

A role for gene-environment interactions in Autism Spectrum Disorder is suggested by variants in genes regulating exposure to environmental factors

João Xavier Santos^{1,2}, Célia Rasga^{1,2}, Ana Rita Marques^{1,2}, Hugo F. M. C. Martiniano^{1,2},
Muhammad Asif^{1,2}, Joana Vilela^{1,2}, Guiomar Oliveira^{3,4,5}, Astrid Moura Vicente^{1,2*}

* Correspondence: Astrid M. Vicente, Instituto Nacional de Saúde Doutor Ricardo Jorge, Avenida Padre Cruz, 1649-016 Lisboa, Portugal. Email: astrid.vicente@insa.min-saude.pt

¹Instituto Nacional de Saúde Doutor Ricardo Jorge, Avenida Padre Cruz, 1649-016 Lisboa, Portugal

²BioISI: Biosystems & Integrative Sciences Institute, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal

³Unidade de Neurodesenvolvimento e Autismo (UNDA), Serviço do Centro de Desenvolvimento da Criança, Centro de Investigação e Formação Clínica, Hospital Pediátrico, Centro Hospitalar e Universitário de Coimbra, Coimbra, Portugal

⁴Institute for Biomedical Imaging and Life Sciences, Faculty of Medicine, Universidade de Coimbra, Coimbra, Portugal

⁵University Clinic of Pediatrics, Faculty of Medicine, University of Coimbra, Portugal

Abstract

Introduction: Autism Spectrum Disorder (ASD) is a clinically heterogeneous neurodevelopmental disorder defined by deficits in social communication and interaction and repetitive and stereotyped interests and behaviors. ASD heritability estimates of 50-83% support a strong role of genetics in its onset, with large sequencing studies reporting a high burden of rare potentially pathogenic copy number variants (CNVs) and single nucleotide variants (SNVs) in affected subjects. Recent data strongly suggests that prenatal to postnatal exposure to ubiquitous environmental factors (e.g. environmental toxins, medications and nutritional factors) contribute to ASD risk. Detoxification processes and physiological permeability barriers (i.e. blood-brain barrier, placenta and respiratory cilia) are crucial in regulating exposure and response to external agents during early development. Thus, the objectives of this study were: 1) to find genes involved in detoxification and regulation of barriers permeability with a high load of relevant CNVs and SNVs in ASD subjects; 2) to explore interactions between the identified genes and environmental factors relevant for the disorder.

Material and Methods: Through literature and databases review we searched for genes involved in detoxification and regulation of barriers permeability processes. Genetic data collected from large datasets of subjects with ASD (Autism Genome Project (AGP), Simmons Simplex Collection (SSC), and Autism Sequencing Consortium (ASC)) was used to identify potentially pathogenic variants targeting detoxification and barrier genes. Data from control subjects without neuropsychiatric disorder history was used for comparison purposes. The Comparative Toxicogenomics Database (CTD) was interrogated to identify putatively relevant gene-environment interactions reported in humans throughout the literature.

Results: We compiled a list of 519 genes involved in detoxification and regulation of permeability barriers. The analysis of AGP and SSC data resulted in the identification of 7 genes more-frequently targeted by CNVs in ASD subjects from both datasets, after Bonferroni correction for multiple testing

(AGP: $P < 3.5211 \times 10^{-4}$; SSC: $P < 4.587 \times 10^{-4}$). Moreover, 8 genes were exclusively targeted by CNVs from ASD subjects. Regarding SNVs analyses using the ASC dataset, we found 40 genes targeted by potentially pathogenic loss-of-function and/or missense SNVs exclusive to 6 or more cases. The CTD was interrogated for interactions between 55 identified genes and 54 terms for unique chemicals associated with the disorder. A total of 212 gene-environment interaction pairs, between 51/55 (92.7%) genes and 38/54 (70.4%) chemicals, putatively relevant for ASD, were discovered. *ABCB1*, *ABCG2*, *CYP2C19*, *GSTM1*, *CYP2D6*, and *SLC3A2* were the genes that interacted with more chemicals, while valproic acid, benzo(a)pyrene (b(a)p), bisphenol A, particulate matter and perfluorooctane sulfonic acid (PFOS) were the top chemicals.

Discussion: The identified genes code for functionally diverse proteins, ranging from enzymes that increase the degradability of xenobiotics (CYP450s, UGTs and GSTs), to transporters (ABCs and SLCs), proteins that regulate the correct function of barriers (claudins and dyneins) and placental hormones. The identified gene-environment interactions may reflect the fact that some genes and chemicals are understudied and that the potential neurotoxicity of many substances is unreported. We suggest that environmental factors can have pathogenic effects when individuals carry variants targeting these genes and discuss the potential mechanisms by which these genes can influence ASD risk.

Conclusion: We reinforce the hypothesis that gene-environment interactions are relevant, at least, for a subset of ASD cases. Given that no treatment exists for the pathology, the identification of relevant modifiable exposures can contribute to the development of preventive strategies for health management policies in ASD.

Introduction

Autism spectrum disorder (ASD) is an early onset neurodevelopmental disorder characterized by deficits on social communication and interaction and repetitive and stereotyped interests and behaviors (1,2). These two core features often appear associated with other symptoms, such as intellectual disability, speech delays and attention deficits (2), originating a phenotypically heterogeneous spectrum. In recent years, ASD prevalence estimates have been on the rise, with values of 1-5% being reported in developed countries (3,4). Regarding gender distribution, a male skewness is consistently reported, with a male-to-female ratio of 4:1 to 3:1 being assumed (5,6).

The advent of high-throughput sequencing platforms has identified multiple rare *de novo* or inherited high-effect copy number variants (CNVs) (7–11) and gene-disrupting single nucleotide variants (SNVs) (12–14) associated with pathology. Low-effect common variants are also relevant (15). ASD onset can also happen in a syndromic context, comorbid to pathologies like Fragile X syndrome, Rett Syndrome and Tuberous sclerosis (16). Despite this, most cases of the disorder remain idiopathic. Moreover, family studies, particularly monozygotic and dizygotic twin studies, report ASD heritability estimates of 50-83% (17–19), which clearly leave space for a role of non-genetic factors in the disorder. Current research suggests that ASD is likely explained by a multifactorial etiology that includes genetic and non-genetic risk factors, which interact in a cumulative way to reach a threshold for onset (15,17,18,20).

Since the beginning of the 2000s, studies started to focus on prenatal to postnatal exposure to environmental agents as non-genetic risk factors. Early development is a recognized window of susceptibility to external cues, which can have detrimental effects, potentially modulating the neuropathological events that lead to the onset of the disorder (21). To infer associations between

early exposures and ASD risk, studies either adopt prospective or retrospective designs, resorting to one of four different strategies to measure the exposures: 1) collection of biological samples (blood (22), urine (23) and naturally-shed deciduous teeth (24)) from the pregnant women and/or their offspring to quantify multiple analytes; 2) application of questionnaires answered by the mothers of the children to evaluate self-reported exposures (25,26); 3) geo-referencing studies for the mapping of sources of environmental toxins (e.g. landfill sites (27) and agricultural fields (28)) or understanding of air-quality patterns (29,30); 4) analysis of medical and prescription records to evaluate medications and supplements intake (31).

Multiple associations between environmental exposures and ASD have been reported and, while some, such as vaccination and thimerosal exposure, have been contradicted (32), others warrant further investigation. Meta-analyses, as the ones published by Rossignol *et al* (33) and Modabbernia *et al* (34), are excellent tools to identify environmental factors more consistently associated with the disorder. In this paper we divide these putatively relevant environmental factors into three major classes: 1) environmental toxins; 2) medications; 3) nutritional factors. Environmental toxins include air pollutants (nitrogen dioxide (NO₂), ozone (O₃), particulate matter (PM) and polycyclic aromatic hydrocarbons (PAHs)) (29,30,35,36), bisphenol A (BPA) (37,38), heavy metals (lead, manganese and mercury) (24,39) pesticides (28,40), phthalates (41,42), polybrominated diphenyl ethers (PBDEs) (22,26), polychlorinated biphenyls (PCBs) (22,26,43) and perfluorinated compounds (PFCs) (22,44). Medications include teratogens (valproic acid (45,46) and thalidomide (47,48)) and antidepressants (49,50). Nutritional factors include folic acid (51,52) and vitamin D.

Exposure to most of the referred toxins is ubiquitous, since they are present in environment, everyday household and industrial products and food. PBDEs, PCBs, PFCs and some heavy metals and pesticides are persistent organic pollutants (POPs), being resistant to degradation through chemical or biological processes (53), which increases their risk of bioaccumulation. Contrary, BPA and phthalates are non-POPs and are, thus, rapidly metabolized (53). Nonetheless, given the virtually ubiquitous exposure, they are still relevant. Most of these toxins have neurotoxic properties (63) and many (e.g. BPA, phthalates, pesticides, PCBs, PBDEs and lead) are recognized endocrine-disrupting chemicals (EDCs) (53–55). EDCs, when ingested or absorbed, can mimic estrogens, androgens and thyroid hormones, acting as agonists or antagonists to hormone receptors, potentially leading to endocrine dysregulations. Regarding medications, thalidomide, primarily sold as a sedative, was widely used to alleviate morning sickness in pregnant women during the 50s, while valproate is prescribed for epilepsy and bipolar disorder. Periods of critical vulnerability these two teratogens are proposed to occur early in pregnancy, concomitant with neural tube closure at 28th day of pregnancy (47). Antidepressants are frequently used to treat maternal depression, with selective serotonin reuptake inhibitors (SSRIs) being the most prescribed ones. Contrary to the other factors, it is deficient gestational or at birth levels of nutritional factors that seem to increase the risk of developing ASD. This is not unexpected, since folic acid is used as a supplement by pregnant women in order to prevent neural tube defects in the developing fetus, and vitamin D is a steroid hormone that plays a crucial role in calcium and phosphorous metabolism. Circulatory levels of 25-hydroxyvitamin D – 25(OH)D – the precursor of the active form of vitamin D, are usually quantified to assess vitamin D deficiency.

Physiological permeability barriers, such as the placenta, the blood-brain barrier (BBB) and the motile cilia of the human airway epithelia are crucial in limiting the exposure of the organism to chemicals. The BBB, which is formed by brain endothelial cells, functions as a semipermeable

membrane to various neurotoxins, thanks to the tight junctions between these cells (56). The placenta establishes an interface between the mother and the developing fetus that, among other functions, regulates transfer of nutrients and waste products between maternal and fetal plasma (57). Finally, the respiratory epithelium serves as a barrier to potential xenobiotics by the action of mucociliary clearance carried by the cilia (58). Most of the referred environmental factors are able to cross these structures (59–64). Meanwhile, detoxification pathways involve a series of enzymatic reactions that act to detoxify xenobiotics and remove them from cells. Thus, these structures and processes are of crucial importance during neurodevelopment, when the organism is particularly vulnerable to exogenous influences.

The objectives of this study were: 1) to identify if genes involved in detoxification and regulation of barriers permeability are targeted by potentially pathogenic CNVs and SNVs in subjects with ASD; 2) to understand if such genes interact with environmental factors relevant for the disorder. We hypothesize that environmental factors can have pathogenic effects in genetically susceptible individuals. For such, we questioned large population datasets composed by individuals with ASD for the presence of variants targeting genes involved in detoxification and regulation of barriers permeability, and compared the results with control populations without history of neuropsychiatric disorder. We further explored interactions between such genes and the environmental factors potentially relevant for ASD.

Materials and Methods

Population datasets

The primary ASD dataset used for CNVs identification was obtained from the Autism Genome Project (AGP) consortium (N=2446 ASD cases) (8,9) (table 1). As control, data from two population cohorts (Cooper *et al* (65) and Shaikh *et al* (66)) composed by subjects without clinical history of neuropsychiatric disease was used (N=9649). From Shaikh *et al*, 694 African-American and 12 Asian-American individuals were not considered (table 1). We had no access to the ethnicity of Cooper *et al* subjects, and thus all individuals were analyzed. CNV data from these control datasets are publicly available through the Database of Genomic Variants (DGV) (67). Additionally, for results validation, data from the Simons Simplex Collection (SSC) (N=1124 ASD cases), a resource of the Simons Foundation Autism Research (SFARI), was used (table 1) (10). All these populations were genotyped through various Illumina platforms. AGP population was genotyped using Illumina 1M SNP arrays, control populations were genotyped using 550k to 1.2 platforms and SSC population was genotyped using 1Mv3 or 1Mv1 arrays.

For SNVs identification we used data from 3426 subjects (2674 cases and 752 controls) with, at least, 80% Caucasian ethnicity, genotyped through the Autism Sequencing Consortium (ASC) (68). ASC exome-sequencing data was obtained from dbGaP portal (accession code: phs000294.v4.p3). Samples were sequenced at the Broad Institute.

The Autism Diagnostic Interview-Revised (ADI-R) and Autism Diagnostic Observation Schedule (ADOS) tools were applied for assessment of ASD diagnosis in AGP, SSC and ASC datasets.

Table 1: Data regarding characteristics of the ASD- and control population datasets used in this study. *AGP* – Autism Genome Project; *SSC* – Simons Simplex Collection; *ASC* – Autism Sequencing Consortium

Dataset	Population type	Number of analyzed subjects	Ethnicity (%)	References
AGP	Simplex and multiplex families	2446 subjects diagnosed with ASD	Caucasians (87.8%); African-Americans (3.1%); Latin-Americans (6.3%); Asians (2.8%); Mixed (0.1%)	(8,9)
SSC	Simplex families	1124 subjects diagnosed with ASD	Caucasians (78.6%); African-Americans (3.8%); Latin-Americans (4%); Asians (4.3%); Mixed (9.3%)	(10)
Cooper <i>et al</i> 2011 control cohort	Unrelated subjects	8329 subjects without history of neurological disease	Caucasians (81.2%); African-Americans (2%); individuals of other or unknown ancestry (16.5%)	(65)
Shaikh <i>et al</i> 2009 control cohort	Unrelated subjects	1320 subjects without history of neurological disease	Caucasians (100%) (706 African- or Asian-Americans were not considered)	(66)
ASC	ASD-families and unrelated control subjects	2674 subjects with ASD and 752 controls	Caucasians (100%)	(68)

Identification of genes involved in detoxification and regulation of barriers permeability processes

To define a list of genes involved in detoxification processes and regulation of blood-brain barrier, placenta or respiratory cilia permeability we performed a literature review, by querying PubMed with the following terms: “detoxification”, “placenta”, “blood-brain barrier” and “respiratory cilia”. To perform this list we limited our query output to reviews, as these offer an excellent compendium for information regarding pathways organization. Additionally, publicly available databases, such as *The Human Protein Atlas* (69) and *Toxin and Toxin-Target Database (T3DB)* (70,71) were also used. *The Human Protein Atlas* contains protein expression data derived from the annotation of immunohistochemical staining of specific cell populations in human tissues and organs, including brain and placenta, allowing for the detection of genes expressed in these organs. *T3DB* provides mechanisms of toxicity and target proteins for a wide variety of toxins, allowing us select genes that code for proteins targeted by relevant environmental factors.

CNVs quality control and characterization

For this study, only genic variants were considered. Variants from the AGP project were initially predicted using three different algorithms (*QuantisNP*, *PennCNV* and *iPattern*) (8). As a quality control, variants predicted by only one algorithm or that corresponded to amplification artifacts resultant from the used methodology were excluded. Variants that did not pass quality filters, but were experimentally validated by real time quantitative PCR, as previously described by Pinto *et al* (8), were analyzed. This way, only genic high-confidence, or experimentally-validated, variants from AGP subjects with a diagnosis of ASD were considered. Considering the SSC dataset, no quality

control was performed, since the available data already only contained high-quality rare variants (10). In this, variants were defined as rare when up to 50% of their sequence overlapped regions present at >1% frequency in DGV (10).

AGP, SSC and control datasets were analyzed for the frequency with which the studied genes were targeted by CNVs, and putatively relevant genes were then divided into two categories: genes exclusively targeted by CNVs in ASD patients and genes more-frequently targeted by CNVs in ASD patients, after Bonferroni correction for multiple testing. A resume of the used methodology is shown in figure 1. Genes exclusively targeted by CNVs from ASD-subjects may be relevant, even in very low-frequencies, since ultra-rare mutations unique to patients with the disorder are known to play a role in ASD etiology (7–9). Meanwhile, genes more-frequently found in CNVs from ASD-patients, when compared to control subjects, may reveal a better proxy for the effect of exposure to environmental factors (in a gene-environment interaction model for ASD, subjects that carry CNVs targeting the genes studied in this article only develop ASD when exposed to an external trigger; subjects that have CNVs targeting these genes, but are not exposed, show a normal development).

SNVs quality control and variants prioritization

Regarding exome-sequencing data, quality control was done by filtering out samples with depth <8 and genotype quality ≤ 20 and by excluding variants with missingness >10%. Very common variants (MAF>5%), based on frequencies from The Genome Aggregation Database (gnomAD) (72), were not considered. Variant annotation was done using Variant Effect Predictor (VEP) tool from Ensembl (73) which, among other information, allowed us to know the impact predictions attributed to nonsynonymous mutations by SIFT (74) and PolyPhen-2 (75) *in silico* tools.

For variants prioritization, only loss-of-function (l-o-f) and missense variants, respectively defined as having high and moderate impact by VEP, were considered. Concerning moderate impact variants, only the ones defined as deleterious by SIFT and probably or possibly damaging by PolyPhen-2 were kept. L-o-f variants include frameshift mutations, loss of start or stop codons, gain of a stop codon and mutations in splice donor and acceptor sites. This process is resumed in figure 1. To further refine this prioritization, we ranked the remaining l-o-f and missense variants according to their frequency in ASC cases and controls, establishing six ranks (A-F). Variants ranked highly (rank A) were the ones exclusively present in 6 or more cases, while the ones ranked lower (rank F) were those more-frequently present in controls. After this, in order to identify genes with high burden of potentially pathogenic SNVs, we counted the numbers of l-o-f and damaging and deleterious missense SNVs, ranked by class, targeting each gene.

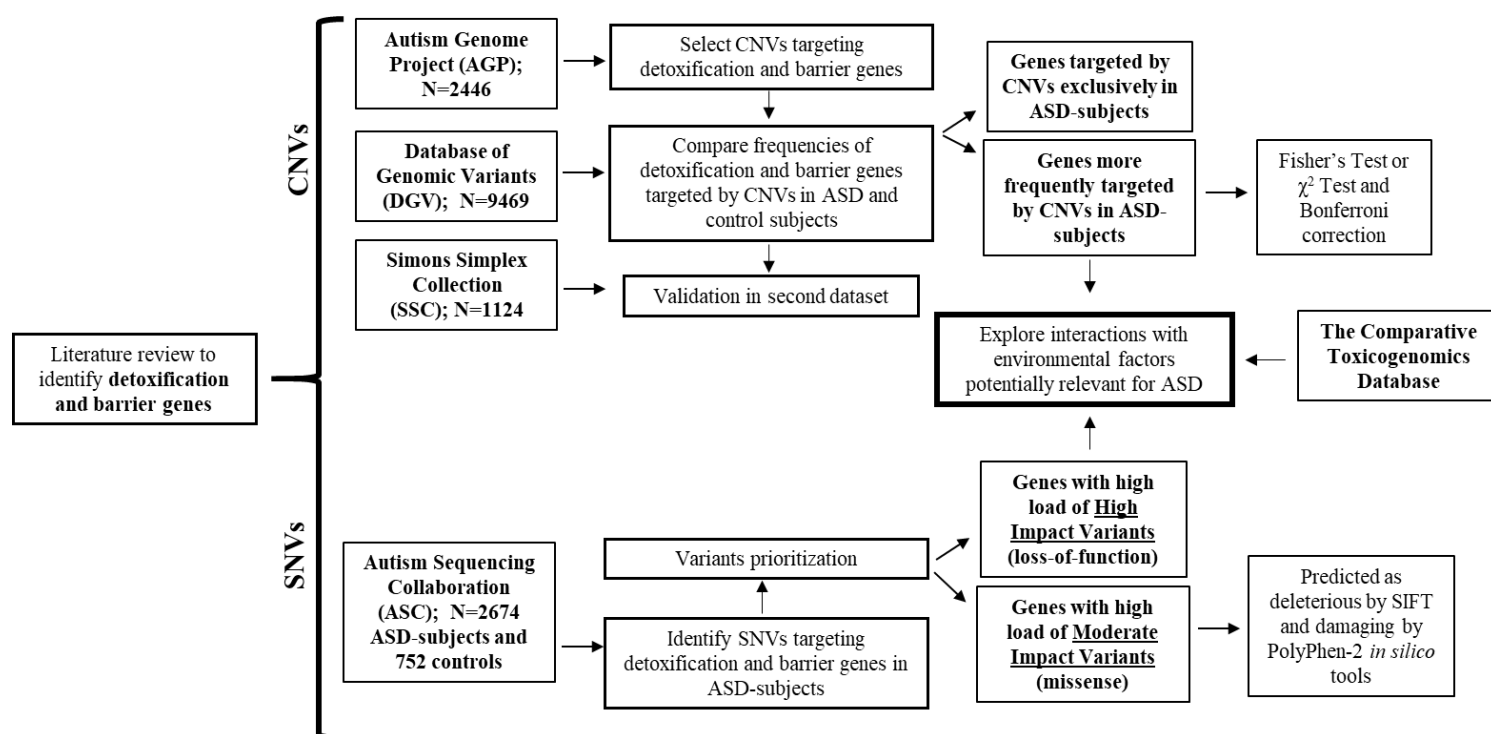


Figure 1: Methodological flowchart resuming the workflow of this study. Large population datasets were used for discovery of detoxification and barrier permeability genes with a high burden of potentially pathogenic CNVs and SNVs in individuals with ASD. The Comparative Toxicogenomics Database was used to explore interactions between the identified genes and environmental factors relevant for the disorder.

