polluted regions. In these samples, the greatest increases in observed over expected gene mutations were in genes related to ASD, schizophrenia, and obesity (Figure 3C). However, contrasting with the mutational patterns in PAH-exposed iPSCs, ADHD genes were not mutated more than expected, and genes associated with congenital heart defects were mutated more than expected even after Bonferroni correction (Figure 3B).

Average gene lengths of disease gene sets were consistent with patterns of mutability. For instance, ASD, schizophrenia, obesity, and ADHD genes, which were on average longer than other disease-related genes, had the greatest increases in observed versus expected chemical-induced mutations, while ALS, coronary artery disease, oral cleft, and Alzheimer's disease genes, which are much shorter in length, were not mutated more than expected (*cf.* Figures 1B, 3A). Genes mutated entirely at random (modeled by randomly sampling nucleotides from the genome) were on average longer than all disease-associated genes, further indicating that gene length was a strong driver of mutability (dashed horizontal line, Figure 1B).

A sensitivity analysis stratifying NDD genes by sequence length quartiles demonstrated a strong role of sequence length in NDD gene mutability (Supplementary Figure S2). Genes in the top quartile had an average length of 678,000 nucleotides and were mutated more than the expected per-gene mutation rate by all chemical exposure classes. By contrast, the bottom quartile genes averaged 20,000 nucleotides in length and were mutated less than the expected mutation rate by several exposures (Supplementary Figure S2).

Another sensitivity analysis explored the mutability of cancer driver genes, and genes overlapping between cancer and

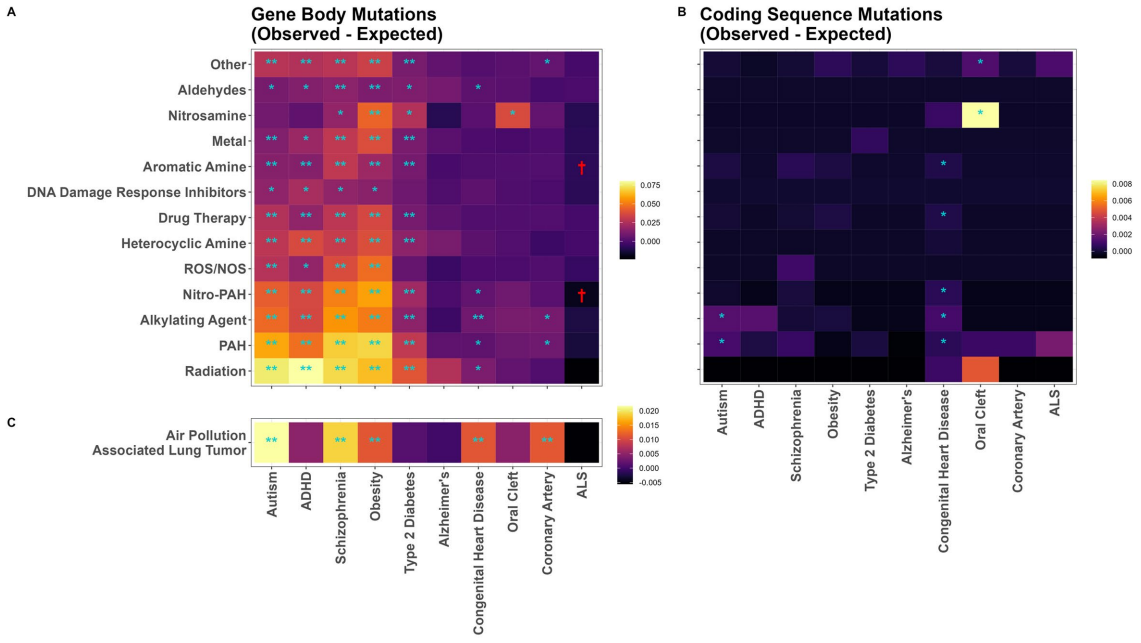


FIGURE 3
Mutation vulnerability of disease associated genes to various carcinogens. Heatmaps of observed minus expected mutations in disease gene sets for **(A)** mutations in gene body sequence following chemical treatment in iPSC, **(B)** mutations in coding sequences following chemical treatment in iPSC, and **(C)** mutations in 14 human lung cancers from individuals living in highly polluted regions. Significant levels from exact binomial tests. Higher mutation rates: *$p < 0.05$; **Bonferroni-adjusted $p < 0.05$. Lower mutation rates: †$p < 0.05$.

neurodevelopmental processes (Supplementary Figure S3). NDD genes were mutated at higher-than-expected rates by all chemical classes except for nitrosamines, while cancer driver genes were never mutated more than expected. Furthermore, genes overlapping between the NDD and cancer lists were never mutated more than expected.

## Gene ontology

Gene ontology analysis on the 692 genes which contained coding sequence (CDS) variants in PAH-treated iPSCs revealed enriched gene ontologies closely related to neurodevelopment: neuron projection, neuron part, and calcium ion binding (Supplementary Table S4). Furthermore, the enriched plasma membrane term may be related to metabolic diseases including obesity and type 2 diabetes (Cheng et al., 2018). Similar results were obtained when the analysis included all 2,061 CDS mutations in iPSCs exposed to all environmental mutagens rather than just PAH-treated iPSCs. For instance, the top three gene ontology terms were neuron projection guidance, sensory organ morphogenesis, and cell morphogenesis involved in neuron differentiation (Supplementary Table S5), supporting our hypothesis that NDD genes are particularly vulnerable to environmental mutagens.

In addition to the vulnerability of NDD genes to environmental mutagens uncovered here, genes associated with obesity and type 2 diabetes were mutated by chemical treatment more than expected. We therefore hypothesized that these genes might be linked to neurodevelopmental processes. To explore this possibility, we performed GO analysis on our list of obesity- and type 2 diabetes-associated genes, but found no evidence for enrichment of NDD processes (Supplementary Table S6).