Identification of interactions between detoxification and barrier genes and environmental factors potentially relevant for ASD

To identify gene-environment interactions potentially relevant for the disorder we resorted to The Comparative Toxicogenomics Database (CTD) (figure 1), a manually curated platform that provides information about interactions between chemicals and gene products (76). The CTD organizes chemicals in classes, with broader classes consecutively branching into less inclusive classes. Furthermore, it lists all published references that support each interaction. As of January, 2019, 1854704 curated chemical-gene interactions between 46990 unique genes and 13015 unique chemicals in 590 organisms were recorded by the CTD.

We uploaded the HUGO Gene Nomenclature Committee (HGCN) gene symbols of the studied genes to the CTD query interface. All symbols were found by the CTD. The output files were then manually interrogated for the presence of the MeSH IDs correspondent to each chemical. The MeSH ID is a unique identifier assigned to each chemical by the Medical Subjects Headings (<https://www.nlm.nih.gov/mesh/>). Only interactions observed in *Homo sapiens* were considered. We surveyed a total of 54 individual chemicals putatively relevant for ASD (table S1), organized into environmental toxins, medications and nutritional factors. The abbreviations used for each chemical also listed in table S1. A similar approach has been used by Carter and Blizard, where the authors queried 206 candidate genes for autism for interactions with chemicals relevant for ASD (77).

Statistical Analysis

Regarding CNVs identification, statistical analysis (Fisher Test or χ^2 test) was performed using open source programming language and software environment R. When necessary, the Bonferroni test for multiple comparisons was applied.

Results

Identification of genes involved in detoxification and regulation of barriers permeability processes

Through systematic literature review we identified 519 genes involved in detoxification and regulation of barriers permeability processes (table 2). Some of these genes overlap in their roles. For example, *GSTP1* is a main glutathione S-transferase that is expressed at the BBB. Meanwhile, *ACHE* is a gene involved in response to pesticides that is also highly expressed at the placenta. For a full list of the 519 genes identified see figures S1a and S1b.

Table 2: Number of detoxification and permeability regulation of the BBB, placenta or respiratory cilia genes identified through literature review and access to online databases. In italic and parentheses is represented the number of genes solely involved in detoxification or permeability regulation of each of the barriers.

Class of genes	Number of genes (Genes unique to the process or barrier)	Class of genes	Number of genes
Detoxification genes	297 (<i>284</i>)	Detoxification and BBB genes	7
Blood-brain barrier genes	91 (<i>81</i>)	Detoxification and placenta genes	5
Placenta genes	120 (<i>112</i>)	BBB and placenta genes	2
Respiratory cilia genes	27 (<i>27</i>)	Detoxification, BBB and placenta genes	1

Identification of detoxification and regulation of barrier permeability genes targeted by CNVs from AGP and SSC ASD-subjects

Using data from AGP, we searched for CNVs from 2446 patients with ASD targeting genes involved in detoxification and regulation of barriers permeability. Moreover, we calculated the frequency with each gene is targeted by CNVs in ASD and control subjects. From the 519 genes identified, 173 (33.3%) were targeted by CNVs from 555/2446 (22.7%) ASD-subjects (table 3). Of these 173 genes, 31 (17.9%) were exclusively targeted by CNVs from 62/2446 (2.5%) ASD-subjects, while 23 (13.3%) were more frequently-targeted by CNVs from 261/2446 (10.7%) ASD-subjects after Bonferroni correction for multiple testing ($P < 3.5211 \times 10^{-4}$) (table 3).

Among the genes exclusively targeted by CNVs from ASD-subjects, *STS* was the most frequent (n=12, F=0.50%), followed by *CYP2D6* (n=9, F=0.37%) and *ARSF* (n=5, F=0.20%) (table 4). The genes more-frequently targeted by CNVs in ASD-subjects include nine genes coding UDP-glucuronosyltransferases (UGTs), three genes coding for glutathione S-transferases (GSTs), two

genes coding for members of the ATP-binding cassette (ABC) transporters family and two genes coding for cytochrome P450 family members (table 5). For the additional frequencies for the genes targeted by CNVs in both AGP and control datasets that do not reach statistical significance after correction for multiple testing go to table S2, from supplementary data.

For results validation, we applied the same methodology to the data from the SSC. From the 31 genes found exclusively in CNVs from AGP-subjects, 8(25.8%) were also exclusively found in CNVs from SSC-subjects. Once more, *STS* (n=7, F=0.62%) and *CYP2D6* (n=5, F=0.45%) were the most frequently targeted genes. The other 6 genes common to both datasets were *CLDN3*, *ARSF*, *GUSB*, *CYP2R1*, *SLC3A2* and *SULT2B1* (table 4). When considering the 23 genes more frequently found in CNVs from AGP-subjects, 7 (30.4%) were also more frequently-targeted by CNVs in SSC subjects, after correction for multiple testing ($P < 4.587 \times 10^{-4}$) (table 5). These were *CSH1*, *MAGEA8*, *CYP4X1*, *CHST5*, *CSH2*, *GH2* and *ABCC1*. Interestingly, two genes coding for glutathione S-transferases (*GSTA1* and *GSTA2*) and three genes coding for sulfotransferases (*SULT1A2*, *SULT1A3* and *SULT1A4*) were also more-frequently found in CNVs from SSC ASD-subjects, with statistical significance after correction for multiple testing, even though the same was not observed in the AGP dataset (table 5). Again, refer to table S2 for frequency data on additional genes that did not reach statistical significance.

Altogether, we identify a new set of ASD candidate genes, with emphasis on 15 genes which were found associated with the disorder in both AGP and SSC datasets. For a resume of the obtained results see figure S2.

Table 3: Numbers and percentages of detoxification and barrier genes targeted by CNVs in individuals from the AGP and SSC datasets and validation results.

	AGP dataset (N=2446)		SSC dataset (N=1124)		Validation
	Number of genes (and %)	Number of individuals with CNVs targeting the genes (and %)	Number of genes (and %)	Number of individuals with CNVs targeting the genes (and %)	Number of common genes between AGP and SSC (and % relative to number of AGP targeted genes)
Any detoxification and/or barrier gene	173/519 (33.3)	555/2446 (22.7)	132/519 (25.4)	238/1124 (21.2)	84/173 (48.6)
Genes targeted by CNVs exclusively in ASD subjects	31/173 (17.9)	62/2446 (2.5)	23/132 (17.4)	43/1124 (3.8)	8/31 (25.8)
Genes more-frequently targeted by CNVs in ASD subjects	23/173 (13.3)	261/2446 (10.7)	12/132 (9.1)	80/1124 (7.1)	7/23 (30.4)

Table 4: Frequencies observed for genes exclusively targeted by CNVs from individuals with ASD from AGP and/or SSC datasets

Genes exclusively targeted by CNVs from both AGP and SSC subjects			Genes exclusively targeted by CNVs only from AGP subjects		Genes exclusively targeted by CNVs only from SSC subjects	
Gene	AGP N (%)	SSC N (%)	Gene	AGP N (%)	Gene	SSC N (%)
<i>STS</i>	12 (0.491)	7 (0.623)	<i>ARSE</i>	3 (0.123)	<i>CCDC101</i>	4 (0.356)
<i>CYP2D6</i>	9 (0.368)	5 (0.445)	<i>ARSH</i>	3 (0.123)	<i>CHST14</i>	3 (0.267)
<i>ARSF</i>	5 (0.204)	1 (0.089)	<i>SLC16A1</i>	3 (0.123)	<i>ADSL</i>	2 (0.178)
<i>GUSB</i>	3 (0.123)	1 (0.089)	<i>TRIM64B</i>	3 (0.123)	<i>CYP4F22</i>	2 (0.178)
<i>CLDN3</i>	1 (0.041)	4 (0.356)	<i>ABCG2</i>	2 (0.082)	<i>HSD17B1</i>	2 (0.178)
<i>CYP2R1</i>	1 (0.041)	1 (0.089)	<i>AKR7A3</i>	2 (0.082)	<i>CHST4</i>	1 (0.089)
<i>SLC3A2</i>	1 (0.041)	1 (0.089)	<i>ALDH3A2</i>	2 (0.082)	<i>CYP4A11</i>	1 (0.089)
<i>SULT2B1</i>	1 (0.041)	1 (0.089)	<i>CHST12</i>	2 (0.082)	<i>CYP4F2</i>	1 (0.089)
			<i>XAGE3</i>	2 (0.082)	<i>CYP4F8</i>	1 (0.089)
			<i>ARSG</i>	1 (0.041)	<i>DNMT3B</i>	1 (0.089)
			<i>CES3</i>	1 (0.041)	<i>JAM2</i>	1 (0.089)
			<i>CYP11B2</i>	1 (0.041)	<i>PTGES</i>	1 (0.089)
			<i>CYP7A1</i>	1 (0.041)	<i>PTGS1</i>	1 (0.089)
			<i>FMO3</i>	1 (0.041)	<i>SLC22A5</i>	1 (0.089)

			<i>GPX2</i>	1 (0.041)	<i>SYN1</i>	1 (0.089)
			<i>ILIRL1</i>	1 (0.041)		
			<i>KIFC1</i>	1 (0.041)		
			<i>NOTUM</i>	1 (0.041)		
			<i>PTGES3</i>	1 (0.041)		
			<i>SLC25A20</i>	1 (0.041)		
			<i>SUOX</i>	1 (0.041)		
			<i>UGT1A1</i>	1 (0.041)		
			<i>UGT2A3</i>	1 (0.041)		

Table 5: Frequencies observed for genes more-frequently targeted by CNVs from individuals with ASD from AGP and/or SSC datasets, after Bonferroni correction for multiple testing. Presented are genes that are more-frequently targeted in both AGP and SSC datasets, and also genes that are more-frequently targeted only in AGP or SSC datasets.

	AGP dataset				SSC dataset				Control dataset
	AGP N (%)	Test Statistic	p-value	OR (95% CI)	SSC N (%)	Test Statistic	p-value	OR (95% CI)	DGV N (%)
<i>ABCBI</i>	20 (0.818)	51.83	6.04x10 ⁻¹³	15.9 (5.96-42.41)	2 (0.178)	0.91	1.61x10 ⁻⁰¹	3.44 (0.67-17.74)	5 (0.052)
<i>GSTM1</i>	20 (0.818)	51.83	6.04x10 ⁻¹³	15.9 (5.96-42.41)	2 (0.178)	0.91	1.61x10 ⁻⁰¹	3.44 (0.67-17.74)	5 (0.052)
<i>CYP21A2</i>	67 (2.739)	44.14	3.05x10 ⁻¹¹	2.83 (2.07-3.88)	23 (2.046)	9.52	2.03x10 ⁻⁰³	2.1 (1.33-3.33)	95 (0.985)
<i>CSH1</i>	19 (0.777)	41.89	9.64x10 ⁻¹¹	10.78 (4.53-25.68)	9 (0.801)	31.25	2.27x10 ⁻⁰⁸	11.12 (4.13-29.91)	7 (0.073)
<i>MAGEA8</i>	16 (0.654)	53.12	1.04x10 ⁻¹⁰	63.53 (8.42-479.27)	15 (1.335)	110.27	2.51x10 ⁻¹⁴	130.5 (17.22-988.88)	1 (0.010)
<i>CYP4X1</i>	17 (0.695)	52.35	1.70x10 ⁻¹⁰	33.76 (7.79-146.22)	7 (0.623)	36.80	3.94x10 ⁻⁰⁶	30.23 (6.27-145.69)	2 (0.021)
<i>GSTT2</i>	16 (0.654)	37.44	9.42x10 ⁻¹⁰	12.7 (4.65-34.7)	1 (0.089)	0.48	4.84x10 ⁻⁰¹	1.72 (0.2-14.71)	5 (0.052)
<i>CHST5</i>	33 (1.349)	35.92	2.06x10 ⁻⁰⁹	4.11 (2.52-6.7)	23 (2.046)	54.94	1.24x10 ⁻¹³	6.28 (3.66-10.77)	32 (0.332)
<i>UGT1A8</i>	14 (0.572)	40.87	1.47x10 ⁻⁰⁸	27.77 (6.31-122.26)	3 (0.267)	8.38	9.64x10 ⁻⁰³	12.91 (2.15-77.34)	2 (0.021)
<i>UGT2B10</i>	14 (0.572)	40.87	1.47x10 ⁻⁰⁸	27.77 (6.31-122.26)	0 (0)	n/a	n/a	n/a	2 (0.021)
<i>UGT1A10</i>	13 (0.531)	37.08	6.40x10 ⁻⁰⁸	25.77 (5.81-114.29)	3 (0.267)	8.38	9.64x10 ⁻⁰³	12.91 (2.15-77.34)	2 (0.021)
<i>CSH2</i>	14 (0.572)	27.75	1.38x10 ⁻⁰⁷	9.25 (3.55-24.1)	8 (0.712)	27.91	1.27x10 ⁻⁰⁷	11.52 (3.99-33.26)	6 (0.062)
<i>UGT1A6</i>	11 (0.450)	33.70	2.22x10 ⁻⁰⁷	43.58 (5.62-337.76)	3 (0.267)	11.61	4.18x10 ⁻⁰³	25.82 (2.68-248.44)	1 (0.010)
<i>UGT1A7</i>	11 (0.450)	33.70	2.22x10 ⁻⁰⁷	43.58 (5.62-337.76)	3 (0.267)	11.61	4.18x10 ⁻⁰³	25.82 (2.68-248.44)	1 (0.010)
<i>UGT1A9</i>	11 (0.450)	33.70	2.22x10 ⁻⁰⁷	43.58 (5.62-337.76)	3 (0.267)	11.61	4.18x10 ⁻⁰³	25.82 (2.68-248.44)	1 (0.010)
<i>GH2</i>	14 (0.572)	25.32	4.87x10 ⁻⁰⁷	7.93 (3.2-19.67)	8 (0.712)	25.16	5.27x10 ⁻⁰⁷	9.87 (3.57-27.28)	7 (0.073)

<i>UGT1A3</i>	10 (0.409)	29.85	1.06x10 ⁻⁰⁶	39.61 (5.07-309.55)	3 (0.267)	11.61	4.18x10 ⁻⁰³	25.82 (2.68-248.44)	1 (0.010)
<i>UGT1A4</i>	10 (0.409)	29.85	1.06x10 ⁻⁰⁶	39.61 (5.07-309.55)	3 (0.267)	11.61	4.18x10 ⁻⁰³	25.82 (2.68-248.44)	1 (0.010)
<i>UGT1A5</i>	10 (0.409)	29.85	1.06x10 ⁻⁰⁶	39.61 (5.07-309.55)	3 (0.267)	11.61	4.18x10 ⁻⁰³	25.82 (2.68-248.44)	1 (0.010)
<i>GSTT2B</i>	12 (0.491)	26.49	3.67x10 ⁻⁰⁶	11.89 (3.83-36.89)	3 (0.267)	1.72	4.18x10 ⁻⁰³	2.15 (0.24-19.23)	4 (0.041)
<i>ABCC1</i>	11 (0.450)	23.07	1.37x10 ⁻⁰⁵	10.89 (3.47-34.24)	8 (0.712)	34.85	4.62x10 ⁻⁰⁶	17.28 (5.2-57.49)	4 (0.041)
<i>CLDN5</i>	7 (0.286)	18.48	9.05x10 ⁻⁰⁵	27.69 (3.41-225.17)	1 (0.089)	0.45	1.98x10 ⁻⁰¹	8.59 (0.54-137.45)	1 (0.010)
<i>CESI</i>	17 (0.695)	14.68	1.27x10 ⁻⁰⁴	3.55 (1.84-6.83)	1 (0.089)	0.18	7.15x10 ⁻⁰¹	0.45 (0.06-3.37)	19 (0.197)
<i>SULT1A3</i>	6 (0.245)	7.50	6.72x10 ⁻⁰³	5.93 (1.67-21.03)	10 (0.890)	49.47	9.95x10 ⁻⁰⁸	21.64 (6.78-69.13)	4 (0.041)
<i>SULT1A4</i>	6 (0.245)	7.50	6.72x10 ⁻⁰³	5.93 (1.67-21.03)	10 (0.890)	49.47	9.95x10 ⁻⁰⁸	21.64 (6.78-69.13)	4 (0.041)
<i>GSTA1</i>	7 (0.286)	12.44	9.23x10 ⁻⁰⁴	9.23 (2.38-35.71)	6 (0.534)	24.75	8.10x10 ⁻⁰⁵	17.26 (4.31-69.09)	3 (0.031)
<i>GSTA2</i>	7 (0.286)	12.44	9.23x10 ⁻⁰⁴	9.23 (2.38-35.71)	6 (0.534)	24.75	8.10x10 ⁻⁰⁵	17.26 (4.31-69.09)	3 (0.031)
<i>SULT1A2</i>	0 (0)	n/a	n/a	n/a	5 (0.445)	21.74	2.15x10 ⁻⁰⁴	21.55 (4.18-111.22)	2 (0.021)

Identification of potentially pathogenic SNVs targeting genes involved in detoxification and regulation of barrier permeability using ASC dataset

Using exome-sequencing data collected through the ASC, we searched for SNVs from 2674 subjects with ASD and 752 controls targeting the 519 studied genes. After quality control we obtained a total of 52180 variants present in cases, targeting said genes (figure 2). These variants were annotated using VEP, which classifies them into four categories according to the impact on the transcript: low, modifier, moderate and high impact. As previously said, moderate and high impact variants include, respectively, missense and l-o-f mutations. Low impact variants include synonymous, start and stop retained and splice region variants, while modifier impact variants comprehend intronic variants and variants located at the 3' and 5' prime UTRs, among others. For variants prioritization only high impact variants and moderate impact variants predicted as detrimental by both SIFT and PolyPhen-2 *in silico* tools were considered.

From the 52180 variants identified in cases, 381 had a high impact, with 71/381 (18.6%) being also present in controls. A total of 201/519 (38.7%) genes were targeted by high impact variants in cases. 17754 moderate impact variants were identified, of which 3339 were predicted as having a detrimental effect by both *in silico* tools. Of these, 683/3339 (20.5%) were present in controls. An amount of 417/519 (80.3%) genes were targeted by the 3339 variants. A total of 420/519 (80.9%) genes were targeted by the total 3720 high and deleterious and damaging moderate impact variants.