## Sequence characteristics and mutation vulnerability

Quasi-Poisson regressions further supported a stronger role of sequence length in gene but not CDS mutation number. Controlling for expression and GC content, each interquartile range increase (IQR) in gene length was associated with a 1.104-fold increase in gene mutation rate (rate ratio (RR) = 1.104, 95% CI [1.102, 1.105]; Figure 4A), while an IQR-increase in CDS length was associated with a 1.050-fold increased CDS mutation rate (RR = 1.050, 95% CI [1.045, 1.055]; Figure 4D).

Quasi-Poisson regressions also showed significant effects of GC content and expression on gene and CDS mutability. Expression was not associated with mutability for genes (RR = 1.001, 95% CI [0.988, 1.014]; Figure 4B), while each IQR increase in expression was associated with a 13% decreased mutation rate for CDS (RR = 0.870,

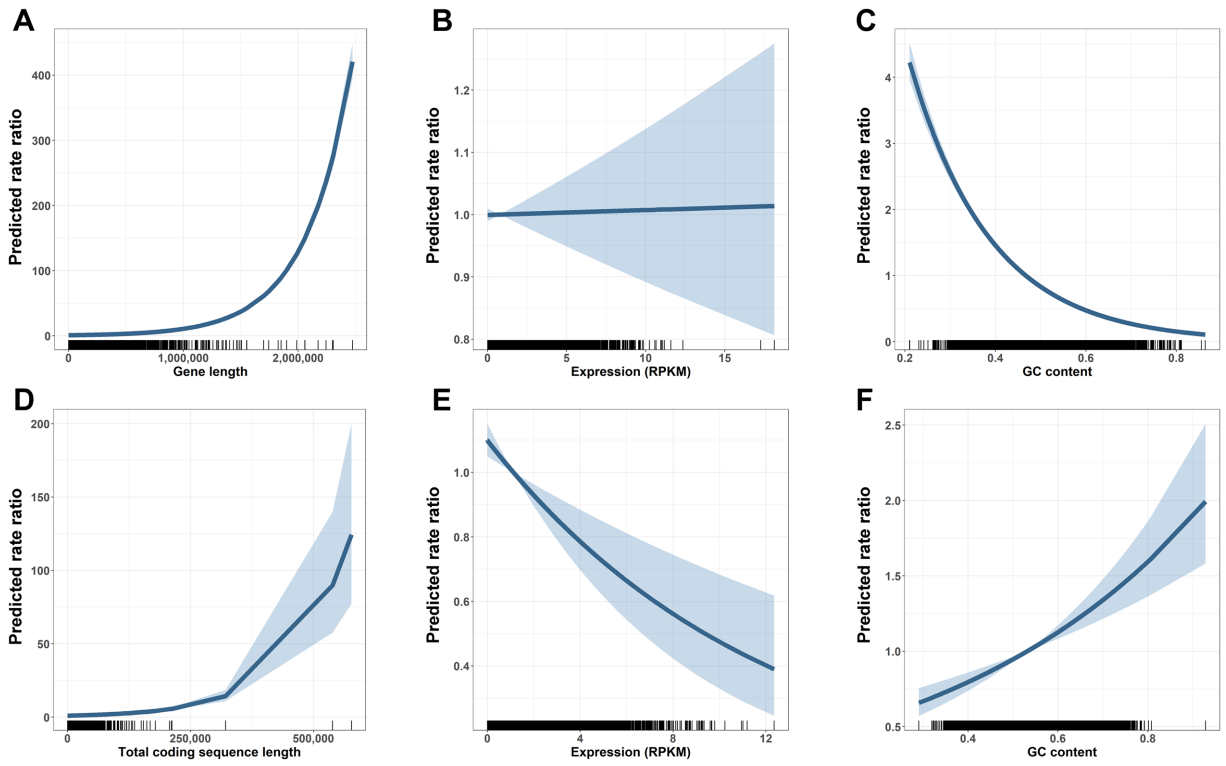


FIGURE 4
Gene length and other determinants of mutation vulnerability across gene bodies and coding sequences. Gene body Quasi-Poisson model results **(A–C)** depict centered rate ratios (lines) and 95% confidence intervals (shaded regions) at each value of *x* (i.e., panel A shows the predicted mutation rate ratio of a gene with gene length indicated on the *x*-axis versus a gene with length, expression, and GC content set to the mean). Distribution of gene characteristics shown along the x axis, with one mark per observation. Longer gene length is associated with higher mutation frequencies across carcinogen treated cells **(A)**. Expression is not associated with altered mutation risk across gene bodies **(B)**. Higher GC content is associated with decreased mutation risk across gene bodies **(C)**. Coding sequence (CDS) Quasi-Poisson model results **(D–F)** depict centered rate ratios (lines) and 95% confidence intervals (shaded regions) at each value of *x*. Distribution of CDS characteristics shown along the *x* axis, with one mark per observation. Longer coding sequence is associated with a modestly increased risk of mutation **(D)**. Higher gene expression is associated with reduced CDS mutations **(E)**. In contrast to the gene body, higher GC content is associated with increased risk of CDS mutation **(F)**.

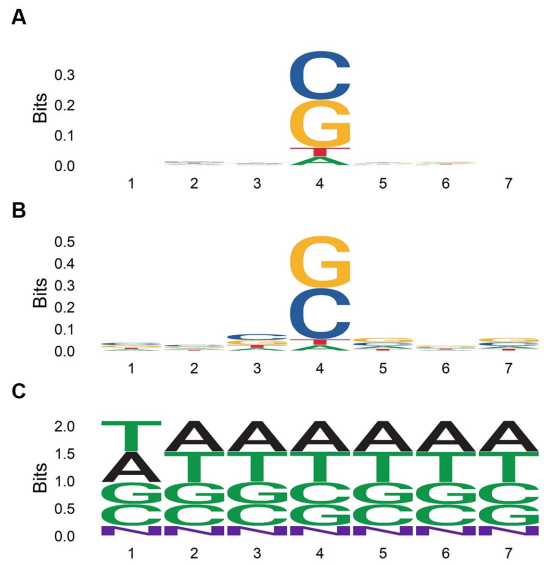


FIGURE 5
Nucleotide content of 7-mers centered on each gene body **(A)** or coding sequence **(B)** mutation among all single base substitutions identified by Kucab et al. (2019), and for 50,000 7-mers randomly sampled from the human genome **(C)**.