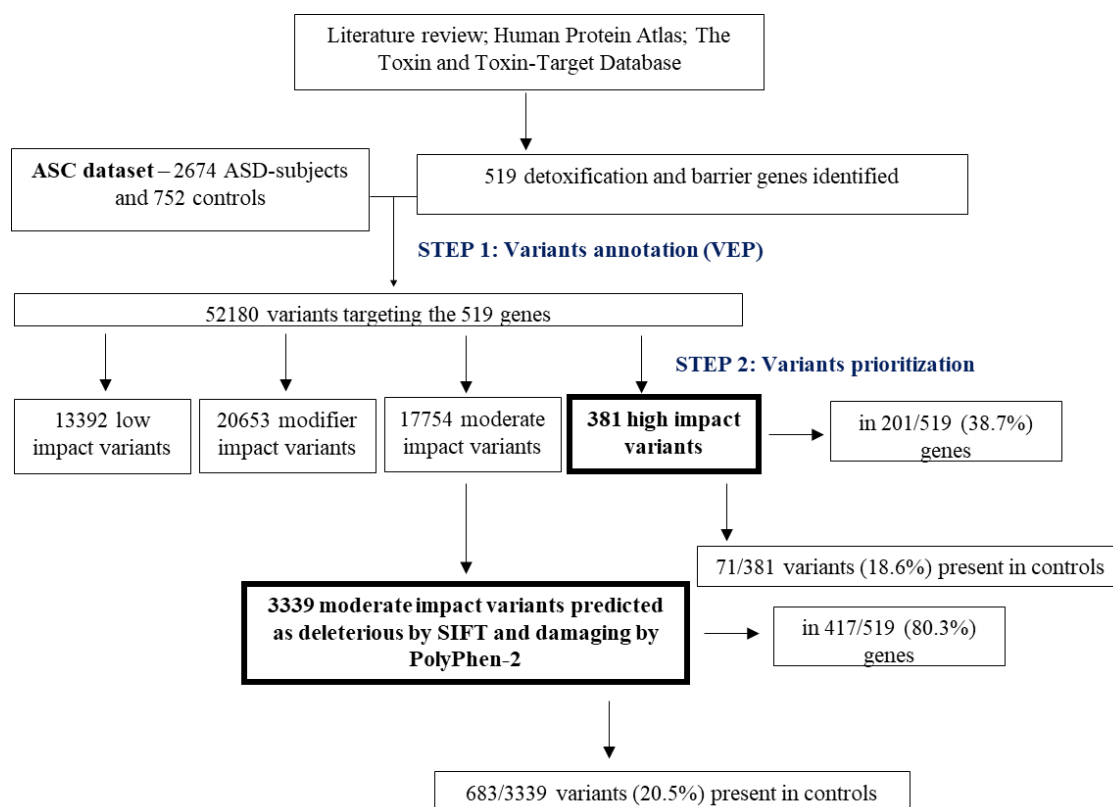


Figure 2: Main results regarding the numbers of high and moderate (predicted as deleterious by SIFT and damaging by PolyPhen-2) impact SNVs targeting detoxification and barrier genes, in cases and controls from ASC dataset.

A major limitation of our study is the small size of ASC control population (n=752), when compared to case population (n=2674). Thus, when a variant is exclusively found in 1 or 2 patients, and not in controls, this might be for two reasons: 1) the variant is, indeed, associated with ASD risk; 2) the control population is too small to find that same variant. To try to overcome this issue, we refined the prioritization step by establishing a rank of the 3720 l-o-f and missense variants according to their frequency in cases and controls. As shown in table 6, six ranks (A-F) were defined. Ranks A and B include variants exclusively found in ≥ 6 or in $3 \leq n \leq 5$ cases, respectively. Given that case population is 2.8x bigger than the control population, variants that appear in 3 or more patients, and not in controls, are hypothetically associated with the pathology. A total of 192 variants are included in the first two ranks (3 of the A rank variants solely appeared in homozygosity, and only in cases) (table 6). Contrary, the 2774 variants that appear only in $1 \leq n \leq 2$ cases and not in controls were classified in rank E, as these may not be associated with the phenotype (table 6). Ranks C and D include 223 SNVs more frequent in cases than controls (table 6). Surprisingly, 531 variants were found to be more common in control population (table 6). As discussed ahead, this can be due to the fact that many of these variants may only have a pathogenic effect in the presence of an external factor.

Table 6: Numbers high and moderate impact variants included in each rank. Again, only missense variants defined as deleterious by SIFT and damaging by PolyPhen-2 were considered. AF – allelic frequency.

Rank	Description of rank	N° of variants	N° of high impact variants	N° of moderate impact variants
A	Variants exclusively present in cases ($n \geq 6$) or only in homozygosity in cases	44	3	41
B	Variants exclusively present in cases ($3 \leq n \leq 5$)	148	15	133
C	Variants more-frequently present in cases (AF cases/AF controls >1.5)	85	5	80
D	Variants more-frequently present in cases (AF cases/AF controls between 1 and 1.5)	138	8	130
E	Variants exclusively present in cases ($1 \leq n \leq 2$)	2774	292	2482
F	Variants more frequently present in controls (AF cases/AF controls <1)	531	58	473
	Total variants	3720	381	3339

Identification of genes with high load of potentially pathogenic SNVs

Upon ranking the 3720 l-o-f and detrimental missense mutations, for each of the 420 genes targeted by such variants, we summed the number of variants per rank.

As listed in table 7, a total of 40 genes were found to be targeted by variants in rank A. Such variants were found in 6 or more cases and never in controls, and thus are the best candidates for an association with ASD, even considering the case vs control size population problem. Notably, only 4 of these 40 genes (*CFTR*, *CYP2D6*, *AKR1B10* and *SLC1A1*) had two variants in rank A. For all other genes, only one variant was observed (table 7). For *GSTO1* and *LGALS16* no other SNVs were found besides a variant in rank A (table 7). 17 genes (*DNAH5*, *CFTR*, *XDH*, *ABCA8*, *ARMC4*, *LOXL4*,

ALDH1L1, *SLC22A5*, *AFDN*, *ALOX15*, *PTGIS*, *CYP4B1*, *ALDH3B1*, *CYP2D6*, *ADH1A*, *AKR1B10* and *SLC1A1*) were also found to carry SNVs from rank 2.

The other 380 genes were not targeted by rank A variants. Still, some, like *DNAH7*, *DNAH11*, *TJP3*, *AOC2*, *CGN*, *SLC12A7* and *TJP1*, were found to have a high load of relevant variants, particularly from ranks B and C, as can be seen in table S3.

Table 7: Numbers of variants by impact and by rank for each gene targeted by, at least, one variant ranked in rank A. Genes are ordered by the amount of high + moderate impact variants they have. In grey are highlighted the 17 genes targeted by both rank A and rank B variants.

Gene	Number of variants by impact			N° of variants by rank					
	High Impact Variants	Moderate Impact Variants	High + Moderate Impact Variants	A	B	C	D	E	F
<i>DNAH5</i>	1	84	85	1	2	1	1	73	7
<i>CFTR</i>	6	52	58	2	6	4	1	39	6
<i>XDH</i>	3	30	33	1	1	1	2	22	6
<i>KIF17</i>	2	27	29	1	0	2	1	18	7
<i>ABCA8</i>	5	22	27	1	2	2	2	16	4
<i>ARMC4</i>	3	24	27	1	1	0	0	20	5
<i>LOXL4</i>	2	25	27	1	3	0	2	18	3
<i>LRP1</i>	0	27	27	1	0	1	0	21	4
<i>ALDH1L1</i>	3	20	23	1	1	0	2	18	1
<i>SLC22A5</i>	3	20	23	1	1	0	0	18	3
<i>APC</i>	0	21	21	1	0	0	1	18	1
<i>AFDN</i>	0	19	19	1	1	0	0	15	2
<i>ALOX15</i>	2	15	17	1	3	0	1	10	2
<i>CYP27A1</i>	2	15	17	1	0	0	1	12	3
<i>PTGIS</i>	1	16	17	1	1	0	2	12	1
<i>ABCG2</i>	3	12	15	1	0	0	1	13	0
<i>CYP4B1</i>	1	14	15	1	2	0	0	11	1
<i>CYP2C19</i>	3	11	14	1	0	1	0	9	3
<i>CBS</i>	1	11	12	1	0	0	1	10	0
<i>IFT74</i>	1	11	12	1	0	0	0	9	2
<i>ALDH3B1</i>	0	11	11	1	1	0	0	8	1
<i>CYP2D6</i>	0	11	11	2	1	0	0	6	2
<i>ADH1A</i>	1	9	10	1	1	0	0	8	0
<i>TXNRD2</i>	2	8	10	1	0	0	0	7	2
<i>UGT2B11</i>	1	9	10	1	0	0	1	5	3
<i>AKR1B10</i>	1	7	8	2	1	0	0	5	0
<i>ALDH18A1</i>	1	7	8	1	0	0	0	7	0
<i>SLC1A1</i>	1	7	8	2	1	0	0	5	0
<i>SULT1C2</i>	1	7	8	1	0	0	1	3	3
<i>GH2</i>	1	6	7	1	0	0	0	4	2
<i>ALDH1A1</i>	0	6	6	1	0	0	0	5	0
<i>CES3</i>	2	4	6	1	0	1	0	2	2
<i>GSTM4</i>	1	3	4	1	0	1	0	2	0
<i>AKRIC2</i>	0	3	3	1	0	0	0	2	0
<i>CHST5</i>	0	3	3	1	0	0	0	1	1
<i>COMT</i>	0	3	3	1	0	0	0	2	0
<i>CES1</i>	1	1	2	1	0	0	0	0	1
<i>DRD1</i>	0	2	2	1	0	0	0	1	0
<i>GSTO1</i>	0	1	1	1	0	0	0	0	0
<i>LGALS16</i>	0	1	1	1	0	0	0	0	0

Interactions between detoxification and barrier genes targeted by potentially pathogenic CNVs and SNVs and environmental factors potentially relevant for ASD

To identify potentially relevant gene-environment interactions for ASD we performed an exploratory analysis using the Comparative Toxicogenomics Database (CTD). We interrogated the CTD for interactions between 54 individual chemicals and 55 detoxification and barrier genes. The 54 analyzed chemicals cluster into the three major classes of environmental factors potentially relevant for ASD, referred in the introduction of this article: 1) environmental toxins; 2) medications; 3) nutritional factors (table S1). The 55 queried genes include the 15 genes more-frequently or exclusively targeted by CNVs in ASD subjects, in both AGP and SSC datasets (tables 4 and 5), and the 40 genes with 1-o-f and missense SNVs from rank A (table 7). As AGP was the main dataset for CNVs discovery, we also added *ABCB1*, *GSTM1* and *CYP21A2*, which were the top genes more-frequently targeted by CNVs in patients from this dataset, but not validated upon SSC analysis (table 5).

We identified a total of 212 gene-environment interaction pairs, between 51/55 (92.7%) genes and 38/54 (70.4%) chemicals (table S4). Four genes (*ARSF*, *CES3*, *LGALS16* and *MAGEA8*) had no reported interactions with the relevant chemicals, with 24 genes interacting only with 1 or 2 environmental factor (figure 3). Notably, for *ABCB1*, *ABCG2*, *GSTM1*, *CYP2D6*, *CYP2C19* and *SLC3A2* we identified interactions with 10 or more chemicals (figure 3). *ABCC1*, another member of the ABC transporters family, interacted with 9 chemicals. Regarding environmental factors, for 16 of them (including 7 PCB congeners, 2 PBDE congeners and 2 phthalates) we did not identify any interplay with the queried genes, which may reflect an understudy of such factors, as discussed below (figure 4). Contrary, 5 of the chemicals (valproic acid, benzo(a)pyrene (b(a)p), bisphenol A, particulate matter and perfluorooctane sulfonic acid (PFOS)) interacted with 10 or more genes (figure 4).

Overall, through this exploratory analysis we identify gene-environment interactions pairs with putative relevance for ASD. Moreover, we present detoxification and barrier genes that interact with more chemicals (and vice-versa), which may pinpoint towards genetic and non-genetic factors that justify further studies in the context of this pathology.

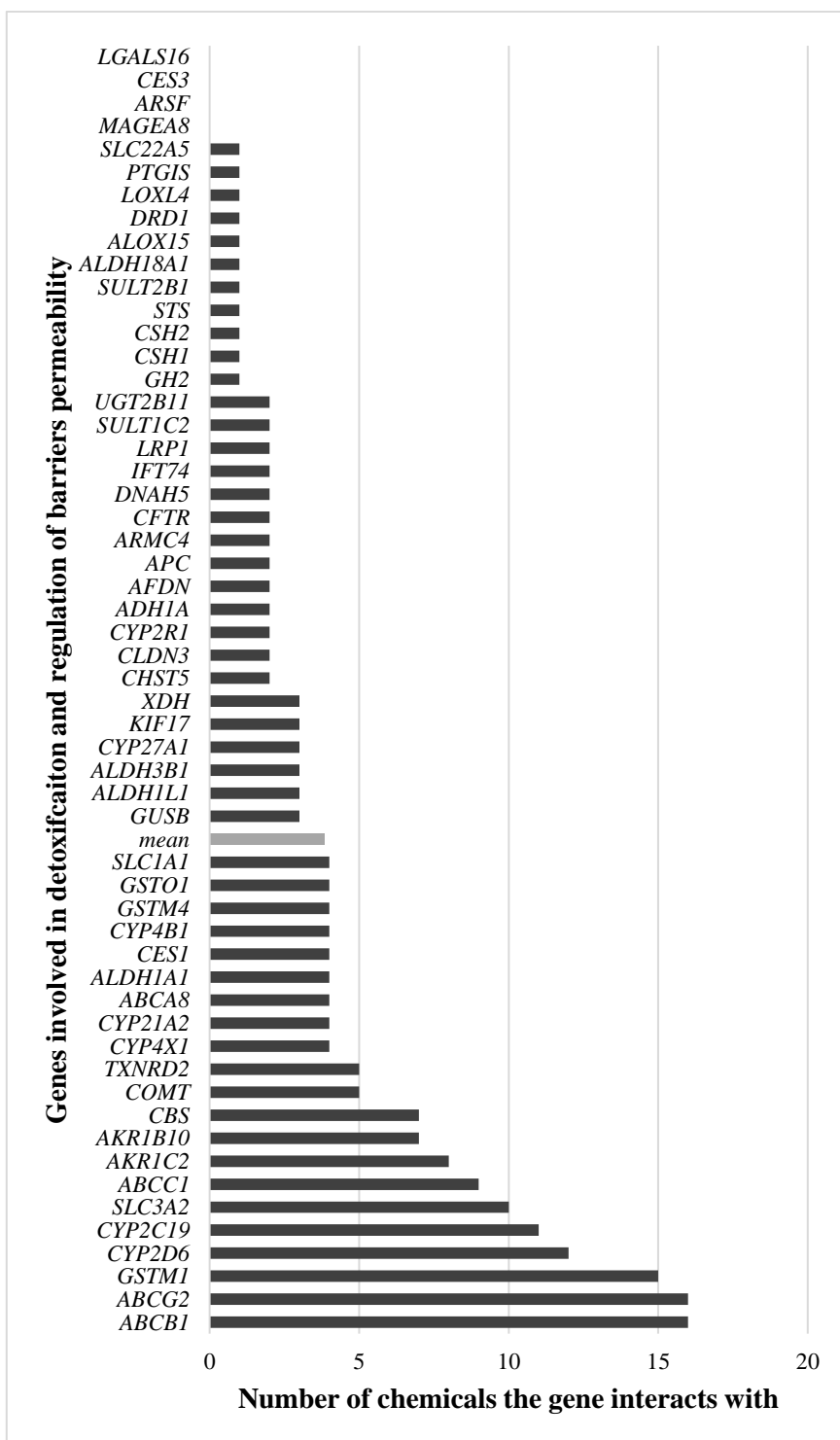


Figure 3: The graphic features the number of individual chemicals potentially relevant for ASD that each of the 55 queried genes interacts with, accordingly to The Comparative Toxicogenomics Database.

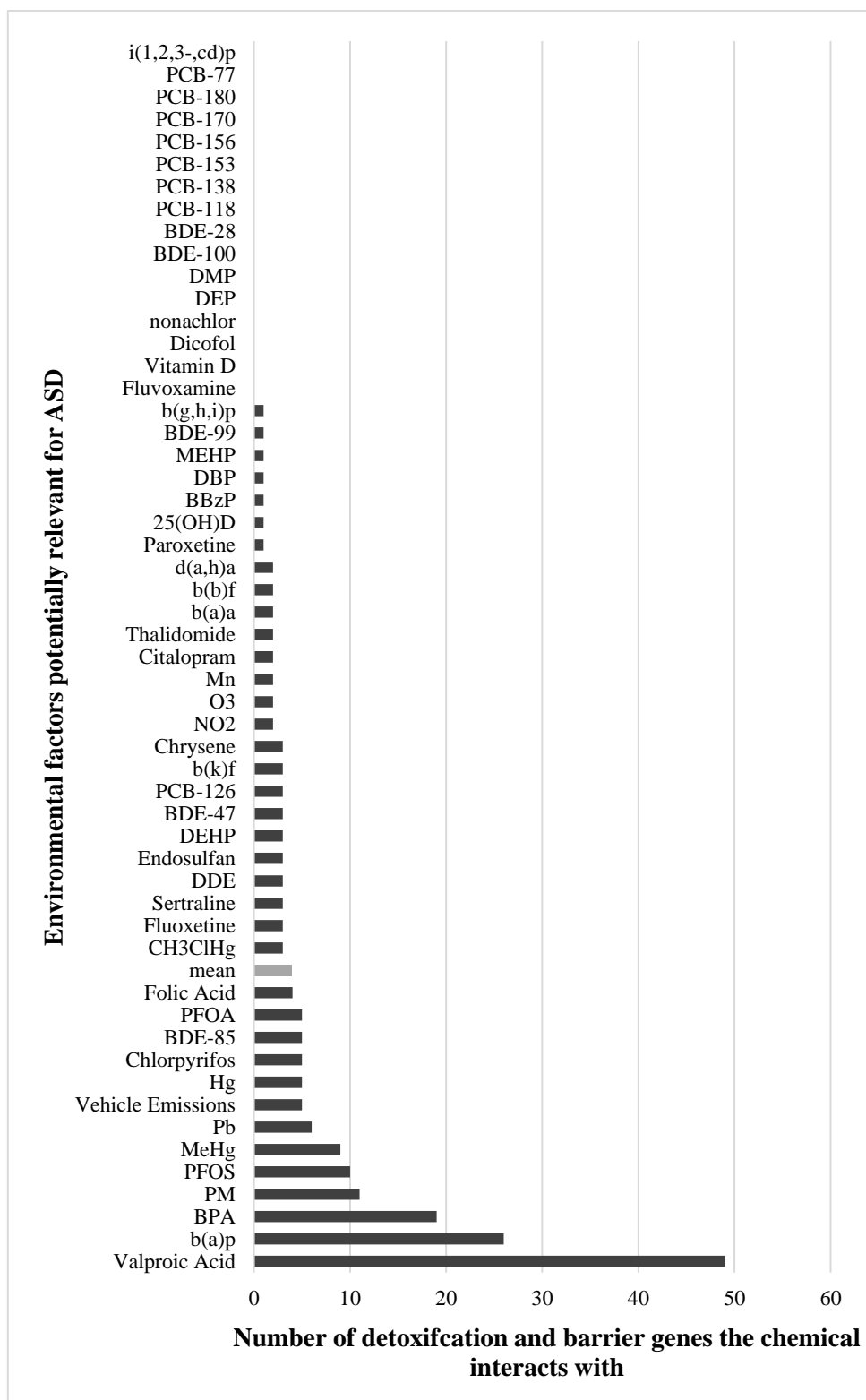


Figure 4: The graphic features number of genes targeted by potentially pathogenic CNVs and/or SNVs that each of the 54 individual chemicals potentially relevant for ASD interacts with, accordingly to The Comparative Toxicogenomics Database.

Discussion

Detoxification and barrier genes with a high burden of potentially pathogenic CNVs and SNVs

Through the analysis of two datasets (AGP and SSC) of subjects diagnosed with ASD, we identified a set of 15 genes involved in detoxification and/or regulation of blood-barrier barrier, placenta or respiratory cilia permeability that are significantly overrepresented in CNVs from these subjects, when comparing with controls. Moreover, using the ASC dataset, also composed of subjects with the pathology, we found that a number of detoxification and barrier genes are targeted by potentially pathogenic loss-of-function and/or missense SNVs.

Genes coding for key enzymes involved in detoxification of xenobiotics were identified by us. Cytochrome P450 enzymes (CYP450s) are a family of monooxygenases implicated in phase I metabolism of most environmental toxins and pharmaceutical drugs. These enzymes oxidize molecules, making them more water-soluble. *CYP2D6* was identified as exclusively-targeted by CNVs in ASD subjects (AGP: n=9, F=0.37%; SSC: n=5, F=0.45%), and was also targeted by 3 potentially pathogenic SNVs exclusive to ASC cases. *CYP2D6* codes for an enzyme that detoxifies multiple toxins, including up to 25% of clinical drugs, with many functional polymorphisms in this gene being known to affect the metabolizer status of their carriers (78). ASD patients who are CYP2D6 poor metabolizers have been shown to exhibit an altered response to therapy using risperidone, an antipsychotic drug (79,80). Another identified CYP450-coding gene was *CYP4X1*, which was significantly more-frequently targeted by CNVs in ASD subjects (AGP: n=17, F=0.695; SSC: n=7, F=0.62). Though this gene is poorly studied, it is suspected to be involved in fatty acids and arachidonic acid metabolism. Interestingly, *CYP4X1* has been suggested to be highly expressed at the late term fetal human brain (81), a pattern that has also been observed in rat brain (82). *CYP21A2* was found to be more-frequently targeted by CNVs in AGP ASD-subjects (n=67, F=2.74). This gene codes for a hydroxylase responsible for steroids biosynthesis. CNVs targeting *CYP2R1* were exclusively found in two patients. *CYP2R1* codes for vitamin D 25-hydroxylase, an enzyme responsible for the conversion of vitamin D acquired from sun exposure or diet to its main circulatory form. This is relevant because a growing number of studies report associations between vitamin D deficiency and ASD.

Genes coding for uridine diphosphate glucuronosyltransferases (UGTs) were found. UGTs are major enzymes of phase II metabolism, and are responsible for glucuronidation reactions, in which substrates are conjugated with a glucuronic acid moiety, increasing their water-solubility. Eight members of the UGT1A gene locus (*UGT1A3-UGT1A10*) were more-frequently targeted by CNVs in AGP ASD-subjects, and would often appear in the same variant. These genes were also found in SSC dataset, but statistical significance did not remain after Bonferroni correction for multiple testing. A CNV targeting *UGT1A1* was also exclusively found in an AGP subject. UGT1A locus, located on chromosome 2, includes nine unique, but highly-similar, transcripts that code for nine enzymes (83). By splicing out in-between sequences, each of the unique exons at the 5' end of the locus is combined with the 3' end exons conserved in all isoforms. Members of the UGT2B locus were also identified: *UGT2B10* was more-frequently found in CNVs from AGP ASD-subjects, and one potentially pathogenic SNV exclusive to cases was present in *UGT2B11* (84). Both proteins are steroid-metabolizing enzymes. Importantly, human and mouse studies show that few UGTs (e.g. *UGT1A4*, *1A6* and *1A7*) are expressed at endothelial cells and astrocytes of the BBB, where they glucuronidate antipsychotic drugs, benzo(a)pyrene and PCBs that cross this barrier (85). Although no associations

between UGT variants and autism have been reported until now, we suggest that attention should be given to them, due to the broad amount of toxins they help degrade.

GSTM1 was more-frequently targeted by CNVs in ASD subjects from the AGP dataset (n=20, F=0.82) and *GSTM4* was found to have two relevant SNVs. These genes code for glutathione S-transferases (GSTs), which are enzymes known to catalyze the conjugation of reduced glutathione to multiple xenobiotics in order to make them more hydrophilic. Notably, homozygous deletion of *GSTM1* has been associated with ASD onset, with the gene being pointed as a candidate for the disorder (86,87). Other GSTs, particularly *GSTT2*, and *GSTA1* and *GSTA2*, were also, respectively, found in higher frequency in AGP and SSC ASD-subjects, when compared to controls.

Another relevant detoxification gene identified was *CHST5*, which was more-frequently found in ASD-subjects (AGP: n=33, F=1.35; SSC n=23, F=2.05) and also carried a missense SNV only present in 6 or more ASC cases. This gene codes for a carbohydrate sulfotransferase, with an exome-sequencing study reporting a rare synonymous mutation associated with the pathology (88). *STS*, that codes for steroid sulfatase, was the top gene exclusively targeted by CNVs in both primary and validation dataset (AGP: n=12, F=0.491; SSC: n=7, F=0.623). Steroid sulfatase is an enzyme that metabolizes steroid hormones and, in the brain, maintains the balance between neurosteroids and their unconjugated forms(89). Point mutations and deletions in *STS* are responsible for X-linked ichthyosis. Notably, males with this dermatological disease are sometimes diagnosed with ASD or Attention Deficit Hyperactivity Disorder (ADHD) or manifest more autistic traits than non-affected subjects (90,91). However, Kent *et al* (90) showed that from 25 males with *STS* deficiency, the 5 that met the criteria for ASD diagnosis all had large deletions that comprised both *STS* and *NLGN4X* genes, with the later coding for a neuroligin already associated with autism. In AGP and SSC datasets, none of the CNVs that included *STS* also targeted *NLGN4X*. Thus, while *NLGN4X* might account for ASD in X-linked ichthyosis, disruptions of *STS* may contribute to the phenotype.

Genes responsible for regulation of barriers permeability, particularly transporters, were also identified as having a high burden of relevant CNVs and SNVs. Multiple members of the ATP-binding cassette (ABCs) transporters family were found. *ABCC1* was more-frequently included in CNVs from AGP and SSC subjects (AGP: n=11, F=0.45; SSC: n=8, F=0.71), when compared to controls, while *ABCB1* was only more-frequently found in AGP subjects (n=20, F=0.82). *ABCA8* was found to have a high load of potentially pathogenic SNVs, with *ABCG2* being identified in both CNVs and SNVs analyses. ABC transporters regulate the flux of xenobiotics across cell membranes. *ABCC1*, *ABCB1* and *ABCG2* are all expressed in both BBB (92) and placenta (93,94), which suggests they may exert a protective role for the fetus, and their substrates include glutathione conjugates and hydrophobic compounds (92). *CFTR*, also an ABC transporter, was the second gene with the highest load of relevant SNVs. Mutations in *CFTR* are the main cause of cystic fibrosis, an autosomal recessive disorder characterized by mucus build-up and reduced mucociliary clearance of the respiratory tract, an important line of defense against airborne pollutants (95). *CFTR* is also expressed in neurons of the developing human brain (96). Other transporters found to carry relevant mutations were members of Solute Carriers (SLCs) family. *SLC3A2* (AGP: n=1, F=0.041; SSC: n=1, F=0.089) and *SLC16A1* (AGP: n=3, F=0.123) were exclusively targeted by CNVs from ASD subjects. *SLC22A5* carried two potentially pathogenic SNVs exclusive to 3 or more cases. At the BBB and placenta, SLCs are responsible for ionic transport and uptake of nutrients and exogenous chemicals. Substrates for *SLC3A2* include methylmercury (97). *SLC22A5* and *SLC16A1* code for proteins involved in the transport of mitochondrial biomarkers, such as carnitine, pyruvate and lactate.

Interestingly, a crescent number of studies report low blood levels of carnitine and elevated blood levels of pyruvate and lactate in some ASD patients (98,99) .

Genes coding for claudins, which are transmembrane proteins essential for the formation and integrity of BBB tight junctions (100), were also found. *CLDN3* was exclusively found in CNVs from subjects with ASD (AGP: $n=1$, $F=0.04$; SSC: $n=4$, $F=0.36$). Meanwhile, *CLDN5* was more-frequently targeted by CNVs from AGP ($n=7$, $F=0.286$), but not SSC, subjects. *CLDN3* and *CLDN5* code for the predominant claudins expressed at the BBB (100). Elevated levels of claudin-5 have been described in postmortem cerebral cortex and cerebellum tissues of subjects with ASD (101). Additionally, *TJP3* was found to have a high burden of potentially pathogenic SNVs (but none from rank A). This gene codes for a scaffolding protein present at tight junctions.

CNVs targeting *CSH1* (AGP: $n=19$, $F=0.78$; SSC: $n=9$, $F=0.80$), *CSH2* (AGP: $n=14$, $F=0.57$; SSC: $n=8$, $F=0.71$) and *GH2* (AGP: $n=14$, $F=0.57$; SSC: $n=8$, $F=0.71$) were more-frequently found in ASD subjects from both AGP and SSC datasets. For *GH2*, a SNV present in more than 6 cases, and not in controls, was also found. Together with *GHI* and *CSHL1*, these three genes are part of the human growth hormone (GH)/chorionic somatomammotropin (CSH) gene cluster located on 17q22-24 band (102). Thus, in our analyses, CNVs would often target these genes together. *GH2* codes for placental growth hormone and *CSH1* and *CSH2* code for human placental lactogen. During pregnancy, *GHI* expression is abrogated, and syncytiotrophoblast cells synthesize and release these placental hormones, that together act to increase the availability of nutrients to the fetus (103). Notably, *GHI*, which is expressed postnatally, was not found by us. To our knowledge, variants in 17q22-24 locus have not been described in ASD, but given the role of these genes in regulation of fetal growth, such variants could potentially lead to neurodevelopmental issues.

Finally, *DNAH5*, *DNAH7* and *DNAH11* were all found to have a high burden of potentially pathogenic SNVs. These genes code for members of axonemal dynein heavy chain family, which are microtubule-based ATPases that regulate motility of respiratory cilia (104). The correct movement of the cilia is crucial to avoid the contact of exogenous substances with the airways. Of note, these genes all have very large size (>300kb).

Overall, we found that subjects with ASD carry potentially pathogenic variants in genes that code for proteins with a wide variety of functions. If the function of these proteins is compromised during early development this might translate into neurodevelopmental problems. Such proteins range from enzymes that increase water-solubility of xenobiotics (CYP450s, UGTs and GSTs), to transporters (ABCs and SLCs), proteins that secure the correct function of barriers (claudins and dyneins) and placental hormones. This emphasizes the complexity of the biological mechanisms that can lead to the pathology.

As previously referred, we hypothesize that subjects carrying variants in these genes will develop ASD only when exposed to an external trigger (or the environmental factors only function as triggers when the genetic susceptibility is present). This is relevant seeing that we defined as important genes more-frequently targeted by CNVs is ASD subjects, after statistical correction for multiple testing, even though they were targeted by variants also in controls. We argue that, in the affected individuals, such CNVs contributed to ASD onset due to an exposure to an external trigger, while the unaffected subjects, carrying CNVs in the same genes, were not exposed to the trigger. Meanwhile, this may partly explain the large amount of l-o-f and missense variants more-frequently found in control subjects from the ASC dataset. Unaffected subjects that carry such variants may not have been exposed to external cues that trigger the pathology. In both CNVs and SNVs cases, we can

also suggest that the variants described here may have a low to moderate impact and, together with other genetic and non-genetic factors, act in concert to reach the threshold necessary for ASD onset. Some subjects carrying these variants suffered other insults that allowed the threshold to be reached, while other subjects with the same variants did not reach such threshold and experienced a typical development. Protective effects, like the proposed female protective effect (105), can augment the onset threshold in some individuals. In some cases, the dysfunction of a given protein may be compensated by the action of another functional homologue protein.

Identification of gene-environment interactions potentially relevant for ASD

Using the CTD we identified 212 gene-environment interaction pairs, between 51/55 detoxification and barrier genes with a high load of relevant variants and 38/54 chemicals potentially relevant for ASD.

ABCB1 and *ABCG2* were the genes identified to interact with more chemicals relevant for ASD, each with 16/54 (29.6%) interactions. *ABCC1* interacted with 9/54 (16.7%) chemicals. As previously referred, these genes code for transporters responsible for the flux of xenobiotics across cell membranes and their dysregulation might lead to cellular imbalances. *CYP2D6* and *CYP2C19* were also identified as top genes interacting, respectively, with 12/54 (22.2%) and 11/54 (20.4%) chemicals. As *CYP2D6* was identified carrying both relevant CNVs and SNVs exclusive to ASD patients and has already been linked to influence risperidone therapy in ASD (79,80), we suggest this cytochrome as a prime candidate for future gene-environment studies in the disorder. *GSTM1* interacted with 15/54 (27.8%) genes, which was expected since it codes for a main glutathione transferase known to act on multiple substrates. Given the reported associations between null *GSTM1* genotype and the pathology, this gene should also be a first candidate for gene-environment interactions (86,87). *SLC3A2* was also one of the top identified genes, with 10/54 (18.5%) interactions. *AKR1C2* and *AKR1B10* were found to interact with 8/54 (14.8%) and 7/54 (13.0%) chemicals. These genes code for members of the NADPH-dependent aldo-keto reductase family. ABC, CYP450, GST and SLC proteins are all products of important pharmacogenes (106), and thus it is possible that studies on their targets are inflated in relation to lesser clinically relevant proteins. This is supported by the fact that no interactions with chemicals were found for *ARSF*, *CES3*, *LGALS16* and *MAGEA8*, which have poorly defined functions. For example, according to *The Human Protein Atlas*, in healthy tissues MAGEA8 protein is only expressed at the testis and placenta and possibly plays a role in embryonic development (107).

Valproic acid was identified as the top chemical, interacting with 49/55 (89.1%) of the genes. An otherwise safe medication, when ingested during pregnancy valproic acid is known to cause serious birth defects, hence its teratogenic capacity (108). This has likely prompted research on its targets. Even when adjusting for potential confounders (e.g., maternal seizure attacks), fetal exposure to valproic acid remains strongly associated with ASD risk (34,46), a fact that is further reinforced by rodent studies (109). B(a)p, the most well studied PAH (110), interacted with 26/55 (47.2%) genes. B(a)p results from the combustion of organic matter and is found in tobacco smoke, diesel exhaust and grilled food, with prenatal exposure to this toxin leading to neurobehavioral problems in animal models and humans (111). Other PAHs were found to interact with 1-3 chemicals, with i(1,2,3,-cd)p having no reported associations, which may reflect an understudy of these congeners (110). Bisphenol A, a major EDC used as a starting material for plastics manufacture, being present in consumer goods such as water bottles and food and beverage cans, was found to interact with 19/55 (34.5%) of the

queried genes. BPA concentrations are measured in urine and, as such, this toxin has been mostly reported as a biomarker for 3 to 13 years old children with the pathology (38,112). However, by measuring maternal urinary concentrations of this toxin, a study found that children, particularly females, with higher gestational exposure manifested more neurobehavioral problems (37). Particulate matter was identified to interact with 11/55 (20.0%) genes. PM, microscopic particles suspended in atmospheric air, are generally divided into two categories according to their size: 1) coarse particulate matter (PM₁₀) with a diameter between 10µm and 2.5µm; 2) fine particulate matter (PM_{2.5}) with a diameter of 2.5µm or less. As they are smaller, PM_{2.5} are particularly noxious because they easily penetrate the respiratory tract, causing multiple adverse conditions that go beyond respiratory diseases (113). Consistent with this, ASD risk associates stronger with PM_{2.5} exposure than with PM₁₀ exposure (34). Among the top chemicals was also PFOS, interacting with 10/55 (18.2%) genes. PFOS is a PFC used as coating and as a water and stain repellent, with prenatal exposure to this substance linked to multiple neurodevelopmental outcomes (114). Toxic heavy metals, such as lead and mercury (and its derivative methylmercury), and folic acid were also found as top chemicals by the CTD analysis.

While it is expected that some environmental factors have broader targets than others, the current knowledge on neurotoxic effects of chemicals is underestimated (115). Of the hundred thousand chemicals existent, few hundreds have documented neurotoxic effects for humans, with this number decreasing when considering neurodevelopmental effects. This happens because, due to the limitations inherent to address so many compounds, most have not been investigated for potential neurotoxicity (115). Effects of chemicals for which there is more concern regarding human exposure and growing awareness (e.g. BPA, b(a)p, pollutants and valproic acid) may be more investigated. Thus, in the future, we anticipate to identify more than the 212 interaction pairs between the queried genes and chemicals identified by the CTD.

Studies that investigate the interplay between genetic and environmental factors in ASD are sparse. Potential interactions associated with the pathology have been reported for prenatal exposure to PM₁₀ and NO₂ and rs1858830 polymorphism on *MET* gene (116), maternal folic acid intake during gestation and rs1801133 polymorphism on *MTHFR* (52) and O₃ exposure and overall burden of copy number duplications (117). We hypothesize that environmental factors can function as a trigger for ASD onset, only in subjects with a genetic susceptibility. Yet, possible biological mechanisms behind this are only now being proposed. Among these are epigenetic alterations, endocrine dysregulations, hypoxic and oxidative stress events, inflammation and mitochondrial dysfunctions (34,98). Epigenetic changes, such as DNA methylation and altered patterns of microRNA expression, have been associated with ASD onset (118). Interestingly, environmental factors relevant for the pathology (e.g. BPA, PCBs and PM) are known to induce epigenetic alterations (119–122). EDCs are able to mimic hormones, leading to endocrine imbalances, and this may be associated with the high male-to-female ratio observed in ASD diagnosis. *STS*, one of the identified genes, codes for an enzyme involved in sex steroids metabolism. Most probably, the multitude of genetic and non-genetic factors involved in the disorder can combine in multiple ways, affecting different biological pathways and originating similar, but distinct, clinical ASD phenotypes.

Limitations and strengths of this study

An important limitation to note in this study is the large size of the datasets used to identify CNVs, but the low occurrence of events (individuals carrying CNVs), explaining the wide confidence

intervals obtained. Nonetheless, our confidence intervals are always concordant with the obtained *p-values*. Larger populations, allowing for the occurrence of more events, may overcome this limitation (still, since many ASD relevant mutations are ultra-rare variants, this could prove to be unattainable). Another limitation of this study is the fact that the ASC dataset, which was used to identify SNVs, has 2.8x more cases than controls. To counteract this fact we ranked the identified potentially pathogenic missense and l-o-f variants, and attributed few weight to SNVs exclusively found in 1 or 2 cases (n=2774 variants). While some of these variants may be ultra-rare and associated with ASD risk, others may not have been found in control-subjects because of the reduced size of the control population. Again, this could be resolved with a larger dataset. A last limitation of our strategy is the fact that we identified multiple interactions between environmental factors reported to be associated with ASD and detoxification and barrier genes, but the type of interactions were not analyzed. Among many other effects, an external factor can influence a protein expression or localization, its folding, stability and structure and its interaction with other molecules. Future studies should look into these matter, as different types of effects in the same protein can lead to different outcomes.

Strengths of our strategy include the analysis of both CNVs and SNVs targeting the studied genes in ASD subjects. Since both types of genetic variants are known risk factors for the disorder, an integrative approach allows for the detection of genes with high load of both types of variants. This was verified in this study for *CYP2D6* and *GH2*. In our query using the CTD only interactions observed in humans were considered. While this greatly reduced the amount of discovered interactions, as animal models are an excellent tool to study the effects of exogenous substances, the gene-environment interaction pairs found offer more consistence for a possible association with ASD. Finally, by ranking missense and l-o-f SNVs based on their cases and controls frequencies, and giving more importance to variants only present in 6 or more controls, despite reducing the number of considered variants, allowed the maintaining of the ones more strongly associated with the phenotype in the analyzed dataset.

Conclusions

In this study we present a two-step strategy that allowed us to discover gene-environment interactions with putative relevance for ASD. First, by using large datasets of genetic information collected from subjects with the pathology, we were able to identify a novel set of ASD-candidate genes targeted by potentially pathogenic CNVs and SNVs. Secondly, using the CTD, we identified interactions between such genes and environmental risk factors for the disorder. It is likely that the identified gene-environment interactions are not, per se, a single causative force, but together with other genetic and non-genetic factors allow the onset of ASD. Each gene-environment interaction pair can be seen as a unit with low to moderate effect that, alone, does not cause the disorder.

Despite the discussed limitations, we expect a positive outcome from this study, as it may contribute in directing future investigations that aim to tackle specific gene-environment interplays in ASD. Exposure to environmental factors can, at some extent, be mitigated, which is an important issue for personalized medicine approaches. ASD risk in genetically susceptible individuals could be decreased if early exposure to environmental triggers is eliminated. Thus, the identification of gene-environment interactions in ASD may contribute to the implementation of health management policies in the disorder.

Declarations

Acknowledgements

We acknowledge the Simons Foundation for Autism Research (SFARI) and the Database of Genomic Variants (DGV) for the availability of data regarding SSC and DGV datasets. We are grateful to the families recruited through Autism Genome Project, Simons Simplex Collection and Autism Sequencing Consortium projects.

Financial Support

João Xavier Santos is a fellow of the BioSys PhD Program and an awardee of a scholarship funded by *Fundação para a Ciência e Tecnologia*, Portugal (Ref: PD/BD/114386/2016). AGP data was collected from patients genotyped in the context of Autism Genome Project (AGP), funded by NIMH, HRB, MRC, Autism Speaks, Hilibrand Foundation, Genome Canada, OGI and CIHR. ASC data was collected from patients genotyped in the context of Autism Sequencing Consortium supported by NIH grants U01MH100233, U01MH100209, U01MH100229 and U01MH100239.

Conflicts of Interest

The authors report no conflicts of interest.

References

1. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (5th ed.). 2013.
2. Lord C, Elsabbagh M, Baird G, Veenstra-Vanderweele J. Seminar Autism spectrum disorder. *Lancet*. 2018; 392(10146):508–20.
3. Elsabbagh M, Divan G, Koh YJ, Kim YS, Kauchali S, Marcín C, et al. Global Prevalence of Autism and Other Pervasive Developmental Disorders. *Autism Res*. 2012;5(3):160–79.
4. Lyall K, Croen L, Daniels J, Fallin MD, Ladd-Acosta C, Lee B, et al. The Changing Epidemiology of Autism Spectrum Disorders. *Annu Rev Public Health*. 2017;38:81-102
5. Loomes R, Hull L, Mandy WPL. What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis. *J Am Acad Child Adolesc Psychiatry*. 2017;56(6):466–74.
6. Werling DM, Geschwind DH. Sex differences in autism spectrum disorders. *Curr Opin Neurol*. 2013;26(2):146–53.
7. Levy D, Ronemus M, Yamrom B, Lee Y ha, Leotta A, Kendall J, et al. Rare De Novo and Transmitted Copy-Number Variation in Autistic Spectrum Disorders. *Neuron*. 2011;70(5):886–97.
8. Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L, et al. Convergence of Genes and Cellular Pathways Dysregulated in Autism Spectrum Disorders. *Am J Hum Genet*. 2014;94(5):677–94.
9. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, et al. Functional Impact of Global Rare Copy Number Variation in Autism Spectrum Disorder. *Nature*. 2010;466(7304):368–72.