95% CI [0.812, 0.931]; Figure 4E). Similarly, the consequence of GC content was different for genes and CDS. Each IQR-increase in gene GC content was associated with a 0.524-fold decreased mutation rate (RR = 0.524, 95% CI [0.508, 0.539]; Figure 4C), while each IQR-increase in CDS GC content was associated with a 1.274-fold increased mutation rate (RR = 1.274, 95% CI [1.175, 1.381]; Figure 4F).

## Local sequence and mutation vulnerability

We aligned 7-mers centered on each gene or CDS mutation, along with randomly sampled 7-mers from the human genome (Figure 5). Mutated regions were GC enriched, while randomly sampled 7-mers contained equal proportions of each nucleotide. G and C were more highly enriched in 7-mers centered on CDS mutations compared to 7-mers centered on gene mutations. Thus, CDS mutations may be governed more strongly by local GC content. When aligning 7-mers on each gene or CDS mutation stratified by chemical exposure, this pattern held for some but not all chemical classes (Figure 6).

In the COSMIC mutational signatures analysis, single base substitution enrichments for all neuropsychiatric cases and controls

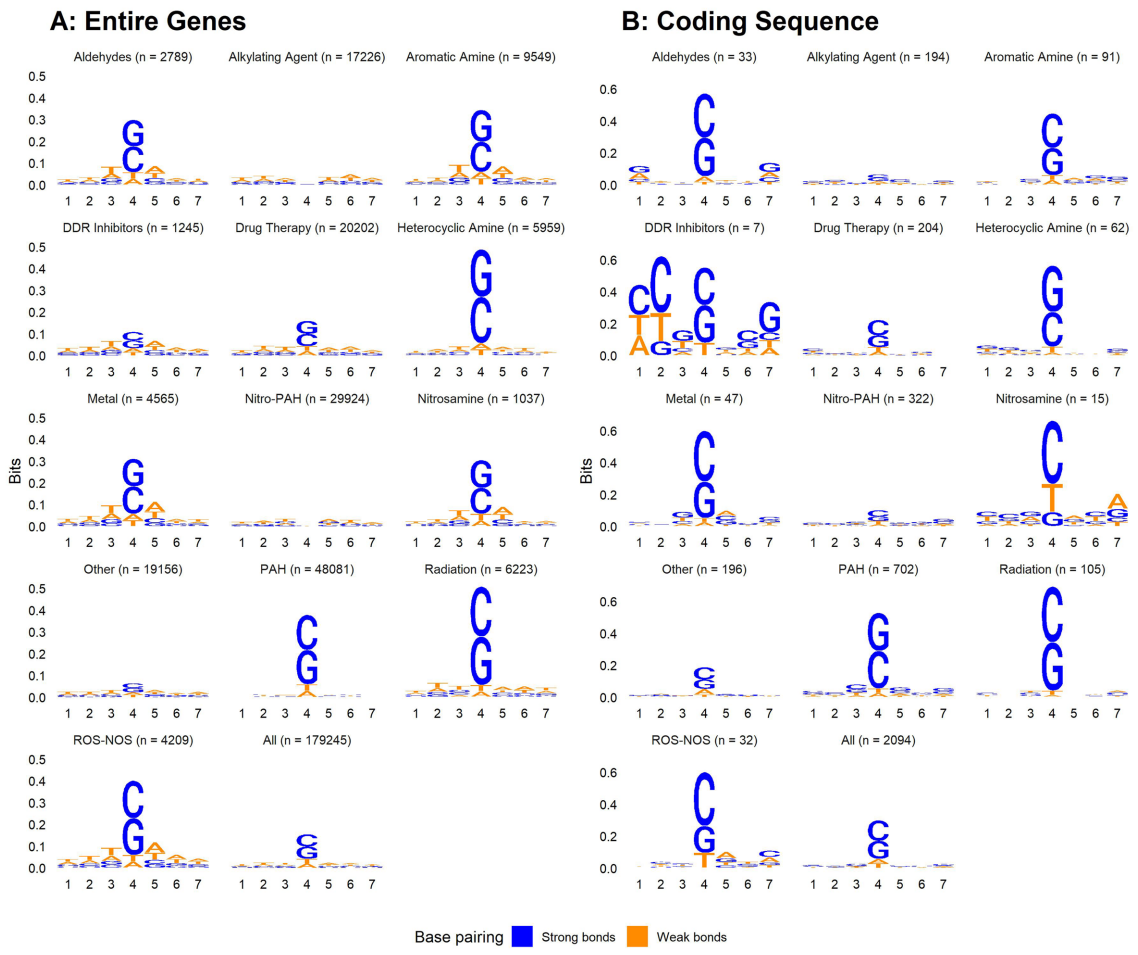


FIGURE 6
Nucleotide content of 7-mers centered on each gene body **(A)** or coding sequence **(B)** mutation among all single base substitutions identified by Kucab et al. (2019), stratified by chemical class.