10. Sanders S, Ercan-sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, et al. Multiple recurrent de novo copy number variations (CNVs), including duplications of the 7q11.23 Williams-Beuren syndrome region, are strongly associated with autism. *Neuron*. 2011;70(5):863–85.
11. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese- C, Walsh T, et al. Strong Association of De Novo Copy Number Mutations with Autism. *Science*. 2007;316(5823):445–9.
12. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, et al. De novo mutations revealed by whole exome sequencing are strongly associated with autism. *Nature*. 2013;485(7397):237–41.
13. Iossifov I, O’Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature*. 2014;515(7526):216-21
14. Neale BM, Kou Y, Liu L, Ma’ayan A, Samocha KE, Lin C, et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*. 2013;485(7397):242–5.
15. Gaugler T, Klei L, Sanders SJ, Bodea CA, Goldberg AP, Lee AB, et al. Most genetic risk for autism resides with common variation. *Nat Genet*. 2014;46(8):881–5.
16. Devlin B, Scherer SW. Genetic architecture in autism spectrum disorder. *Curr Opin Genet Dev*. 2012;22(3):229–37.
17. Sandin S, Lichtenstein P, Kuja-Halkola R, Larsson H, Hultman CM, Reichenberg A. The familial risk of autism. *JAMA*. 2014;311(17):1770–7.
18. Sandin S, Lichtenstein P, Kuja-Halkola R, Hultman C, Larsson H, Reichenberg A. The Heritability of Autism Spectrum Disorder Analysis method B. *JAMA*. 2017;318(12):1182–4.
19. Pettersson E, Lichtenstein P, Larsson H, Song J, Attention Deficit/Hyperactivity Disorder Working Group of the Ipsych-Broad-PGC Consortium, et al. Genetic influences on eight psychiatric disorders based on family data of 4 408 646 full and half-siblings, and genetic data of 333 748 cases and controls. 2018; 1-8
20. Eapen V. Genetic basis of autism: Is there a way forward? *Curr Opin Psychiatry*. 2011;24(3):226–36.
21. Pinson A, Bourguignon JP, Parent AS. Exposure to endocrine disrupting chemicals and neurodevelopmental alterations. *Andrology*. 2016;4(4):706–22.
22. Braun JM, Kalkbrenner AE, Just AC, Yolton K, Calafat AM, Sjödin A, et al. Gestational exposure to endocrine-disrupting chemicals and reciprocal social, repetitive, and stereotypic behaviors in 4- and 5-year-old children: the HOME study. *Environ Health Perspect*. 2014;122(5):513–20.
23. Eskenazi B, Marks AR, Bradman A, Harley K, Barr DB, Johnson C, et al. Organophosphate pesticide exposure and neurodevelopment in young Mexican-American children. *Environ Health Perspect*. 2007;115(5):792–8.
24. Arora M, Reichenberg A, Willfors C, Austin C, Gennings C, Berggren S, et al. Fetal and postnatal metal dysregulation in autism. *Nat Commun*. 2017;8:15493.
25. Hertz-Picciotto I, Schmidt RJ, Walker CK, Bennett DH, Oliver M, Shedd-Wise KM, et al. A Prospective Study of Environmental Exposures and Early Biomarkers in Autism Spectrum Disorder: Design, Protocols, and Preliminary Data from the MARBLES Study. *Environ*

- Health Perspect. 2018;126(11):117004.
26. Kim SM, Han DH, Lyoo HS, Min KJ, Kim KH, Renshaw P. Exposure to environmental toxins in mothers of children with autism spectrum disorder. *Psychiatry Investig*. 2010;7(2):122–7.
 27. Ming X, Brimacombe M, Malek JH, Jani N, Wagner GC. Autism spectrum disorders and identified toxic land fills: co-occurrence across States. *Environ Health Insights*. 2008;2:55–9.
 28. Shelton JF, Geraghty EM, Tancredi DJ, Delwiche LD, Schmidt RJ, Ritz B, et al. Neurodevelopmental Disorders and Prenatal Residential Proximity to Agricultural Pesticides: The CHARGE Study. *Environ Health Perspect*. 2014;122(10):1103–9.
 29. Becerra T, Wilhelm M, Olsen J, Cockburn M, Ritz B. Ambient air pollution and autism in Los Angeles County, California. *Environ Health Perspect*. 2012;121(3):380–6.
 30. Volk HE, Lurmann F, Penfold B, Hertz-Picciotto I, McConnell R. Traffic-related air pollution, particulate matter, and autism. *JAMA psychiatry*. 2013;70(1):71–7.
 31. Croen LA, Grether JK, Yoshida CK, Odouli R, Hendrick V. Antidepressant use during pregnancy and childhood autism spectrum disorders. *Arch Gen Psychiatry*. 2011;68(11):1104–12.
 32. Taylor LE, Swerdfeger AL, Eslick GD. Vaccines are not associated with autism: An evidence-based meta-analysis of case-control and cohort studies. *Vaccine*. 2014;32(29):3623–9.
 33. Rossignol D a, Genuis SJ, Frye RE. Environmental toxicants and autism spectrum disorders: a systematic review. *Transl Psychiatry*. 2014;4(2):e360.
 34. Modabbernia A, Velthorst E, Reichenberg A. Environmental risk factors for autism: an evidence-based review of systematic reviews and meta-analyses. *Mol Autism*. 2017;8(1):1–16.
 35. Jung CR, Lin YT, Hwang BF. Air Pollution and Newly Diagnostic Autism Spectrum Disorders: A Population-Based Cohort Study in Taiwan. *PLoS One*. 2013;8(9):e75510.
 36. Volk HE, Hertz-Picciotto I, Delwiche L, Lurmann F, McConnell R. Residential proximity to freeways and autism in CHARGE study. *Environ Health Perspect*. 2011;119(6):873–7.
 37. Braun JM, Kalkbrenner AE, Calafat AM, Yolton K, Ye Xiaoyun et al. Impact of Early-Life Bisphenol A Exposure on Behavior and Executive Function in Children. 2011;128(5):873–82.
 38. Stein TP, Schluter MD, Steer RA, Guo L, Ming X. Bisphenol A Exposure in Children With Autism Spectrum Disorders. *Autism Res*. 2015;8(3):272–83.
 39. Roberts AL, Lyall K, Hart JE, Laden F, Just AC, Bobb JF, et al. Perinatal air pollutant exposures and autism spectrum disorder in the children of Nurses' Health Study II participants. *Environ Health Perspect*. 2013;121(8):978–84.
 40. Philippat C, Barkoski J, Tancredi DJ, Elms B, Barr DB, Ozonoff S, et al. Prenatal exposure to organophosphate pesticides and risk of autism spectrum disorders and other non-typical development at 3 years in a high-risk cohort. *Int J Hyg Environ Health*. 2018;221(3):548–55.
 41. Testa C, Nuti F, Hayek J, De Felice C, Chelli M, Rovero P, et al. Di-(2-ethylhexyl) phthalate and autism spectrum disorders. *ASN Neuro*. 2012;4(4):223–9.
 42. Miodovnik A, Engel SM, Zhu C, Ye X, Soorya LV, Silva MJ, Calafat AM, Wolf MS.

- Endocrine disruptors and childhood social impairment. *Neurotoxicology*. 2012;32(2):261–7.
43. Cheslack-Postava K, Rantakokko P V., Hinkka-Yli-Salomäki S, Surcel HM, McKeague IW, Kiviranta HA, et al. Maternal serum persistent organic pollutants in the Finnish Prenatal Study of Autism: A pilot study. *Neurotoxicol Teratol*. 2013;38(212):1–5.
 44. Lyall K, Yau VM, Hansen R, Kharrazi M, Yoshida CK, Calafat AM et al. Prenatal maternal serum concentrations of per- and polyfluoroalkyl substances in association with autism spectrum disorder and intellectual disability. *Environ Health Perspect*. 2018;126(1):017001.
 45. Bromley RL, Mawer GE, Briggs M, Cheyne C, Clayton-smith J, García-fiñana M, et al. The prevalence of neurodevelopmental disorders in children prenatally exposed to antiepileptic drugs. *Journal Neurol Neurosurg Psychiatry*. 2014;84(6):637–43.
 46. Christensen J, Gronborg TK, Sorensen MJ, Schendel D, Parner ET, Pedersen LH, et al. Prenatal valproate exposure and risk of Autism Spectrum Disorders. *Jama*. 2014;309(16):1696-703
 47. Rodier PM, Ingram JL, Tisdale B, Nelson S, Romano J. Embryological origin for autism: Developmental anomalies of the cranial nerve motor nuclei. *J Comp Neurol*. 1996;370(2):247–61.
 48. Stromland K, Nordin V, Miller M, Akerstrom B, Gillberg C. Autism in Thalidomide Embryopathy: A population study. 1994;36(4):351-56.
 49. Rai D, Lee BK, Dalman C, Newschaffer C, Lewis G, Magnusson C. Antidepressants during Pregnancy and Autism in Offspring: Population Based Cohort Study. *BMJ*. 2017;358:j2811
 50. Gidaya NB, Lee BK, Burstyn I, Yudell M, Mortensen EL, Newschaffer CJ. In utero exposure to selective serotonin reuptake inhibitors and risk for autism spectrum disorder. *J Autism Dev Disord*. 2014;44(10):2558–67.
 51. Al-Farsi YM, Waly MI, Deth RC, Al-Sharbati MM, Al-Shafae M, Al-Farsi O, et al. Low folate and vitamin B12 nourishment is common in Omani children with newly diagnosed autism. *Nutrition*. 2013;29(3):537–41.
 52. Schmidt RJ, Tancredi DJ, Ozonoff S, Hansen RL, Hartiala J, Allayee H, et al. Maternal periconceptional folic acid intake and risk of autism spectrum disorders and developmental delay in the CHARGE (CHildhood Autism Risks from Genetics and Environment) case-control study. *Am J Clin Nutr*. 2012;96(1):80–9.
 53. Schug TT, Blawas AM, Gray K, Heindel JJ, Lawler CP. Elucidating the links between endocrine disruptors and neurodevelopment. *Endocrinology*. 2015;156(6):1941–51.
 54. de Cock M, Maas YGH, Van De Bor M. Does perinatal exposure to endocrine disruptors induce autism spectrum and attention deficit hyperactivity disorders? Review. *Acta Paediatr*. 2012;101(8):811–8.
 55. Diamanti-Kandarakis E, Bourguignon JP, Giudice LC, Hauser R, Prins GS, Soto AM, et al. Endocrine-disrupting chemicals: An Endocrine Society scientific statement. *Endocr Rev*. 2009;30(4):293–342.
 56. Ballabh P, Braun A, Nedergaard M. The blood-brain barrier: An overview: Structure, regulation, and clinical implications. *Neurobiol Dis*. 2004;16(1):1–13.
 57. Griffiths SK, Campbell JP. Placental structure, function and drug transfer. *Contin Educ*

- Anaesthesia, *Crit Care Pain*. 2015;15(2):84–9.
58. Antunes MB, Cohen NA. Mucociliary clearance -- a critical upper airway host defense mechanism and methods of assessment. *Curr Opin Allergy Clin Immunol*. 2007;7(1):5-10.
 59. Zheng W, Aschner M, Ghersi-Egea J-F. Brain barrier systems: a new frontier in metal neurotoxicological research. *Toxicol Appl Pharmacol*. 2003;192(1):1–11.
 60. Schonfelder G, Wittfoht W, Hopp H, Talsness CE, Paul M, Chahoud I. Parent bisphenol a accumulation in the human maternal-fetal-placental unit. *Environ Health Perspect*. 2002;110(11):703–7.
 61. Mose T, Mortensen GK, Hedegaard M, Knudsen LE. Phthalate monoesters in perfusate from a dual placenta perfusion system, the placenta tissue and umbilical cord blood. *Reprod Toxicol*. 2007;23(1):83–91.
 62. Gundacker C, Hengstschläger M. The role of the placenta in fetal exposure to heavy metals. *Wien Med Wochenschrift*. 2012;162(9–10):201–6.
 63. Acosta-Maldonado B, Sánchez-Ramírez B, Reza-López S, Levario-Carrillo M. Effects of exposure to pesticides during pregnancy on placental maturity and weight of newborns: A cross-sectional pilot study in women from the Chihuahua State, Mexico. *Hum Exp Toxicol*. 2009;28(8):451–9.
 64. Zhang X, Li X, Jing Y, Fang X, Zhang X, Lei B, et al. Transplacental transfer of polycyclic aromatic hydrocarbons in paired samples of maternal serum, umbilical cord serum, and placenta in Shanghai, China. *Environ Pollut*. 2017;222:267–75.
 65. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. A copy number variation morbidity map of developmental delay. *Nat Genet*. 2011;43(9):838–46.
 66. Shaikh T, Gai X, Perin JC, Glessner JT, Xie H, Murphy K, O'Hara R et al. High-resolution mapping and analysis of copy number variations in the human genome: A data resource for clinical and research applications. *Genome Res*. 2009;19(9):1682–90.
 67. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants: A curated collection of structural variation in the human genome. *Nucleic Acids Res*. 2014;42(D1):986–92.
 68. Buxbaum JD, Daly MJ, Devlin B, Lehner T, Roeder K, State MW, Autism Sequencing Consortium. The Autism Sequencing Consortium: Large-Scale, High-Throughput Sequencing in Autism Spectrum Disorders. *Neuron*. 2012;76(6):1052–6.
 69. Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, et al. A subcellular map of the human proteome. *Science*. 2017;356(6340):eaal3321.
 70. Wishart D, Arndt D, Pon A, Sajed T, Guo AC, Djoumbou Y, et al. T3DB: The toxic exposome database. *Nucleic Acids Res*. 2015;43(D1):D928–34.
 71. Lim E, Pon A, Djoumbou Y, Knox C, Shrivastava S, Guo AC, et al. T3DB: A comprehensively annotated database of common toxins and their targets. *Nucleic Acids Res*. 2009;38(D1):781–6.
 72. Lek M, Karczewski KJ, Minikel E V., Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285–91.
 73. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant

- Effect Predictor. *Genome Biol*; 2016;17(1):1–14.
74. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res*. 2012;40(W1):452–7.
 75. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7(4):248–9.
 76. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, McMorran R, Wieggers J, et al. The Comparative Toxicogenomics Database: Update 2019. *Nucleic Acids Res*. 2019;47(D1):D948–54.
 77. Carter CJ, Blizard RA. Autism genes are selectively targeted by environmental pollutants including pesticides, heavy metals, bisphenol A, phthalates and many others in food, cosmetics or household products. *Neurochem Int*; 2016;101:83–109.
 78. Ingelman-Sundberg M, Sim SC, Gomez A, Rodriguez-Antona C. Influence of cytochrome P450 polymorphisms on drug therapies: Pharmacogenetic, pharmacoepigenetic and clinical aspects. *Pharmacol Ther*. 2007;116(3):496–526.
 79. Youngster I, Zachor DA, Gabis LV, Bar-Chaim A, Benveniste-Levkovitz P, Britzi M, et al. CYP2D6 genotyping in paediatric patients with autism treated with risperidone: A preliminary cohort study. *Dev Med Child Neurol*. 2014;56(10):990–4.
 80. Puangpetch A, Vanwong N, Nuntamool N, Hongkaew Y, Chamnanphon M, Sukasem C. CYP2D6 polymorphisms and their influence on risperidone treatment. *Pharmacogenomics Pers Med*. 2016;9:131–47.
 81. Choudhary D, Jansson I, Stoilov I, Sarfarazi M, Schenkman JB. Expression patterns of mouse and human CYP orthologs (families 1-4) during development and in different adult tissues. *Arch Biochem Biophys*. 2005;436(1):50–61.
 82. Bylund J, Zhang C, Harder DR. Identification of a novel cytochrome P450, CYP4X1, with unique localization specific to the brain. *Biochem Biophys Res Commun*. 2002;296(3):677–84.
 83. Gong QH, Cho JW, Huang T, Potter C, Gholami N, Basu NK, et al. Thirteen UDPglucuronosyltransferase genes are encoded at the human UGT1 gene complex locus. *Pharmacogenetics*. 2001;11(4):357–68.
 84. Turgeon D, Carrier JS, Lévesque E, Hum DW, Bélanger A. Relative enzymatic activity, protein stability, and tissue distribution of human steroid-metabolizing UGT2B subfamily members. *Endocrinology*. 2001;142(2):778–87.
 85. Ouzzine M, Gulberti S, Ramalanjaona N, Magdalou J, Fournel-Gigleux S. The UDP-glucuronosyltransferases of the blood-brain barrier: their role in drug metabolism and detoxication. *Front Cell Neurosci*. 2014;8:1–12.
 86. James SJ, Melnyk S, Jernigan S, Cleves MA, Halsted CH, Wong DH, et al. Metabolic endophenotype and related genotypes are associated with oxidative stress in children with autism. *Am J Med Genet B Neuropsychiatr Genet*. 2006;141B(8): 947-56.
 87. Buyske S, Williams TA, Mars AE, Stenroos ES, Ming SX, Wang R, et al. Analysis of case-parent trios at a locus with a deletion allele: association of GSTM1 with autism. *BMC Genet*. 2006;7:8.
 88. O’Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, et al. Exome sequencing

- in sporadic autism spectrum disorders identifies severe de novo mutations. 2012;43(6):585–9.
89. Kříž L, Bičíková M, Hampl R. Roles of steroid sulfatase in brain and other tissues. *Physiol Res*. 2008;57(5):657–68.
 90. Kent L, Emerton J, Bhadravathi V, Weisblatt E, Pasco G, Willatt LR, et al. X-linked ichthyosis (steroid sulfatase deficiency) is associated with increased risk of attention deficit hyperactivity disorder, autism and social communication deficits. *J Med Genet*. 2008;45(8):519–24.
 91. Chatterjee S, Humby T, Davies W. Behavioural and psychiatric phenotypes in men and boys with X-linked ichthyosis: Evidence from a worldwide online survey. *PLoS One*. 2016;11(10):1–16.
 92. Begley DJ. ABC Transporters and the Blood-Brain Barrier. *Curr Pharm Des*. 2004;10(12):1295–312.
 93. Joshi AA, Vaidya SS, St-pierre M V, Mikheev AM, Kelly E, Nyandege AN, et al. Placental ABC transporters: biological impact and pharmaceutical significance. *Pharm Res*. 2016;33(12):2847–78.
 94. Mao Q. BCRP/ABCG2 in the placenta: Expression, function and regulation. *Pharm Res*. 2008;25(6):1244–55.
 95. Wolff RK. Effects of airborne pollutants on mucociliary clearance. *Environ Health Perspect*. 1986;66:223–37.
 96. Marcorelles P, Friocourt G, Uguen A, Ledé F, Férec C, Laquerrière A. Cystic Fibrosis Transmembrane Conductance Regulator Protein (CFTR) Expression in the Developing Human Brain: Comparative Immunohistochemical Study between Patients with Normal and Mutated CFTR. *J Histochem Cytochem*. 2014;62(11):791–801.
 97. Yin Z, Jiang H, Syversen T, Rocha JBT, Farina M, Aschner M. The methylmercury-L-cysteine conjugate is a substrate for the L-type large neutral amino acid transporter. *J Neurochem*. 2008;107(4):1083–90.
 98. Rossignol DA, Frye RE. Mitochondrial dysfunction in autism spectrum disorders: A systematic review and meta-analysis. *Mol Psychiatry*. 2012;17(3):290–314.
 99. Griffiths KK, Levy RJ. Evidence of mitochondrial dysfunction in autism: Biochemical links, genetic-based associations, and non-energy-related mechanisms. *Oxid Med Cell Longev*; 2017;2017:4314025.
 100. Luissint AC, Artus C, Glacial F, Ganeshamoorthy K, Couraud PO. Tight junctions at the blood brain barrier: Physiological architecture and disease-associated dysregulation. *Fluids Barriers CNS*. 2012;9(1):1–12.
 101. Fiorentino M, Sapone A, Senger S, Camhi SS, Kadzielski SM, Buie TM, et al. Blood-brain barrier and intestinal epithelial barrier alterations in autism spectrum disorders. *Mol Autism*. 2016;7(1):1–17.
 102. Shewchuk BM, Liebhaber SA, Cooke NE. Specification of unique Pit-1 activity in the hGH locus control region. *Proc Natl Acad Sci*. 2002;99(18):11784–9.
 103. Handwerker S, Freemark M. The Roles of Placental Growth Hormone and Placental Lactogen in the Regulation of Human Fetal Growth and Development. *J Pediatr Endocrinol Metab*.