10.3389/fnins.2023.1106573

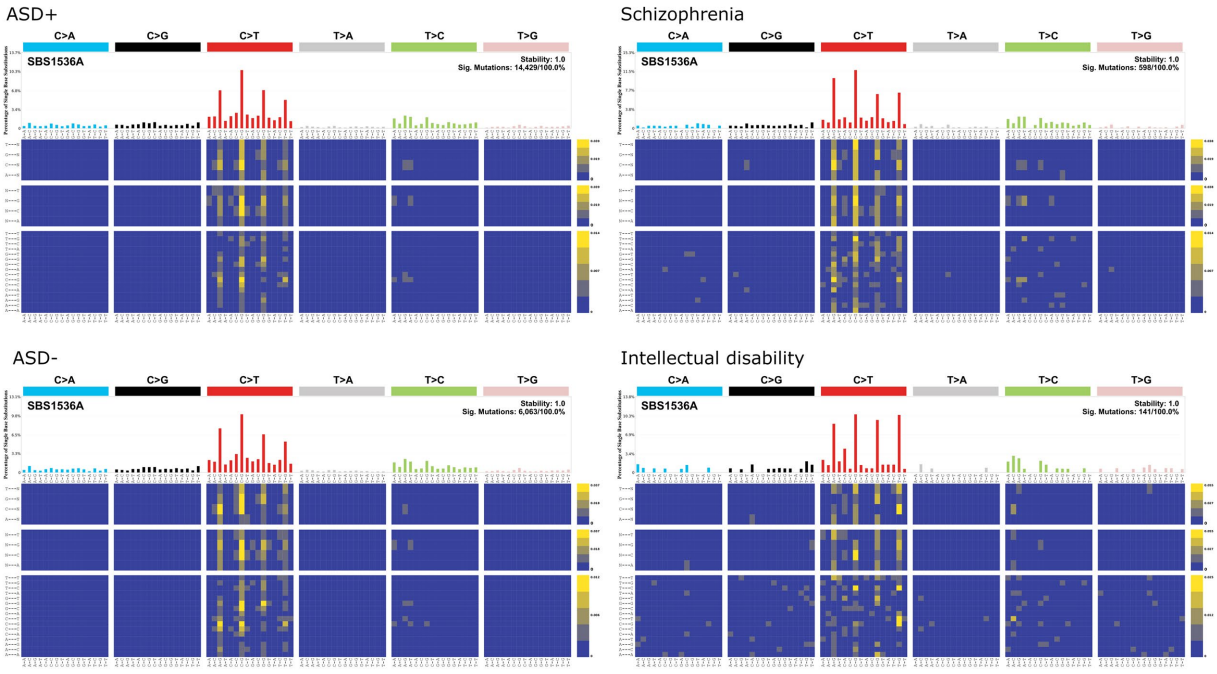


FIGURE 7
COSMIC single base substitution signatures for *de novo* mutations in individuals living with autism spectrum disorders (ASD) and their family members as controls (ASD−; Feliciano et al., 2018), schizophrenia (Fromer et al., 2014), and intellectual disability (De Ligt et al., 2012).

were clock-like/aging associated signatures (i.e., SBS1; Figure 7), which are enriched for NpCpG to NpTpG substitutions. The SBS1 signature does not resemble any of the chemical mutation signatures identified by Kucab et al. (2019). One could interpret this preliminary analysis to suggest that *de novo* mutations in ASD, schizophrenia, and intellectual disability reflect sporadic mutational processes rather than chemical-induced mutation. However, it is also possible that mutational signatures generated from iPSC cultures are not readily comparable to *in vivo* human mutational signatures. For instance, methylated CpG sequences are disproportionately targeted by environmental carcinogens such as PAHs, which form guanine adducts that induce G to T transversions at methylated CpGs (Pfeifer, 2006).

## Polycyclic aromatic hydrocarbon adduct repair associated mutation

An existing genome-wide PAH adduct repair assay dataset (Li et al., 2017) was utilized to determine if NDD genes are linked to PAH adduct repair. Pairwise contrasts revealed that DNA damage following PAH treatment was significantly enriched in genes of ASD-related genes compared to genes associated with coronary artery disease, congenital heart defects, and orofacial cleft (Figures 8A, B; Supplementary Table S7). Schizophrenia genes similarly demonstrated more DNA damage compared to coronary artery disease and congenital heart defect genes (Figures 8A, B). However, this pattern of increased DNA damage in neurodevelopmental diseases was not observed for CDS. In fact, ASD and schizophrenia CDS were among the diseases with the lowest levels of DNA damage (Figures 8C, D). Although over half of the Dunn's Tests contrasts were significant (Supplementary Table S7), differences in the mean and median CDS DNA damage enrichments between diseases were minimal (Figures 8C, D).

# Discussion

We have shown that environmental carcinogens may disproportionately mutate neurodevelopmental and metabolic genes. ASD, ADHD, schizophrenia, obesity, and type-2 diabetes genes were mutated significantly more than expected based on the mutation rate of randomly sampled genes, while GO analyses revealed that genes mutated by PAHs and other environmental carcinogens were overwhelmingly enriched for neurodevelopmental processes. Environmentally induced mutations may play a greater role in neurodevelopmental disease than previously assumed. Rather than attributing sporadic neurodevelopmental diseases to intrinsic mutational processes, this work suggests that some proportion of genetic neurodevelopmental disease risk may be explained by environmental mutagenesis.

Neurodevelopmental genes may be particularly sensitive to mutation because the transcriptome of neural tissues, especially neurons, is biased toward longer genes (King et al., 2013; Zylka et al., 2015). Our analyses revealed that sequence length was a strong driver of mutability, although the association between sequence length and mutability was twice as strong for genes