2000;13:343.

104. Ishikawa T. Structural biology of cytoplasmic and axonemal dyneins. *J Struct Biol*. 2012;179(2):229–34.
105. Hull L, Mandy W. Protective effect or missed diagnosis? Females with autism spectrum disorder. *Future Neurol*. 2017;12(3):159–69.
106. Ingelman-Sundberg M, Mkrтчian S, Zhou Y, Lauschke VM. Integrating rare genetic variants into pharmacogenetic drug response predictions. *Hum Genomics*. 2018;12(1):26.
107. Lifantseva N, Koltsova A, Krylova T, Yakovleva T, Poljanskaya G, Gordeeva O. Expression patterns of cancer-testis antigens in human embryonic stem cells and their cell derivatives indicate lineage tracks. *Stem Cells Int*. 2011;2011:795239.
108. Macfarlane A, Greenhalgh T. Sodium valproate in pregnancy: What are the risks and should we use a shared decision-making approach? *BMC Pregnancy Childbirth*. *BMC Pregnancy and Childbirth*; 2018;18(1):1–11.
109. Nicolini C, Fahnstock M. The valproic acid-induced rodent model of autism. *Exp Neurol* 2018;299(Pt A):217–27.
110. Moorthy B, Chu C, Carlin DJ. Polycyclic aromatic hydrocarbons: From metabolism to lung cancer. *Toxicol Sci*. 2015;145(1):5–15.
111. Das SK, Patel B, Patri M. Neurotoxic effect of benzo[a]pyrene and its possible association with 6-hydroxydopamine induced neurobehavioral changes during early adolescence period in rats. *J Toxicol*. 2016;2016:8606410.
112. Kardas F, Bayram AK, Demirci E, Akin L, Ozmen S, Kendirci M, et al. Increased Serum Phthalates (MEHP, DEHP) and Bisphenol A Concentrations in Children with Autism Spectrum Disorder. *J Child Neurol*. 2016;31(5):629–35.
113. Kim KH, Kabir E, Kabir S. A review on the human health impact of airborne particulate matter. *Environ Int*. 2015;137(3):136–43.
114. Rappazzo KM, Coffman E, Hines EP. Exposure to perfluorinated alkyl substances and health outcomes in children: A systematic review of the epidemiologic literature. *Int J Environ Res Public Health*. 2017;14(7):1–22.
115. Landrigan PJ. What causes autism? Exploring the environmental contribution. *Curr Opin Pediatr*. 2010;22(2):219–25.
116. Volk HE, Kerin T, Lurmann F, Hertz-Picciotto I, McConnell R, Campbell DB. Autism spectrum disorder: interaction of air pollution with the MET receptor tyrosine kinase gene. *Epidemiology*. 2014;25(1):44–7.
117. Kim D, Volk H, Girirajan S, Pendergrass S, Hall MA, Verma SS, et al. The joint effect of air pollution exposure and copy number variation on risk for autism. *Autism Res*. 2017;10(9):1470–80.
118. Eshraghi AA, Liu G, Kay SS, Eshraghi RS, Mittal J, Moshiree B, et al. Epigenetics and Autism Spectrum Disorder: Is There a Correlation?. *Front Cell Neurosci*. 2018;12:78.
119. Li J, Li WX, Bai C, Song Y. Particulate matter-induced epigenetic changes and lung cancer. *Clin Respir J*. 2017;11(5):539–46.

120. Yaoi T, Itoh K, Nakamura K, Ogi H, Fujiwara Y, Fushiki S. Genome-wide analysis of epigenomic alterations in fetal mouse forebrain after exposure to low doses of bisphenol A. *Biochem Biophys Res Commun.* 2008;376(3):563–7.
121. Mileva G, Baker SL, Konkle ATM, Bielajew C. Bisphenol-A: Epigenetic Reprogramming and Effects on Reproduction and Behavior. *Int J Environ Res Public Health.* 2014;11(7):7537–61.
122. Casati L, Colciago A, Negri-Cesi P, Celotti F. Polychlorinated biphenyls affect histone modification pattern in early development of rats: a role for androgen receptor-dependent modulation?. *Epigenomics.* 2012;4(1):101-12

Supplementary data

Table S1: Individual chemicals potentially relevant for ASD studied for interactions with genes involved in detoxification and regulation of barriers permeability, using the Comparative Toxicogenomics Database.

Category	Individual chemical	MeSH ID	Category	Individual chemical	MeSH ID
Air Pollutants	Nitrogen dioxide (NO ₂)	D009585	Phthalates	Butylbenzyl phthalate (BBzP)	C027561
	Ozone (O ₃)	D010126		Dibutyl phthalate (DBP)	D003993
	Particulate matter (PM)	D052638		Diethyl phthalate (DEP)	C007379
	Vehicle emissions	D001335		Diethylhexyl phthalate (DEHP)	D004051
Bisphenol A	Bisphenol A (BPA)	C006780		Dimethyl phthalate (DMP)	C024629
				Mono-(2-ethylhexyl)phthalate (MEHP)	C016599
Heavy Metals	Lead (Pb)	D007854	Polybrominated diphenyl ethers	2,2',3,4,4'-pentabromodiphenyl ether (BDE-85)	C086401
	Manganese (Mn)	D008345		2,2',4,4',5-brominated diphenyl ether (BDE-99)	C477694
	Mercury (Hg)	D008628		2,2',4,4',6-brominated diphenyl ether (BDE-100)	C517827
	Methylmercuric chloride (CH ₃ ClHg)	C004925		2,2',4,4'-tetrabromodiphenyl ether (BDE-47)	C511295
	Methylmercury compounds (MeHg)	D008767		2,2',4-tribromodiphenyl ether (BDE-28)	C533760
Medications (Selective serotonin reuptake inhibitors (SSRIs) and teratogens)	Citalopram	D015283	Polychlorinated biphenyls	2,2',3,3',4,4',5-heptachlorobiphenyl (PCB-170)	C541131
	Fluoxetine	D005473		2,2',3,4,4',5,5'-heptachlorobiphenyl (PCB-180)	C410127
	Fluvoxamine	D016666		2,2',3',4,4',5-hexachlorobiphenyl (PCB-138)	C029790
	Paroxetine	D017374		2,3,3',4,4',5-hexachlorobiphenyl (PCB-156)	C087667
	Sertraline	D020280		2,3',4,4',5-pentachlorobiphenyl (PCB-118)	C070055
	Thalidomide	D013792		2,4,5,2',4',5'-hexachlorobiphenyl (PCB-153)	C014024
	Valproate (VPA)	D014635		3,4,3',4'-tetrachlorobiphenyl (PCB-77)	C028451
Nutritional factors	25-hydroxyvitamin D [25(OH)D]	C104450		3,4,5,3',4'-pentachlorobiphenyl (PCB-126)	C023035
	Folic acid	D005492	Polycyclic Aromatic Hydrocarbons	Benzo(a)anthracene [b(a)a]	C030935
	Vitamin D	D014807		Benzo(a)pyrene [b(a)p]	D001564
Pesticides	Chlorpyrifos	D004390		Benzo(b)fluoranthene [b(b)f]	C006703
	Dichlorodiphenyldichloroethylene (DDE)	D003633		Benzo(g,h,i)perylene [b(g,h,i)p]	C006718
	Dicofol	D004010		Benzo(k)fluoranthene [b(k)f]	C022921
	Endosulfan	D004726		Chrysene	C031180
	Nonachlor	C001870		Dibenzo(a,h)anthracene [d(a,h)a]	C026486
Perfluorinated compounds	Perfluorooctane sulfonic acid (PFOS)	C076994		Indeno(1,2,3,-cd)pyrene [i(1,2,3,-cd)p]	C041508
	Perfluorooctanoic acid (PFOA)	C023036			

Table S2: Frequency of genes targeted by CNVs from individuals with ASD (from both AGP and/or SSC datasets) and control datasets, but that failed to achieve statistical significance after Bonferroni correction for multiple testing. Presented are only genes that do not appear in table 5.

Gene list	AGP dataset				SSC dataset				Control dataset
	AGP N (%)	Test statistic	p-value	OR (95% CI)	SSC N (%)	Test statistic	p-value	OR (95% CI)	DGV N (%)
<i>ABCA2</i>	7 (0.286)	8.58	0.003402	5.54 (1.76-17.46)	2 (0.178)	0.91	0.160502	3.44 (0.67-17.74)	5 (0.052)
<i>ADAMTS18</i>	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	2 (0.178)	5.03	0.030365	17.2 (1.56-189.82)	1 (0.01)
<i>ADORA2A</i>	1 (0.041)	3.75x10 ⁻²⁵	1	0.99 (0.11-8.83)	1 (0.089)	1.72x10 ⁻²⁴	0.423657	2.15 (0.24-19.23)	4 (0.041)
<i>AHR</i>	2 (0.082)	7.16x10 ⁻²⁷	1	1.13 (0.23-5.43)	1 (0.089)	3.80x10 ⁻³¹	0.585968	1.23 (0.15-9.98)	7 (0.073)
<i>AKR1C2</i>	2 (0.082)	0.08	0.350127	1.97 (0.36-10.78)	1 (0.089)	1.72x10 ⁻²⁴	0.423657	2.15 (0.24-19.23)	4 (0.041)
<i>ALDH2</i>	9 (0.368)	7.09	0.00776	3.56 (1.44-8.77)	6 (0.53)	9.83	0.001718	5.17 (1.88-14.26)	10 (0.104)
<i>ALDH3B2</i>	3 (0.123)	0.32	0.398461	1.97 (0.49-7.9)	1 (0.089)	1.22x10 ⁻²⁴	0.537703	1.43 (0.17-11.9)	6 (0.062)
<i>ALDH5A1</i>	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	1 (0.089)	0.45	0.197793	8.59 (0.54-137.45)	1 (0.01)
<i>ALDH7A1</i>	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	2 (0.178)	5.03	0.030365	17.2 (1.56-189.82)	1 (0.01)
<i>ALPP</i>	11 (0.45)	2.40	0.121247	1.89 (0.92-3.88)	1 (0.089)	0.45	0.506186	0.37 (0.05-2.76)	23 (0.238)
<i>ANKRD11</i>	1 (0.041)	2.56x10 ⁻³⁰	1	0.79 (0.09-6.76)	1 (0.089)	5.42x10 ⁻³¹	0.483817	1.72 (0.2-14.71)	5 (0.052)
<i>ARSD</i>	2 (0.082)	1.65	0.106128	7.9 (0.72-87.11)	3 (0.267)	11.61	0.004178	25.82 (2.68-248.44)	1 (0.01)
<i>ATP1B4</i>	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	1 (0.089)	0.45	0.197793	8.59 (0.54-137.45)	1 (0.01)
<i>B3GAT1</i>	3 (0.123)	0.60	0.207965	2.37 (0.57-9.92)	1 (0.089)	5.42x10 ⁻³¹	0.483817	1.72 (0.2-14.71)	5 (0.052)
<i>B3GAT2</i>	2 (0.082)	0.74	0.184213	3.95 (0.56-28.04)	1 (0.089)	0.12	0.281506	4.3 (0.39-47.41)	2 (0.021)
<i>CFTR</i>	2 (0.082)	1.65	0.106128	7.9 (0.72-87.11)	1 (0.089)	0.45	0.197793	8.59 (0.54-137.45)	1 (0.01)
<i>CGNLI</i>	2 (0.082)	3.06	0.047601	0.26 (0.06-1.1)	6 (0.534)	0.91	0.340893	1.72 (0.71-4.14)	30 (0.311)
<i>CHST11</i>	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	2 (0.178)	5.03	0.030365	17.2 (1.56-189.82)	1 (0.01)
<i>CHST6</i>	2 (0.082)	0.085	0.350127	1.97 (0.36-10.78)	1 (0.089)	1.72x10 ⁻²⁴	0.423657	2.15 (0.24-19.23)	4 (0.041)
<i>CHST8</i>	1 (0.041)	2.56x10 ⁻³⁰	1	0.79 (0.09-6.76)	1 (0.089)	5.42x10 ⁻³¹	0.483817	1.72 (0.2-14.71)	5 (0.052)
<i>COMT</i>	9 (0.368)	12.36	0.000438	5.94 (2.11-16.69)	1 (0.089)	1.22x10 ⁻²⁴	0.537703	1.43 (0.17-11.9)	6 (0.062)
<i>CYP2B6</i>	3 (0.123)	0.043	0.47322	1.48 (0.39-5.58)	1 (0.089)	9.37x10 ⁻²⁵	1	1.07 (0.13-8.59)	8 (0.083)
<i>CYP4F12</i>	9 (0.368)	0.63	0.428012	1.48 (0.69-3.19)	1 (0.089)	9.52	0.50918	0.36 (0.05-2.64)	24 (0.249)
<i>CYP4F3</i>	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	1 (0.089)	0.45	0.197793	8.59 (0.54-137.45)	1 (0.01)
<i>CYP4V2</i>	2 (0.082)	0.0063	0.635006	1.58 (0.31-8.14)	1 (0.089)	5.42x10 ⁻³¹	0.483817	1.72 (0.2-14.71)	5 (0.052)
<i>DHRS4</i>	6 (0.245)	11.68	0.001305	11.86 (2.39-58.8)	1 (0.089)	0.12	0.281506	4.3 (0.39-47.41)	2 (0.021)
<i>DNAI2</i>	4 (0.164)	2.21	1	1.05 (0.35-3.17)	1 (0.089)	0.019	1	0.57 (0.08-4.33)	15 (0.155)
<i>DYX1C1</i>	1 (0.041)	2.39	0.492306	1.97 (0.18-21.77)	1 (0.089)	0.12	0.281506	4.3 (0.39-47.41)	2 (0.021)
<i>FMO5</i>	5 (0.204)	11.16	0.001683	19.76 (2.31-169.24)	3 (0.267)	11.61	0.004178	25.82 (2.68-248.44)	1 (0.01)
<i>GABRA5</i>	7 (0.286)	0.21	0.649191	0.77 (0.34-1.72)	1 (0.089)	1.62	0.174416	0.24 (0.03-1.74)	36 (0.373)

GABRB3	7 (0.286)	10.31	0.002094	6.92 (2.02-23.66)	1 (0.089)	1.72x10 ⁻²⁴	0.423657	2.15 (0.24-19.23)	4 (0.041)
GABRG3	7 (0.286)	4.00	1	0.95 (0.42-2.18)	2 (0.178)	0.19	0.766294	0.59 (0.14-2.48)	29 (0.301)
GAL3ST2	5 (0.204)	11.16	0.001683	19.76 (2.31-169.24)	4 (0.356)	18.99	0.000541	34.46 (3.85-308.57)	1 (0.01)
GSTM5	1 (0.041)	2.01	0.100068	0.21 (0.03-1.55)	2 (0.178)	1.43x10 ⁻²⁷	1	0.9 (0.21-3.88)	19 (0.197)
HBG1	4 (0.164)	1.95	0.088543	3.16 (0.85-11.77)	3 (0.267)	3.71	0.042465	5.16 (1.23-21.63)	5 (0.052)
HBG2	4 (0.164)	1.95	0.088543	3.16 (0.85-11.77)	3 (0.267)	3.71	0.042465	5.16 (1.23-21.63)	5 (0.052)
IDS	2 (0.082)	0.74	0.184213	3.95 (0.56-28.04)	1 (0.089)	0.12	0.281506	4.3 (0.39-47.41)	2 (0.021)
INSL4	1 (0.041)	6.58x10 ⁻²⁵	1	1.32 (0.14-12.65)	1 (0.089)	0.018	0.356491	2.86 (0.3-27.55)	3 (0.031)
MIF	2 (0.082)	0.085	0.350127	1.97 (0.36-10.78)	1 (0.089)	1.72x10 ⁻²⁴	0.423657	2.15 (0.24-19.23)	4 (0.041)
MTNRIA	2 (0.082)	1.65	0.106128	7.9 (0.72-87.11)	2 (0.178)	5.03	0.030365	17.2 (1.56-189.82)	1 (0.01)
NOS3	1 (0.041)	0.17	0.697933	0.44 (0.06-3.46)	1 (0.089)	6.41x10 ⁻²⁸	1	0.95 (0.12-7.54)	9 (0.093)
OCLN	26 (1.063)	9.85	0.0017	2.19 (1.36-3.55)	5 (0.445)	3.37x10 ⁻²⁷	1	0.91 (0.36-2.3)	47 (0.487)
PARD3	3 (0.123)	6.25x10 ⁻²⁶	1	0.85 (0.24-2.94)	1 (0.089)	0.003	1	0.61 (0.08-4.66)	14 (0.145)
PRKCE	5 (0.204)	0.25	0.613609	1.52 (0.54-4.26)	3 (0.267)	0.46	0.229386	1.98 (0.56-6.97)	13 (0.135)
SCNN1D	9 (0.368)	10.75	0.001043	5.09 (1.89-13.67)	4 (0.356)	5.39	0.021342	4.92 (1.44-16.83)	7 (0.073)
SLCIA1	2 (0.082)	0.0063	0.635006	1.58 (0.31-8.14)	4 (0.356)	7.80	0.009658	6.89 (1.85-25.69)	5 (0.052)
SLCO1C1	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	3 (0.267)	11.61	0.004178	25.82 (2.68-248.44)	1 (0.01)
TAC3	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	1 (0.089)	0.45	0.197793	8.59 (0.54-137.45)	1 (0.01)
TFRC	2 (0.082)	1.65	0.106128	7.9 (0.72-87.11)	1 (0.089)	0.45	0.197793	8.59 (0.54-137.45)	1 (0.01)
TXNRD2	8 (0.327)	13.31	0.000625	7.91 (2.38-26.3)	1 (0.089)	1.72x10 ⁻²⁴	0.423657	2.15 (0.24-19.23)	4 (0.041)
AKR1B10	1 (0.041)	2.39x10 ⁻²⁷	0.492306	1.97 (0.18-21.77)	0 (0)	n/a	n/a	n/a	2 (0.021)
AKR1B15	1 (0.041)	2.39x10 ⁻²⁷	0.492306	1.97 (0.18-21.77)	0 (0)	n/a	n/a	n/a	2 (0.021)
AKR1C1	1 (0.041)	2.39x10 ⁻²⁷	0.492306	1.97 (0.18-21.77)	0 (0)	n/a	n/a	n/a	2 (0.021)
ALDH1A3	1 (0.041)	8.48	0.000717	0.09 (0.01-0.62)	0 (0)	n/a	n/a	n/a	46 (0.477)
ALDH3A1	2 (0.082)	0.74	0.184213	3.95 (0.56-28.04)	0 (0)	n/a	n/a	n/a	2 (0.021)
ALOX5	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	0 (0)	n/a	n/a	n/a	1 (0.01)
ANKRD33	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	0 (0)	n/a	n/a	n/a	1 (0.01)
ARSA	2 (0.082)	0.30	0.267282	2.63 (0.44-15.76)	0 (0)	n/a	n/a	n/a	3 (0.031)
ARSJ	1 (0.041)	2.39x10 ⁻²⁷	0.492306	1.97 (0.18-21.77)	0 (0)	n/a	n/a	n/a	2 (0.021)
BCHE	1 (0.041)	2.56x10 ⁻³⁰	1	0.79 (0.09-6.76)	0 (0)	n/a	n/a	n/a	5 (0.052)
CBR3	7 (0.286)	1.87	0.171303	2.13 (0.85-5.34)	0 (0)	n/a	n/a	n/a	13 (0.135)
CBS	2 (0.082)	0.085	0.350127	1.97 (0.36-10.78)	0 (0)	n/a	n/a	n/a	4 (0.041)
CES2	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	0 (0)	n/a	n/a	n/a	1 (0.01)
CHAT	3 (0.123)	4.43	0.028043	11.85 (1.23-113.95)	0 (0)	n/a	n/a	n/a	1 (0.01)
CHST10	2 (0.082)	0.74	0.184213	3.95 (0.56-28.04)	0 (0)	n/a	n/a	n/a	2 (0.021)

CRYZ	2 (0.082)	1.65	0.106128	7.9 (0.72-87.11)	0 (0)	n/a	n/a	n/a	1 (0.01)
CYP11B1	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	0 (0)	n/a	n/a	n/a	1 (0.01)
CYP24A1	1 (0.041)	2.39x10-27	0.492306	1.97 (0.18-21.77)	0 (0)	n/a	n/a	n/a	2 (0.021)
CYP27B1	1 (0.041)	2.39x10-27	0.492306	1.97 (0.18-21.77)	0 (0)	n/a	n/a	n/a	2 (0.021)
CYP2C18	2 (0.082)	0.049	0.74907	0.66 (0.15-2.94)	0 (0)	n/a	n/a	n/a	12 (0.124)
CYP2C19	2 (0.082)	0.12	0.749765	0.61 (0.14-2.69)	0 (0)	n/a	n/a	n/a	13 (0.135)
CYP2W1	1 (0.041)	3.75x10-25	1	0.99 (0.11-8.83)	0 (0)	n/a	n/a	n/a	4 (0.041)
CYP46A1	1 (0.041)	2.39x10-27	0.492306	1.97 (0.18-21.77)	0 (0)	n/a	n/a	n/a	2 (0.021)
CYP7B1	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	0 (0)	n/a	n/a	n/a	1 (0.01)
DCXR	2 (0.082)	0.085	0.350127	1.97 (0.36-10.78)	0 (0)	n/a	n/a	n/a	4 (0.041)
DRD4	1 (0.041)	4.44x10-26	1	0.66 (0.08-5.46)	0 (0)	n/a	n/a	n/a	6 (0.062)
EPHX4	1 (0.041)	4.44x10-26	1	0.66 (0.08-5.46)	0 (0)	n/a	n/a	n/a	6 (0.062)
FBN2	2 (0.082)	0.74	0.184213	3.95 (0.56-28.04)	0 (0)	n/a	n/a	n/a	2 (0.021)
FCGR2B	5 (0.204)	2.22	0.136027	2.82 (0.89-8.9)	0 (0)	n/a	n/a	n/a	7 (0.073)
GPR32	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	0 (0)	n/a	n/a	n/a	1 (0.01)
GPX5	1 (0.041)	2.39x10-27	0.492306	1.97 (0.18-21.77)	0 (0)	n/a	n/a	n/a	2 (0.021)
GPX6	1 (0.041)	2.39x10-27	0.492306	1.97 (0.18-21.77)	0 (0)	n/a	n/a	n/a	2 (0.021)
GSTO1	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	0 (0)	n/a	n/a	n/a	1 (0.01)
GSTO2	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	0 (0)	n/a	n/a	n/a	1 (0.01)
GSTP1	2 (0.082)	0.74	0.184213	3.95 (0.56-28.04)	0 (0)	n/a	n/a	n/a	2 (0.021)
IGF2	1 (0.041)	6.58x10-25	1	1.32 (0.14-12.65)	0 (0)	n/a	n/a	n/a	3 (0.031)
IYD	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	0 (0)	n/a	n/a	n/a	1 (0.01)
JAM3	1 (0.041)	2.39x10-27	0.492306	1.97 (0.18-21.77)	0 (0)	n/a	n/a	n/a	2 (0.021)
KIF17	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	0 (0)	n/a	n/a	n/a	1 (0.01)
LRP1	1 (0.041)	1.17	0.221204	0.26 (0.03-1.99)	0 (0)	n/a	n/a	n/a	15 (0.155)
MAPK8IP2	2 (0.082)	0.74	0.184213	3.95 (0.56-28.04)	0 (0)	n/a	n/a	n/a	2 (0.021)
NOTCH1	6 (0.245)	6.05	0.0139	4.74 (1.45-15.55)	0 (0)	n/a	n/a	n/a	5 (0.052)
PAOX	4 (0.164)	6.49	0.004026	0.27 (0.1-0.75)	0 (0)	n/a	n/a	n/a	58 (0.601)
PARD6G	1 (0.041)	2.39x10-27	0.492306	1.97 (0.18-21.77)	0 (0)	n/a	n/a	n/a	2 (0.021)
PHLDA2	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	0 (0)	n/a	n/a	n/a	1 (0.01)
PSG8	25 (1.022)	2.75	0.097063	1.52 (0.96-2.42)	0 (0)	n/a	n/a	n/a	65 (0.674)
PSG9	30 (1.226)	6.96	0.008334	1.83 (1.19-2.83)	0 (0)	n/a	n/a	n/a	65 (0.674)
SHANK2	3 (0.123)	0.32	0.398461	1.97 (0.49-7.9)	0 (0)	n/a	n/a	n/a	6 (0.062)
SLC12A7	3 (0.123)	1.71	0.18304	0.41 (0.12-1.34)	0 (0)	n/a	n/a	n/a	29 (0.301)

<i>SLC19A1</i>	6 (0.245)	0.11	0.742279	1.32 (0.52-3.32)	0 (0)	n/a	n/a	n/a	18 (0.187)
<i>SLC7A5</i>	1 (0.041)	2.39x10 ⁻²⁷	0.492306	1.97 (0.18-21.77)	0 (0)	n/a	n/a	n/a	2 (0.021)
<i>SULF1</i>	1 (0.041)	2.39x10 ⁻²⁷	0.492306	1.97 (0.18-21.77)	0 (0)	n/a	n/a	n/a	2 (0.021)
<i>SULF2</i>	2 (0.082)	1.65	0.106128	7.9 (0.72-87.11)	0 (0)	n/a	n/a	n/a	1 (0.01)
<i>SULT1B1</i>	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	0 (0)	n/a	n/a	n/a	1 (0.01)
<i>SULT2A1</i>	2 (0.082)	0.085	0.350127	1.97 (0.36-10.78)	0 (0)	n/a	n/a	n/a	4 (0.041)
<i>TJP1</i>	3 (0.123)	4.43	0.028043	11.85 (1.23-113.95)	0 (0)	n/a	n/a	n/a	1 (0.01)
<i>TXNRD3</i>	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	0 (0)	n/a	n/a	n/a	1 (0.01)
<i>UGT2A1</i>	2 (0.082)	6.02	0.004728	0.18 (0.04-0.76)	0 (0)	n/a	n/a	n/a	43 (0.446)
<i>UGT2A2</i>	2 (0.082)	5.79	0.007233	0.19 (0.05-0.77)	0 (0)	n/a	n/a	n/a	42 (0.435)
<i>UGT2B11</i>	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	0 (0)	n/a	n/a	n/a	1 (0.01)
<i>UGT2B4</i>	1 (0.041)	0.028	0.36358	3.95 (0.25-63.11)	0 (0)	n/a	n/a	n/a	1 (0.01)
<i>ABCA8</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	0.12	0.281506	4.3 (0.39-47.41)	2 (0.021)
<i>AKR1C4</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	0.018	0.356491	2.86 (0.3-27.55)	3 (0.031)
<i>ALDH18A1</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	0.45	0.197793	8.59 (0.54-137.45)	1 (0.01)
<i>ALDH1A2</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	1.72x10 ⁻²⁴	0.423657	2.15 (0.24-19.23)	4 (0.041)
<i>ALDH3B1</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	3.80x10 ⁻²⁴	0.585968	1.23 (0.15-9.98)	7 (0.073)
<i>ALDH8A1</i>	0 (0)	n/a	n/a	n/a	4 (0.356)	14.74	0.001487	17.23 (3.15-94.16)	2 (0.021)
<i>ALOX12</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	0.45	0.197793	8.59 (0.54-137.45)	1 (0.01)
<i>ALOXE3</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	0.12	0.281506	4.3 (0.39-47.41)	2 (0.021)
<i>CHST9</i>	0 (0)	n/a	n/a	n/a	2 (0.178)	5.03	0.030365	17.2 (1.56-189.82)	1 (0.01)
<i>CYP2U1</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	0.45	0.197793	8.59 (0.54-137.45)	1 (0.01)
<i>DHRS3</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	1.72x10 ⁻²⁴	0.423657	2.15 (0.24-19.23)	4 (0.041)
<i>DNAH5</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	5.42x10 ⁻³¹	0.483817	1.72 (0.2-14.71)	5 (0.052)
<i>GALNS</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	1.72x10 ⁻²⁴	0.423657	2.15 (0.24-19.23)	4 (0.041)
<i>GHI</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	0.12	0.281506	4.3 (0.39-47.41)	2 (0.021)
<i>HSD3B1</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	0.45	0.197793	8.59 (0.54-137.45)	1 (0.01)
<i>IFT74</i>	0 (0)	n/a	n/a	n/a	4 (0.356)	7.80	0.009658	6.89 (1.85-25.69)	5 (0.052)
<i>JAG2</i>	0 (0)	n/a	n/a	n/a	4 (0.356)	0.02	1	0.82 (0.29-2.28)	42 (0.435)
<i>KISS1</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	0.12	0.281506	4.3 (0.39-47.41)	2 (0.021)
<i>NAT1</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	0.018	0.356491	2.86 (0.3-27.55)	3 (0.031)
<i>NOS1AP</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	9.37x10 ⁻²⁵	1	1.07 (0.13-8.59)	8 (0.083)
<i>NOTCH3</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	0.45	0.197793	8.59 (0.54-137.45)	1 (0.01)
<i>NQO2</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	0.45	0.197793	8.59 (0.54-137.45)	1 (0.01)

<i>NTRK2</i>	0 (0)	n/a	n/a	n/a	3 (0.267)	11.61	0.004178	25.82 (2.68-248.44)	1 (0.01)
<i>PTGES2</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	0.45	0.197793	8.59 (0.54-137.45)	1 (0.01)
<i>PTGS2</i>	0 (0)	n/a	n/a	n/a	3 (0.267)	3.71	0.042465	5.16 (1.23-21.63)	5 (0.052)
<i>SGSH</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	0.12	0.281506	4.3 (0.39-47.41)	2 (0.021)
<i>SLC25A25</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	0.12	0.281506	4.3 (0.39-47.41)	2 (0.021)
<i>SLC7A1</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	0.45	0.197793	8.59 (0.54-137.45)	1 (0.01)
<i>SULT1C2</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	0.12	0.281506	4.3 (0.39-47.41)	2 (0.021)
<i>SULT1C3</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	0.45	0.197793	8.59 (0.54-137.45)	1 (0.01)
<i>SULT1C4</i>	0 (0)	n/a	n/a	n/a	1 (0.089)	0.45	0.197793	8.59 (0.54-137.45)	1 (0.01)

Table S3: Numbers of variants by impact and by rank for the 380 genes not targeted by variants from rank A. Genes are ordered by the amount of high + moderate impact variants they have.

Gene	Number of variants by impact			N° of variants by rank				
	High Impact Variants	Moderate Impact Variants	High + Moderate Impact Variants	B	C	D	E	F
<i>DNAH7</i>	10	98	108	4	3	1	76	24
<i>DNAH11</i>	7	68	75	4	0	0	64	7
<i>VWF</i>	3	40	43	0	1	2	35	5
<i>TJP3</i>	4	30	34	2	2	2	25	3
<i>ABCC1</i>	0	31	31	1	0	0	26	4
<i>ATPIA4</i>	5	26	31	0	1	1	20	9
<i>ABCC3</i>	4	26	30	1	1	2	23	3
<i>CCDC40</i>	2	27	29	1	1	0	25	2
<i>NOTCH3</i>	0	29	29	0	1	2	22	4
<i>NOTCH1</i>	0	26	26	0	0	0	21	5
<i>AOC2</i>	4	20	24	2	0	0	20	2
<i>CGN</i>	3	21	24	3	1	2	12	6
<i>ADAMTS18</i>	3	20	23	0	1	0	19	3
<i>AOX1</i>	9	14	23	1	0	1	21	0
<i>PAPPA2</i>	1	22	23	0	1	0	19	3
<i>SLC12A7</i>	2	21	23	2	0	0	20	1
<i>MPDZ</i>	1	21	22	0	1	3	16	2
<i>ANKRD11</i>	0	21	21	0	1	1	18	1
<i>CGNL1</i>	2	19	21	0	0	1	15	5
<i>LOXL2</i>	1	20	21	1	0	0	18	2
<i>ADCY10</i>	3	17	20	0	0	0	18	2
<i>FBN2</i>	0	20	20	0	0	0	18	2
<i>ALDH3B2</i>	1	18	19	3	1	3	8	4
<i>FMO2</i>	3	16	19	2	2	1	10	4
<i>NOTCH4</i>	2	17	19	0	3	1	14	1
<i>ALDH4A1</i>	0	18	18	0	0	1	13	4
<i>CYP2C8</i>	3	15	18	0	1	1	13	3
<i>CYP4F12</i>	1	17	18	2	0	0	14	2
<i>FMO4</i>	1	17	18	1	1	1	13	2
<i>NOS3</i>	0	18	18	1	1	0	12	4
<i>PAPPA</i>	1	17	18	0	0	0	16	2
<i>TBXAS1</i>	2	15	17	0	0	0	12	5
<i>TJPI</i>	0	17	17	3	0	1	7	6
<i>UGT2B10</i>	3	14	17	0	1	1	12	3
<i>CYP1A2</i>	4	12	16	1	1	0	10	4
<i>CYP24A1</i>	2	14	16	1	1	0	10	4
<i>CYP2A13</i>	1	15	16	5	1	0	8	2
<i>CYP2B6</i>	3	13	16	0	1	2	11	2
<i>CYP4F11</i>	1	15	16	0	1	0	13	2
<i>DHRS2</i>	3	13	16	1	0	2	12	1
<i>NOS2</i>	2	14	16	1	0	2	12	1
<i>NOTCH2</i>	0	16	16	0	1	1	13	1
<i>SCNN1A</i>	1	15	16	0	1	1	12	2
<i>UGT2B15</i>	4	12	16	1	1	0	8	6
<i>CP</i>	2	13	15	2	0	3	8	2
<i>CYP1A1</i>	0	15	15	4	2	0	7	2
<i>CYP4F3</i>	1	14	15	0	2	0	8	5

<i>JUP</i>	1	14	15	0	0	0	13	2
<i>UGT2A1</i>	3	12	15	0	0	1	11	3
<i>ARSI</i>	1	13	14	0	0	0	12	2
<i>CYP2F1</i>	1	13	14	0	0	1	11	2
<i>POMT2</i>	1	13	14	0	0	0	13	1
<i>ALDH9A1</i>	3	10	13	0	0	1	9	3
<i>ALOX15B</i>	2	11	13	0	0	1	10	2
<i>CYP2C9</i>	1	12	13	2	2	0	7	2
<i>CYP7A1</i>	0	13	13	1	0	0	10	2
<i>FMO5</i>	4	9	13	0	1	0	11	1
<i>MTR</i>	0	13	13	0	0	0	11	2
<i>PARD3</i>	0	13	13	0	0	0	11	2
<i>ABCA2</i>	0	12	12	0	0	0	10	2
<i>ALOXE3</i>	1	11	12	1	1	0	7	3
<i>AOC1</i>	1	11	12	1	0	0	10	1
<i>CDH5</i>	2	10	12	0	0	0	10	2
<i>CES2</i>	0	12	12	1	0	0	10	1
<i>CHST6</i>	0	12	12	1	1	0	6	4
<i>CYP4V2</i>	1	11	12	0	0	1	10	1
<i>POMT1</i>	3	9	12	0	0	1	9	2
<i>PSG3</i>	2	10	12	2	0	0	6	4
<i>SLCO1C1</i>	1	11	12	0	0	1	10	1
<i>UGT2A3</i>	3	9	12	0	0	0	7	5
<i>ALDH3A1</i>	2	9	11	0	0	0	8	3
<i>ALDH5A1</i>	5	6	11	0	0	2	8	1
<i>BCHE</i>	0	11	11	1	1	1	8	0
<i>CYP2C18</i>	1	10	11	0	0	0	9	2
<i>CYP3A43</i>	1	10	11	0	0	2	8	1
<i>DNAI1</i>	3	8	11	2	0	0	8	1
<i>PSG7</i>	2	9	11	0	1	1	8	1
<i>RSPH4A</i>	1	10	11	1	0	0	8	2
<i>SHANK2</i>	1	10	11	0	0	0	10	1
<i>UGT1A6</i>	0	11	11	0	0	0	10	1
<i>UGT2B7</i>	0	11	11	0	0	1	9	1
<i>ABCB1</i>	0	10	10	0	0	0	10	0
<i>AKR1A1</i>	2	8	10	1	0	1	5	3
<i>AKR1C3</i>	3	7	10	1	0	1	6	2
<i>AKR7A3</i>	2	8	10	0	0	0	7	3
<i>ALDH1B1</i>	2	8	10	0	0	0	6	4
<i>ALOX12B</i>	1	9	10	0	0	0	10	0
<i>ALPP</i>	1	9	10	0	1	0	7	2
<i>CYP11A1</i>	0	10	10	0	0	0	10	0
<i>CYP3A4</i>	1	9	10	0	0	0	6	4
<i>DRD4</i>	0	10	10	0	0	0	10	0
<i>DYX1C1</i>	1	9	10	0	0	1	7	2
<i>EPHX1</i>	0	10	10	1	1	0	8	0
<i>EPHX2</i>	2	8	10	0	1	1	6	2
<i>FMO1</i>	3	7	10	0	0	0	9	1
<i>FMO3</i>	1	9	10	0	0	0	8	2
<i>IL1RL1</i>	0	10	10	0	0	0	9	1
<i>PSG9</i>	4	6	10	0	1	0	6	3
<i>SCNN1B</i>	1	9	10	0	0	0	8	2
<i>TJP2</i>	0	10	10	1	0	0	9	0
<i>ALDH2</i>	0	9	9	1	0	0	8	0

<i>ALDH8A1</i>	0	9	9	0	0	0	7	2
<i>ARSJ</i>	2	7	9	0	0	0	7	2
<i>CCDC114</i>	0	9	9	0	0	0	6	3
<i>CCDC39</i>	1	8	9	2	0	2	5	0
<i>CHAT</i>	1	8	9	0	0	3	3	3
<i>CHST4</i>	0	9	9	0	0	0	8	1
<i>CYP11B2</i>	0	9	9	0	0	0	9	0
<i>CYP26B1</i>	0	9	9	1	0	1	6	1
<i>CYP2A6</i>	2	7	9	0	0	1	8	0
<i>CYP39A1</i>	1	8	9	0	0	2	4	3
<i>GPX6</i>	3	6	9	2	0	0	6	1
<i>GSTA3</i>	0	9	9	0	2	1	3	3
<i>GSTO2</i>	4	5	9	1	0	0	6	2
<i>HSD17B1</i>	0	9	9	1	0	0	8	0
<i>HTRA4</i>	0	9	9	1	1	0	7	0
<i>INMT</i>	2	7	9	0	0	2	6	1
<i>LRRC6</i>	1	8	9	0	0	1	8	0
<i>MAPK8IP2</i>	0	9	9	0	0	2	6	1
<i>MET</i>	0	9	9	1	0	1	6	1
<i>MTRR</i>	2	7	9	0	0	0	7	2
<i>PON2</i>	2	7	9	0	0	0	7	2
<i>PON3</i>	3	6	9	0	1	0	6	2
<i>PSG8</i>	5	4	9	0	0	0	7	2
<i>SCNN1D</i>	0	9	9	0	0	0	7	2
<i>SLC27A1</i>	0	9	9	0	0	0	8	1
<i>SLCO2A1</i>	0	9	9	0	0	0	8	1
<i>SULF2</i>	0	9	9	0	0	0	9	0
<i>TPMT</i>	0	9	9	0	1	3	5	0
<i>ADAM12</i>	1	7	8	0	1	0	7	0
<i>ADH1B</i>	1	7	8	0	1	1	5	1
<i>AKR1C4</i>	3	5	8	0	0	1	7	0
<i>ALDH6A1</i>	1	7	8	1	0	0	5	2
<i>ALOX12</i>	0	8	8	0	1	0	6	1
<i>ANKRD33</i>	0	8	8	0	0	0	5	3
<i>CYP2A7</i>	2	6	8	0	0	2	4	2
<i>CYP2E1</i>	0	8	8	0	0	0	7	1
<i>CYP2J2</i>	1	7	8	0	0	0	6	2
<i>CYP3A5</i>	1	7	8	1	0	0	5	2
<i>CYP4A22</i>	0	8	8	1	2	0	2	3
<i>CYP4F22</i>	2	6	8	0	0	0	7	1
<i>GFAP</i>	0	8	8	0	0	1	6	1
<i>HSD3B1</i>	0	8	8	0	0	0	5	3
<i>INSR</i>	0	8	8	0	0	0	6	2
<i>LGALS13</i>	2	6	8	0	1	0	7	0
<i>NAT2</i>	1	7	8	0	0	1	7	0
<i>NQO2</i>	3	5	8	0	0	0	8	0
<i>SCNN1G</i>	0	8	8	0	0	0	7	1
<i>SLC25A25</i>	0	8	8	0	0	0	7	1
<i>UGT2B4</i>	0	8	8	1	0	0	6	1
<i>ADH5</i>	2	5	7	0	1	0	5	1
<i>ADSL</i>	0	7	7	0	0	0	6	1
<i>AHR</i>	0	7	7	0	0	0	7	0
<i>ALOX5</i>	0	7	7	0	0	0	6	1
<i>ARSB</i>	0	7	7	0	0	0	7	0