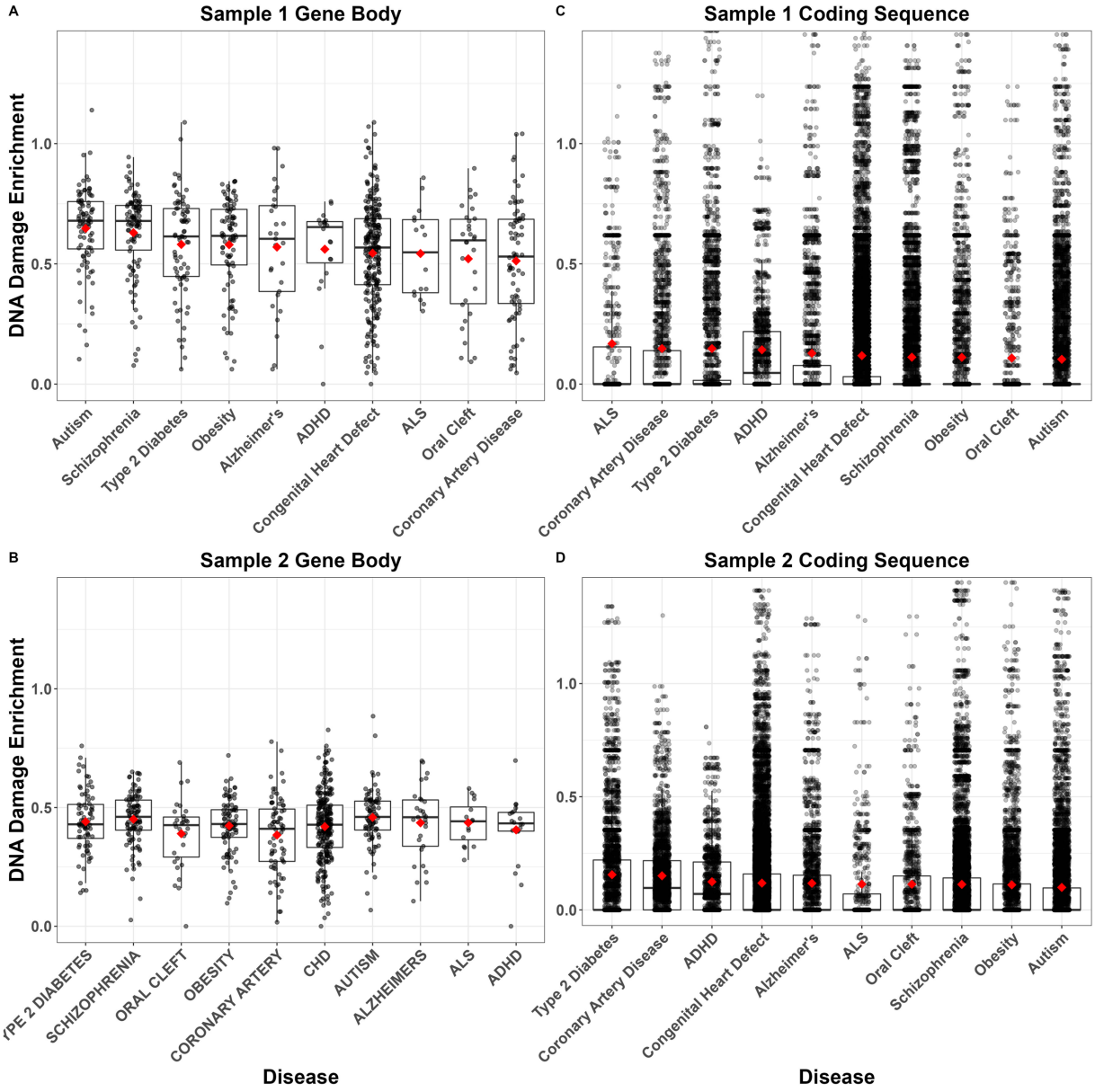


FIGURE 8
Average DNA damage enrichment following polycyclic aromatic hydrocarbon (PAH) treatment of GM12878 cells across each gene body **(A,B)** or coding sequence **(C,D)** in each of 10 disease-related gene sets. Enrichment values indicate the number of tXR-seq counts per gene/coding sequence, controlling for sequence length. Boxplots show median and interquartile range (IQR) with 1.5 IQR whiskers. Gene sets are ordered left–right within each panel from highest to lowest mean enrichment (red diamonds). Data come from two samples of treated cells from Li et al. (2017).

compared to CDS. Other factors may more strongly govern the mutability of protein coding sequences. For instance, we found that higher expression was associated with lower CDS mutation rate, while expression had no effect on the mutability of entire genes. This corroborates prior work showing that lowly expressed genes harbor more mutations (Pleasance et al., 2010), a phenomenon that might be attributable to transcription-coupled DNA repair (Fousteri and Mullenders, 2008; Hanawalt and Spivak, 2008).

Similarly, the effect of GC content was different for genes and CDS. Increased GC content was associated with fewer mutations in genes, but more mutations in CDS. These results are consistent with prior studies indicating that the effect of GC content on mutation rate varies over different genomic scales. GC content across entire genes may reflect higher order DNA structure, and increased GC content has been shown to correlate with decreased mutation rate at higher genomic scales (Wolfe et al., 1989; Hodgkinson and Eyre-Walker, 2011). CDS GC content, however, may more accurately reflect the effect of local GC content on mutability. Cytosines may experience higher mutation rates than other bases because methylated cytosines in CpG dinucleotides are vulnerable to deamination into thymidine. Furthermore, these mutations occur

at higher rates in regions with higher local GC content (Fryxell and Moon, 2005).

Our analyses of the relationship between gene characteristics and mutability show that sequence length is a strong driver, but not the only factor contributing to the elevated mutability of neurodevelopmental disease genes in this dataset. However, our models excluded several characteristics known to be associated with mutation. Studies of cancer driver genes have more comprehensively examined associations of gene characteristics with mutability (e.g., Lawrence et al., 2013; Gorlov et al., 2018). Additional gene and/or sequence characteristics examined in relation to mutability include open versus closed chromatin state (Ying et al., 2010; Schuster-Böckler and Lehner, 2012; Thurman et al., 2012), epigenetic markers (Coarfa et al., 2014), replication timing (earlier replicating regions have a lower mutation rate: Lang and Murray, 2011), di- and/or tri-nucleotide composition (Millar et al., 2002; Samocha et al., 2014), evolutionary conservation (Michaelson et al., 2012), and protein-DNA interactions identified *via* ChIP-seq (e.g., transcription factor binding) (Yang et al., 2018).

Although *de novo* mutation has previously been hypothesized as a pathway linking environmental exposures to increased NDD risk, particularly ASD (Kinney et al., 2010; Pugsley et al., 2021), this hypothesis has not been explicitly tested. For instance, epidemiologic research has linked many known carcinogens, such as air pollutants and heavy metals, with elevated ASD rates at the population level, but none of these studies include mutation data [reviewed by Pugsley et al. (2021)]. Addressing this limitation will require formal mediation analyses showing associations of environmental exposures with increased *de novo* mutation rates, which in turn result in elevated incidence of neurodevelopmental disease. We are unaware of any studies employing this type of mediation approach for environmental exposures, although the mediating role of *de novo* mutations has been investigated for paternal age (Gratten et al., 2016; Taylor et al., 2019). In the future, whole genome/exome sequencing studies of neurodevelopmental diseases such as ASD will need to collect data on environmental exposures to assess this hypothesis.

This work has several limitations. First, our findings that environmental chemicals may disproportionately mutate neurodevelopmental disease genes supports but is not an explicit test of the hypothesis described above. Second, the methods of gene set curation could bias our analyses comparing observed rates of exposure-induced mutations in disease gene sets with the mutation rate of control genes randomly sampled from the genome. To partially address this limitation, we created an online tool allowing researchers to query their own gene sets. Another limitation was our reliance on mutations called in cultured human iPSCs rather than *in vivo*. Because PAHs are a major air pollutant, we attempted to externally validate the results from PAH-exposed iPSCs by analyzing lung tumor mutations from humans living in highly polluted regions. However, future studies might better validate these results using animal models dosed with comparable levels of the environmental chemicals examined by Kucab et al. (2019). Additionally, the use of one iPSC line precludes an examination of potential genetic variability in gene-set mutation vulnerability. Future research should account for diverse genetic backgrounds in genomic instability in disease specific *de novo* mutations.

## Author's note

Custom gene lists can be queried using the algorithm we generated at www.environmentalmutation.com.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: Data are available from the primary sources cited. The SPARK gene variants are available to approved researchers through SFARI upon review. The code generated during this study are available on GitHub at: https://github.com/brennanhilton/environmental-mutation-calculator.

## Author contributions

BB, JS, WC, and BP contributed to conception and design of the study. BB and SZ performed the statistical analysis. BB wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

## Funding

This work was supported by the NIH grants: P30ES009089, R21ES032913, and R24ES029489.

## Acknowledgments

The authors would like to thank to Dr. Barbara Corneo of the Columbia Stem Cell Initiative provided valuable intellectual contribution to the conception of our study.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note



## Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2023.1106573/full#supplementary-material

# References


Abrahams, B. S., Arking, D. E., Campbell, D. B., Mefford, H. C., Morrow, E. M., Weiss, L. A., et al. (2013). SFARI gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol. Autism.* 4:36. doi: 10.1186/2040-2392-4-36

Association T.A. (2019). *Genetics [online]*. Washington, DC: The ALS Association.

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell* 173, 371–385. e318. doi: 10.1016/j.cell.2018.02.060

Beaty, T. H., Marazita, M. L., and Leslie, E. J. J. F. (2016). Genetic factors influencing risk to orofacial clefts: today's challenges and tomorrow's opportunities. *F1000Res*:5:2800. doi: 10.12688/f1000research.9503.1

Bellinger, D. C. (2012). A strategy for comparing the contributions of environmental chemicals and other risk factors to neurodevelopment of children. *Environ. Health Perspect.* 120, 501–507. doi: 10.1289/ehp.1104170

Boffetta, P., Jourenkova, N., and Gustavsson, P. (1997). Cancer risk from occupational and environmental exposure to polycyclic aromatic hydrocarbons. *Cancer Causes Control* 8, 444–472. doi: 10.1023/A:1018465507029

Cheng, M., Mei, B., Zhou, Q., Zhang, M., Huang, H., Han, L., et al. (2018). Computational analyses of obesity associated loci generated by genome-wide association studies. *PLoS One* 13:e0199987. doi: 10.1371/journal.pone.0199987

Churko, J. M., Lee, J., Ameen, M., Gu, M., Venkatasubramanian, M., Diecke, S., et al. (2017). Transcriptomic and epigenomic differences in human induced pluripotent stem cells generated from six reprogramming methods. *Nat. Biomed. Eng.* 1:826. doi: 10.1038/s41551-017-0141-6

Coarfa, C., Pichot, C. S., Jackson, A., Tandon, A., Amin, V., Raghuraman, S., et al. (2014). Analysis of interactions between the epigenome and structural mutability of the genome using Genboree workbench tools. *BMC Bioinformatics* 15, 1–12. doi: 10.1186/1471-2105-15-S7-S2

Crawley, J. N., Heyer, W.-D., and Lasalle, J. M. J. T. I. G. (2016). Autism and Cancer share risk genes, pathways, and drug targets. *Trends Genet.* 32, 139–146. doi: 10.1016/j.tig.2016.01.001

De Ligt, J., Willemsen, M. H., Van Bon, B. W., Kleefstra, T., Yntema, H. G., Kroes, T., et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* 367, 1921–1929. doi: 10.1056/NEJMoa1206524

Demontis, D., Walters, R. K., Martin, J., Mattheisen, M., Als, T. D., Agerbo, E., et al. (2019). Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.* 51:63. doi: 10.1038/s41588-018-0269-7

Edwards, S. C., Jedrychowski, W., Butscher, M., Camann, D., Kieltyka, A., Mroz, E., et al. (2010). Prenatal exposure to airborne polycyclic aromatic hydrocarbons and children's intelligence at 5 years of age in a prospective cohort study in Poland. *Environ. Health Perspect.* 118, 1326–1331. doi: 10.1289/ehp.0901070

Emberti Gialloreti, L., Mazzone, L., Benvenuto, A., Fasano, A., Garcia Alcon, A., Kraneveld, A., et al. (2019). Risk and protective environmental factors associated with autism Spectrum disorder: evidence-based principles and recommendations. *J. Clin. Med.* 8:217. doi: 10.3390/jcm8020217

Feliciano, P., Daniels, A. M., Snyder, L. G., Beaumont, A., Camba, A., Esler, A., et al. (2018). SPARK: a US cohort of 50,000 families to accelerate autism research. *Signal. Synapse* 97, 488–493. doi: 10.1016/j.neuron.2018.01.015

Fousteri, M., and Mullenders, L. H. (2008). Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. *Cell Res.* 18, 73–84. doi: 10.1038/cr.2008.6

Fromer, M., Pocklington, A. J., Kavanagh, D. H., Williams, H. J., Dwyer, S., Gormley, P., et al. (2014). De novo mutations in schizophrenia implicate synaptic networks. *Nature* 506:179. doi: 10.1038/nature12929

Fryxell, K. J., and Moon, W.-J. J. M. B. (2005). CpG mutation rates in the human genome are highly dependent on local GC content. *Mol. Biol. Evol.* 22, 650–658. doi: 10.1093/molbev/msi043

Gabel, H. W., Kinde, B., Stroud, H., Gilbert, C. S., Harmin, D. A., Kastan, N. R., et al. (2015). Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nat. Cell Biol.* 522, 89–93. doi: 10.1038/nature14319

Giri, M., Zhang, M., and Lü, Y. (2016). Genes associated with Alzheimer's disease: an overview and current status. *Clin. Interv. Aging* 11:665. doi: 10.2147/CIA.S105769

Gorlov, I. P., Pikielny, C. W., Frost, H. R., Her, S. C., Cole, M. D., Strohbehn, S. D., et al. (2018). Gene characteristics predicting missense, nonsense and frameshift mutations in tumor samples. *BMC Bioinformatics* 19, 1–14. doi: 10.1186/s12859-018-2455-0

Gratten, J., Wray, N. R., Peyrot, W. J., Mcgrath, J. J., Visscher, P. M., and Goddard, M. E. J. N. G. (2016). Risk of psychiatric illness from advanced paternal age is not predominantly from de novo mutations. *Nature Genet.* 48, 718–724. doi: 10.1038/ng.3577

Hanawalt, P. C., and Spivak, G. (2008). Transcription-coupled DNA repair: two decades of progress and surprises. *Nat. Rev. Mol. Cell Biol.* 9, 958–970. doi: 10.1038/nrm2549

Hodgkinson, A., and Eyre-Walker, A. J. N. R. G. (2011). Variation in the mutation rate across mammalian genomes. *Nature Rev. Genet.* 12, 756–766. doi: 10.1038/nrg3098

Iossifov, I., O'roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 515:216. doi: 10.1038/nature13908

Jedrychowski, W. A., Perera, F. P., Camann, D., Spengler, J., Butscher, M., Mroz, E., et al. (2015). Prenatal exposure to polycyclic aromatic hydrocarbons and cognitive dysfunction in children. *Environ. Sci. Pollut. Res.* 22, 3631–3639. doi: 10.1007/s11356-014-3627-8

Jin, S. C., Homsy, J., Zaidi, S., Lu, Q., Morton, S., Depalma, S. R., et al. (2017). Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nature Genet.* 49, 1593–1601. doi: 10.1038/ng.3970

Kim, K.-H., Jahan, S. A., Kabir, E., and Brown, R. J. (2013). A review of airborne polycyclic aromatic hydrocarbons (PAHs) and their human health effects. *Environ. Int.* 60, 71–80. doi: 10.1016/j.envint.2013.07.019

King, I. F., Yandava, C. N., Mabb, A. M., Hsiao, J. S., Huang, H.-S., Pearson, B. L., et al. (2013). Topoisomerases facilitate transcription of long genes linked to autism. *Nat. Cell Biol.* 501, 58–62. doi: 10.1038/nature12504

Kinney, D. K., Barch, D. H., Chayka, B., Napoleon, S., and Munir, K. M. J. M. H. (2010). Environmental risk factors for autism: do they help cause de novo genetic mutations that contribute to the disorder? *Med. Hypotheses* 74, 102–106. doi: 10.1016/j.mehy.2009.07.052

Kriek, E., Rojas, M., Alexandrov, K., and Bartsch, H. (1998). Polycyclic aromatic hydrocarbon-DNA adducts in humans: relevance as biomarkers for exposure and cancer risk. *Mutat. Res.* 400, 215–231. doi: 10.1016/S0027-5107(98)00065-7

Kucab, J. E., Zou, X., Morganella, S., Joel, M., Nanda, A. S., Nagy, E., et al. (2019). A compendium of mutational signatures of environmental agents. *Cells* 177:e816, 821–836.e16. doi: 10.1016/j.cell.2019.03.001

Landrigan, P. J. (2010). What causes autism? Exploring the environmental contribution. *Curr. Opin. Pediatr.* 22, 219–225. doi: 10.1097/MOP.0b013e328336eb9a

Lang, G. I., and Murray, A. W. J. G. B. (2011). Mutation rates across budding yeast chromosome VI are correlated with replication timing. *Genome Biol. Evol.* 3, 799–811. doi: 10.1093/gbe/evr054

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nat. Cell Biol.* 499, 214–218. doi: 10.1038/nature12213

Li, W., Hu, J., Adebali, O., Adar, S., Yang, Y., Chiou, Y.-Y., et al. (2017). Human genome-wide repair map of DNA damage caused by the cigarette smoke carcinogen benzo[a]pyrene. *Proc. Natl. Acad. Sci. U. S. A.* 114, 6752–6757. doi: 10.1073/pnas.1706021114

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nat. Cell Biol.* 518, 197–206. doi: 10.1038/nature14177

Mahajan, A., Go, M. J., Zhang, W., Below, J. E., Gaulton, K. J., Ferreira, T., et al. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genet.* 46, 234–244. doi: 10.1038/ng.2897

Mclaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., et al. (2016). The ensembl variant effect predictor. *Genome Biol.* 17:122. doi: 10.1186/s13059-016-0974-4

Michaelson, J. J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., et al. (2012). Whole-genome sequencing in autism identifies hot spots for De novo germline mutation. *Cell (Cambridge, MA)* 151, 1431–1442. doi: 10.1016/j.cell.2012.11.019

Millar, C. B., Guy, J., Sansom, O. J., Selfridge, J., Macdougall, E., Hendrich, B., et al. (2002). Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice. *Science* 297, 403–405. doi: 10.1126/science.1073354

Nikpay, M., Goel, A., Won, H.-H., Hall, L. M., Willenborg, C., Kanoni, S., et al. (2015). A comprehensive 1000 genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* 47:1121. doi: 10.1038/ng.3396

Perera, F., and Herbstman, J. J. R. T. (2011). Prenatal environmental exposures, epigenetics, and disease. *Reprod. Toxicol.* 31, 363–373. doi: 10.1016/j.reprotox.2010.12.055

Perera, F. P., Rauh, V., Whyatt, R. M., Tsai, W.-Y., Tang, D., Diaz, D., et al. (2006). Effect of prenatal exposure to airborne polycyclic aromatic hydrocarbons on neurodevelopment in the first 3 years of life among inner-city children. *Environ. Health Perspect.* 114, 1287–1292. doi: 10.1289/ehp.9084

Pfeifer, G. (2006). "Mutagenesis at methylated CpG sequences" in *DNA Methylation: Basic Mechanisms*. eds. W. Doerfler and P. Böhm (Berlin: Springer), 259–281.

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., Mcbride, D. J., Humphray, S. J., Greenman, C. D., et al. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 463, 191–196. doi: 10.1038/nature08658

Pugsley, K., Scherer, S. W., Bellgrove, M. A., and Hawi, Z. J. M. P. (2021). Environmental exposures associated with elevated risk for autism spectrum disorder may augment the burden of deleterious de novo mutations among probands. *Mol. Psychiatry.* 27, 710–730. doi: 10.1038/s41380-021-01142-w

Qi, H., Dong, C., Chung, W. K., Wang, K., and Shen, Y. J. H. M. (2016). Deep genetic connection between Cancer and developmental disorders. *Hum. Mutat.* 37, 1042–1050. doi: 10.1002/humu.23040

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., et al. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Symp. Ser.* 44, W160–W165. doi: 10.1093/nar/gkw257

Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* 380, 1674–1682. doi: 10.1016/S0140-6736(12)61480-9

Rauh, V. A., and Margolis, A. E. (2016). Research review: environmental exposures, neurodevelopment, and child mental health–new paradigms for the study of brain and behavioral effects. *J. Child Psychol. Psychiatry* 57, 775–793. doi: 10.1111/jcpp.12537

Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., Mcgrath, L. M., et al. (2014). A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* 46, 944–950. doi: 10.1038/ng.3050

Schuster-Böckler, B., and Lehner, B. J. N. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nat. Cell Biol.* 488, 504–507. doi: 10.1038/nature11273

Shields, P. G., and Harris, C. C. (2000). Cancer risk and low-penetrance susceptibility genes in gene-environment interactions. *J. Clin. Oncol.* 18, 2309–2315. doi: 10.1200/JCO.2000.18.11.2309

Sugino, K., Hempel, C. M., Okaty, B. W., Arnson, H. A., Kato, S., Dani, V. S., et al. (2014). Cell-type-specific repression by methyl-CpG-binding protein 2 is biased toward long genes. *J. Neurosci.* 34, 12877–12883. doi: 10.1523/JNEUROSCI.2674-14.2014

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., et al. (2019). COSMIC: the catalogue of somatic mutations in Cancer. *Nucleic Acids Symp. Ser.* 47, D941–D947. doi: 10.1093/nar/gky1015

Taylor, J. L., Debost, J.-C. P., Morton, S. U., Wigdor, E. M., Heyne, H. O., Lal, D., et al. (2019). Paternal-age-related de novo mutations and risk for five disorders. *Nat. Commun.* 10, 1–9. doi: 10.1038/s41467-019-11039-6

Team, R. C. (2018). *R: A Language and Environment For Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; (*2018*).

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., et al. (2012). The accessible chromatin landscape of the human genome. *Nat. Cell Biol.* 489, 75–82. doi: 10.1038/nature11232

Tran, N. Q. V., and Miyake, K. J. I. J. O. G. (2017). Neurodevelopmental disorders and environmental toxicants: epigenetics as an underlying mechanism. *Comp. Funct. Genom.* 2017, 1–23. doi: 10.1155/2017/7526592

Von Ehrenstein, O. S., Aralis, H., Cockburn, M., and Ritz, B. (2014). In utero exposure to toxic air pollutants and risk of childhood autism. *Epidemiology* 25, 851–858. doi: 10.1097/EDE.0000000000000150

Wang, Q., Chen, R., Cheng, F., Wei, Q., Ji, Y., Yang, H., et al. (2019). A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. *Nat. Neurosci.* 22:691. doi: 10.1038/s41593-019-0382-7

Watanabe, K., Taskesen, E., Van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* 8:1826. doi: 10.1038/s41467-017-01261-5

Whyatt, R., Bell, D., Jedrychowski, W., Santella, R., Garte, S., Cosma, G., et al. (1998). Polycyclic aromatic hydrocarbon-DNA adducts in human placenta and modulation by CYP1A1 induction and genotype. *Carcinogenesis* 19, 1389–1392. doi: 10.1093/carcin/19.8.1389

Wolfe, K. H., Sharp, P. M., and Li, W.-H. J. N. (1989). Mutation rates differ among regions of the mammalian genome. *Nat. Cell Biol.* 337, 283–285. doi: 10.1038/337283a0

Xu, B., Ionita-Laza, I., Roos, J. L., Boone, B., Woodrick, S., Sun, Y., et al. (2012). De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat. Genet.* 44:1365. doi: 10.1038/ng.2446

Yang, J., Wei, X., Tufan, T., Kuscu, C., Unlu, H., Farooq, S., et al. (2018). Recurrent mutations at estrogen receptor binding sites alter chromatin topology and distal gene expression in breast cancer. *GBC* 19, 1–15. doi: 10.1186/s13059-018-1572-4

Ying, H., Epps, J., Williams, R., and Huttley, G. J. M. B. (2010). Evidence that localized variation in primate sequence divergence arises from an influence of nucleosome placement on DNA repair. *Mol. Biol. Evol.* 27, 637–649. doi: 10.1093/molbev/msp253

Yu, X.-J., Yang, M.-J., Zhou, B., Wang, G.-Z., Huang, Y.-C., Wu, L.-C., et al. (2015). Characterization of somatic mutations in air pollution-related lung cancer. *EBioMedicine* 2, 583–590. doi: 10.1016/j.ebiom.2015.04.003

Zhang, F., and Lupski, J. R. J. H. M. G. (2015). Non-coding genetic variants in human disease. *Hum. Mol. Genet.* 24, R102–R110. doi: 10.1093/hmg/ddv259

Zhou, X., Feliciano, P., Wang, T., Shu, C., Astrovskaya, I., Hall, J., et al. (2021). Integrating de novo and inherited variants in over 42,607 autism cases identifies mutations in new moderate risk genes. *medRxiv*. doi: 10.1101/2021.10.08.21264256 [Epub ahead of preprint].

Zylka, M. J., Simon, J. M., and Philpot, B. D. J. N. (2015). Gene length matters in neurons. *Signal. Synapse* 86, 353–355. doi: 10.1016/j.neuron.2015.03.059