<i>ARSG</i>	3	4	7	0	0	0	6	1
<i>ARSK</i>	0	7	7	1	1	0	4	1
<i>CBR3</i>	3	4	7	0	0	0	6	1
<i>CTH</i>	2	5	7	0	0	1	5	1
<i>CYP19A1</i>	0	7	7	1	0	0	6	0
<i>CYP2R1</i>	0	7	7	0	0	0	7	0
<i>CYP3A7</i>	1	6	7	0	0	0	6	1
<i>CYP7B1</i>	0	7	7	1	0	0	6	0
<i>CYP8B1</i>	0	7	7	0	0	0	5	2
<i>DNMT1</i>	0	7	7	1	0	0	6	0
<i>FOXP2</i>	1	6	7	0	1	0	6	0
<i>GSTA5</i>	2	5	7	0	0	0	4	3
<i>JAG2</i>	0	7	7	0	0	0	5	2
<i>LOXL1</i>	0	7	7	0	0	0	6	1
<i>NOTUM</i>	1	6	7	0	0	1	6	0
<i>NQO1</i>	1	6	7	1	2	0	3	1
<i>PNMT</i>	0	7	7	0	0	1	2	4
<i>PSG4</i>	1	6	7	1	0	0	5	1
<i>PSG5</i>	2	5	7	0	0	0	6	1
<i>SGSH</i>	0	7	7	1	0	1	5	0
<i>SLC1A3</i>	0	7	7	0	0	0	7	0
<i>SULT1C3</i>	1	6	7	3	0	0	2	2
<i>SULT2A1</i>	1	6	7	1	0	1	5	0
<i>TF</i>	0	7	7	0	0	1	5	1
<i>ACHE</i>	0	6	6	0	0	1	5	0
<i>ACKR2</i>	0	6	6	0	0	0	4	2
<i>ADH4</i>	0	6	6	2	0	0	3	1
<i>AKR1C1</i>	2	4	6	0	0	0	5	1
<i>AKR7A2</i>	1	5	6	1	0	0	5	0
<i>ALAD</i>	0	6	6	0	0	0	6	0
<i>ALDH7A1</i>	0	6	6	0	0	0	4	2
<i>CBR1</i>	1	5	6	0	0	1	5	0
<i>CCDC151</i>	1	5	6	0	0	0	6	0
<i>CHST9</i>	0	6	6	0	0	1	4	1
<i>CYP27C1</i>	1	5	6	0	0	0	6	0
<i>CYP4A11</i>	3	3	6	0	0	0	5	1
<i>CYP4F2</i>	1	5	6	2	0	1	2	1
<i>CYP4X1</i>	1	5	6	0	0	0	6	0
<i>DHRS1</i>	0	6	6	0	0	0	5	1
<i>DNAAF1</i>	1	5	6	1	0	0	4	1
<i>DNAI2</i>	2	4	6	2	0	1	3	0
<i>EPYC</i>	1	5	6	0	0	2	2	2
<i>FOLR2</i>	1	5	6	0	0	0	5	1
<i>GAL3ST3</i>	0	6	6	1	0	0	4	1
<i>GSTA1</i>	0	6	6	0	0	0	6	0
<i>GSTA2</i>	2	4	6	0	0	0	5	1
<i>LTA4H</i>	0	6	6	1	0	0	4	1
<i>MTHFR</i>	0	6	6	0	0	0	6	0
<i>NOS1</i>	0	6	6	0	0	0	5	1
<i>PAOX</i>	0	6	6	1	0	1	4	0
<i>PRKCE</i>	0	6	6	0	0	0	5	1
<i>PSG2</i>	1	5	6	0	1	0	4	1
<i>PSG6</i>	1	5	6	0	0	1	4	1
<i>PTGRI</i>	2	4	6	0	0	2	4	0

<i>PTGS1</i>	0	6	6	1	0	0	4	1
<i>SIGLEC6</i>	0	6	6	0	0	0	2	4
<i>SLC13A4</i>	0	6	6	0	0	0	6	0
<i>SLC6A12</i>	0	6	6	0	0	0	5	1
<i>SULF1</i>	0	6	6	0	0	0	5	1
<i>SULT1A1</i>	2	4	6	0	0	0	6	0
<i>SULT1E1</i>	2	4	6	0	0	0	6	0
<i>TFRC</i>	0	6	6	0	0	0	5	1
<i>TXNRD1</i>	1	5	6	0	0	0	5	1
<i>UGT1A10</i>	1	5	6	0	0	0	3	3
<i>ADH6</i>	0	5	5	1	0	0	4	0
<i>ARSA</i>	1	4	5	0	0	1	2	2
<i>ATPIA3</i>	0	5	5	1	0	0	4	0
<i>CHST1</i>	1	4	5	0	0	0	5	0
<i>CHST12</i>	0	5	5	0	0	0	5	0
<i>CHST14</i>	0	5	5	0	0	0	5	0
<i>CYP11B1</i>	0	5	5	0	0	0	5	0
<i>CYP26A1</i>	0	5	5	0	0	0	4	1
<i>DNMT3B</i>	0	5	5	0	1	0	4	0
<i>FLT1</i>	0	5	5	0	0	0	5	0
<i>FOLR1</i>	2	3	5	0	0	0	4	1
<i>GAL3ST2</i>	1	4	5	0	0	0	2	3
<i>GPX5</i>	0	5	5	0	0	1	3	1
<i>GPX7</i>	0	5	5	0	0	0	5	0
<i>GUSB</i>	0	5	5	0	0	1	4	0
<i>IFT81</i>	1	4	5	0	0	1	3	1
<i>JAG1</i>	0	5	5	0	0	0	5	0
<i>JAM3</i>	1	4	5	0	0	0	5	0
<i>NTRK2</i>	0	5	5	0	0	0	4	1
<i>PON1</i>	0	5	5	0	0	0	5	0
<i>PSG11</i>	1	4	5	0	0	0	3	2
<i>PTGR2</i>	0	5	5	0	0	0	4	1
<i>SLC1A2</i>	1	4	5	0	0	0	4	1
<i>SLC1A4</i>	1	4	5	1	0	0	3	1
<i>SLC25A20</i>	1	4	5	0	0	0	3	2
<i>SLC3A2</i>	0	5	5	0	0	1	1	3
<i>SMOX</i>	1	4	5	0	1	0	4	0
<i>SULT1B1</i>	3	2	5	0	0	0	4	1
<i>VCL</i>	0	5	5	0	0	0	5	0
<i>ABCC5</i>	0	4	4	1	0	0	3	0
<i>AKRID1</i>	2	2	4	0	0	0	4	0
<i>ALDH1A2</i>	0	4	4	0	0	0	3	1
<i>CHST10</i>	0	4	4	1	0	0	2	1
<i>CHST11</i>	1	3	4	0	0	0	3	1
<i>CHST13</i>	1	3	4	0	1	0	3	0
<i>CRYZ</i>	3	1	4	0	0	0	2	2
<i>CYP20A1</i>	1	3	4	0	0	0	2	2
<i>CYP27B1</i>	1	3	4	0	0	0	4	0
<i>CYP2S1</i>	1	3	4	1	0	0	2	1
<i>CYP2U1</i>	2	2	4	0	0	0	4	0
<i>CYP2W1</i>	0	4	4	0	0	0	2	2
<i>DNAAF2</i>	1	3	4	0	0	0	4	0
<i>GALNS</i>	0	4	4	0	0	0	4	0
<i>GH1</i>	0	4	4	0	0	0	4	0

<i>GPR32</i>	0	4	4	0	0	0	4	0
<i>GPX4</i>	0	4	4	0	0	0	3	1
<i>HGF</i>	2	2	4	0	0	0	3	1
<i>MFSD2A</i>	0	4	4	0	0	0	3	1
<i>MTTP</i>	0	4	4	1	0	0	3	0
<i>NAT1</i>	1	3	4	0	0	1	0	3
<i>NFKB1</i>	0	4	4	0	0	0	4	0
<i>PRG2</i>	0	4	4	0	0	0	3	1
<i>PROK1</i>	1	3	4	0	0	1	3	0
<i>PSG1</i>	3	1	4	1	0	0	3	0
<i>SERPINE2</i>	1	3	4	0	0	0	3	1
<i>SLC2A14</i>	0	4	4	0	1	0	3	0
<i>SLC35D2</i>	0	4	4	0	0	0	4	0
<i>SULT2B1</i>	0	4	4	0	1	1	1	1
<i>TPST1</i>	0	4	4	0	0	0	3	1
<i>UGT2B28</i>	1	3	4	0	0	0	2	2
<i>ACTN1</i>	0	3	3	0	0	0	2	1
<i>ADA</i>	0	3	3	0	0	0	2	1
<i>ADRB2</i>	1	2	3	1	0	0	2	0
<i>AKR1B1</i>	0	3	3	0	0	0	3	0
<i>AKR1B15</i>	0	3	3	0	0	1	2	0
<i>ALDH1A3</i>	0	3	3	0	0	0	2	1
<i>ALDH3A2</i>	0	3	3	0	0	0	1	2
<i>AQP9</i>	0	3	3	0	0	0	2	1
<i>ATP1B2</i>	1	2	3	2	0	0	1	0
<i>ATP1B3</i>	0	3	3	0	0	0	3	0
<i>B3GAT2</i>	0	3	3	0	0	0	2	1
<i>CHST8</i>	0	3	3	0	0	0	3	0
<i>CLDN3</i>	0	3	3	0	0	0	1	2
<i>CYP1B1</i>	0	3	3	0	0	0	3	0
<i>CYP4Z1</i>	0	3	3	0	0	0	3	0
<i>CYP51A1</i>	0	3	3	0	0	0	3	0
<i>DCXR</i>	0	3	3	0	0	0	3	0
<i>EPHX3</i>	0	3	3	0	0	0	3	0
<i>FAM26D</i>	0	3	3	0	0	1	2	0
<i>GABBR2</i>	0	3	3	0	0	0	3	0
<i>GABRA5</i>	0	3	3	0	0	0	2	1
<i>GAL3ST1</i>	0	3	3	0	0	0	2	1
<i>GNS</i>	0	3	3	0	0	0	3	0
<i>GSTM5</i>	2	1	3	0	0	0	3	0
<i>GSTP1</i>	0	3	3	0	0	0	3	0
<i>HSD11B1</i>	0	3	3	0	0	0	3	0
<i>INSL4</i>	2	1	3	0	0	0	3	0
<i>IYD</i>	0	3	3	0	0	0	3	0
<i>NNMT</i>	0	3	3	0	0	0	3	0
<i>PAR6B</i>	0	3	3	1	0	0	2	0
<i>PRKCZ</i>	0	3	3	0	0	0	3	0
<i>PTGES2</i>	0	3	3	0	0	0	3	0
<i>SLC16A1</i>	0	3	3	0	0	0	3	0
<i>SLC19A1</i>	0	3	3	1	0	1	1	0
<i>SLC2A3</i>	0	3	3	0	0	1	2	0
<i>SLC6A4</i>	0	3	3	0	0	0	3	0
<i>SULT1C4</i>	0	3	3	0	0	0	2	1
<i>TP53I3</i>	1	2	3	1	0	0	1	1

<i>TPST2</i>	0	3	3	0	0	0	3	0
<i>UGT1A8</i>	1	2	3	0	0	0	2	1
<i>AOC3</i>	2	0	2	0	0	0	2	0
<i>AQP4</i>	0	2	2	0	0	0	2	0
<i>ATPIA2</i>	0	2	2	0	0	0	1	1
<i>B3GAT3</i>	0	2	2	0	0	0	2	0
<i>CHST3</i>	0	2	2	0	0	0	2	0
<i>CLDN1</i>	0	2	2	0	0	1	1	0
<i>CYP46A1</i>	0	2	2	0	0	0	2	0
<i>EBI3</i>	0	2	2	0	0	0	2	0
<i>EPHX4</i>	0	2	2	0	0	0	1	1
<i>FCGR2B</i>	0	2	2	0	0	0	2	0
<i>GABRA1</i>	0	2	2	0	0	0	2	0
<i>GNGT1</i>	0	2	2	0	0	0	2	0
<i>GSTA4</i>	0	2	2	0	0	0	2	0
<i>GSTM2</i>	1	1	2	0	0	0	1	1
<i>GSTZ1</i>	1	1	2	0	0	0	0	2
<i>HNMT</i>	0	2	2	0	0	0	2	0
<i>LGALS14</i>	1	1	2	1	0	0	1	0
<i>MGST1</i>	1	1	2	0	0	0	2	0
<i>MGST2</i>	1	1	2	0	0	1	1	0
<i>MGST3</i>	0	2	2	0	0	0	2	0
<i>MTF1</i>	0	2	2	0	0	0	2	0
<i>OCLN</i>	0	2	2	0	0	1	0	1
<i>PAR6G</i>	0	2	2	0	0	0	1	1
<i>PTGES</i>	0	2	2	0	0	0	2	0
<i>QDPR</i>	0	2	2	0	0	0	2	0
<i>RAB3B</i>	0	2	2	0	0	0	2	0
<i>SLC39A8</i>	0	2	2	0	0	1	1	0
<i>SLC7A1</i>	0	2	2	0	0	0	2	0
<i>TFPI2</i>	0	2	2	0	0	0	2	0
<i>WNT3A</i>	0	2	2	0	0	0	2	0
<i>ADH1C</i>	0	1	1	0	0	0	1	0
<i>ATPIA1</i>	0	1	1	0	0	0	1	0
<i>ATP1B1</i>	0	1	1	0	0	0	1	0
<i>B3GAT1</i>	1	0	1	0	0	0	1	0
<i>CGA</i>	0	1	1	0	0	0	1	0
<i>CGB1</i>	0	1	1	0	0	0	1	0
<i>CLDN12</i>	0	1	1	0	0	0	1	0
<i>CRH</i>	0	1	1	0	0	0	1	0
<i>CSH1</i>	0	1	1	0	0	0	1	0
<i>CSH2</i>	0	1	1	0	0	1	0	0
<i>CYP26C1</i>	0	1	1	0	0	0	1	0
<i>DHRS3</i>	0	1	1	0	1	0	0	0
<i>DHRS4</i>	0	1	1	0	0	0	0	1
<i>GABRB3</i>	0	1	1	0	0	0	1	0
<i>GSTK1</i>	0	1	1	0	0	0	1	0
<i>GSTM3</i>	0	1	1	0	0	0	1	0
<i>JAM2</i>	0	1	1	0	0	0	0	1
<i>KRTAP26-1</i>	0	1	1	0	0	0	0	1
<i>LOX</i>	0	1	1	0	0	0	1	0
<i>MEST</i>	0	1	1	0	0	0	1	0
<i>MT1A</i>	0	1	1	0	0	0	1	0
<i>MT1B</i>	0	1	1	0	0	1	0	0

<i>NOS1AP</i>	0	1	1	0	0	0	1	0
<i>PTGS2</i>	0	1	1	0	0	0	1	0
<i>RSPH9</i>	0	1	1	0	0	0	1	0
<i>SLC25A12</i>	0	1	1	0	0	0	1	0
<i>SLC25A24</i>	1	0	1	0	0	0	1	0
<i>SLC29A1</i>	0	1	1	0	0	0	1	0
<i>SLC7A5</i>	0	1	1	0	0	0	1	0
<i>TAC3</i>	0	1	1	0	0	1	0	0
<i>VEGFA</i>	0	1	1	0	0	0	1	0

Cytochrome P450 enzymes	
CYP1A1	CYP1A2
CYP1B1	CYP2A6
CYP2A7	CYP2A13
CYP2B6	CYP2C8
CYP2C9	CYP2C18
CYP2C19	CYP2D6
CYP2E1	CYP2F1
CYP2J2	CYP2R1
CYP2S1	CYP2U1
CYP2W1	CYP3A4
CYP3A5	CYP3A7
CYP3A43	CYP4A11
CYP4A22	CYP4B1
CYP4F2	CYP4F3
CYP4F8	CYP4F11
CYP4F12	CYP4F22
CYP4V2	CYP4X1
CYP4Z1	TBXAS1
CYP7A1	CYP7B1
PTGS	CYP8B1
CYP11A1	CYP11B1
CYP11B2	CYP17A1
CYP19A1	CYP20A1
CYP21A2	CYP24A1
CYP26A1	CYP26B1
CYP26C1	CYP27A1
CYP27B1	CYP27C1
CYP39A1	CYP46A1
CYP51A1	
FMOs	Amine Oxidases
FMO1	LOX
FMO2	LOXL1
FMO3	LOXL2
FMO4	LOXL3
FMO5	LOXL4
	AOC1
	AOC2
	AOC3
	MAOA
	MAOB
	PAOX
	SMOX
	AOX1
	XDH
	SUOX
PTGs	Lipoxygenases
PTGS1	ALOX5
PTGS2	ALOX12
PTGES	ALOX15
PTGES2	ALOX15B
PTGES3	ALOXE3
	ALOX12B
UDP-glucuronosyltransferases	
UGT1A1	UGT1A3
UGT1A4	UGT1A5
UGT1A6	UGT1A7
UGT1A8	UGT1A9
UGT1A10	UGT2A1
UGT2A2	UGT2A3
UGT2A3	UGT2B4
UGT2B4	UGT2B7
UGT2B10	UGT2B11
UGT2B11	UGT2B15
UGT2B17	UGT2B28
UGT2B28	B3GAT1
B3GAT1	B3GAT2
B3GAT2	B3GAT3
NATs	Glutathione S-transferases
NAT1	GSTA1
NAT2	GSTA2
	GSTA3
	GSTA4
	GSTA5
	GSTK1
	GSTM1
	GSTM2
	GSTM3
	GSTM4
	GSTM5
	GSTO1
	GSTO2
	GSTP1
	GSTT1
	GSTT2
	GSTT2B
	GSTZ1
	MGST1
	MGST2
	MGST3
Sulfotransferases	
CHST1	CHST2
CHST3	CHST4
CHST5	CHST6
CHST7	CHST8
CHST9	CHST10
CHST11	CHST12
CHST13	CHST14
GAL3ST1	GAL3ST2
GAL3ST3	GAL3ST4
TPST1	TPST2
SULT1A1	SULT1A2
SULT1A3	SULT1A4
SULT1B1	SULT1C2
SULT1C3	SULT1C4
SULT1E1	SULT2A1
SULT2B1	SULT4A1
EHs	Sulfatases
EPHX1	ARSA
EPHX2	ARSB
EPHX3	ARSD
EPHX4	ARSE
LT4H	ARSF
MEST	ARSG
	ARSH
	ARSI
	ARSJ
	ARSK
	GALNS
	GNS
	IDS
	SGSH
	STS
	SULF1
	SULF2
NAPH-dependent carbonyl and quinone reductases	NRs
NQO1	IQYD
NQO2	QDPR
CBR1	
CBR3	
DHRS1	
DHRS2	
DHRS3	
DHRS4	
HSD11B1	
HSD11B2	
DCXR	
AKR1A1	
AKR1B1	
AKR1B10	
AKR1B15	
AKR1C1	
AKR1C2	
AKR1C3	
AKR1C4	
AKR1D1	
AKR7A2	
AKR7A3	
TXNRD1	
TXNRD2	
TXNRD3	
CRYZ	
TP53B	
PTGR1	
PTGR2	
Methyltransferases	GPXs
COMT	GPX1
POMT1	GPX2
POMT2	GPX3
NNMT	GPX4
GNNMT	GPX5
HNNMT	GPX6
INNMT	GPX7
PNNMT	
TPMT	
DNMT1	
DNMT3B	
Esterases	Dehydrogenases
CES1	ADH1A
CES2	ADH1B
CES3	ADH1C
ACHE	ADH4
BCHE	ADH5
PON1	ADH6
PON2	ALDH1A1
PON3	ALDH1A2
	ALDH1A3
	ALDH1B1
	ALDH1L1
	ALDH2
	ALDH3A1
	ALDH3A2
	ALDH3B1
	ALDH3B2
	ALDH4A1
	ALDH5A1
	ALDH6A1
	ALDH7A1
	ALDH8A1
	ALDH9A1
	ALDH18A1
NOSs	
NOS1	
NOS2	
NOS3	
NOS1AP	
GABA receptors	Folate metabolism enzymes
GABBR1	FOLR1
GABBR2	FOLR2
GABRA1	DHFR
GABRA5	MTHFR
GABRB3	CBS
GABRG2	TYMS
GABRG3	CTH
	MTR
	MTRR
Other enzymes/receptors	
MTF1	
MTF2	
MT1A	
MT2A	
GUSB	
ALAD	
CP	
TF	
AHR	

Figure S1a: Genes involved in detoxification processes identified through systematic literature review. Gene symbols are in accordance with HUGO Gene Nomenclature Committee (HGCN).

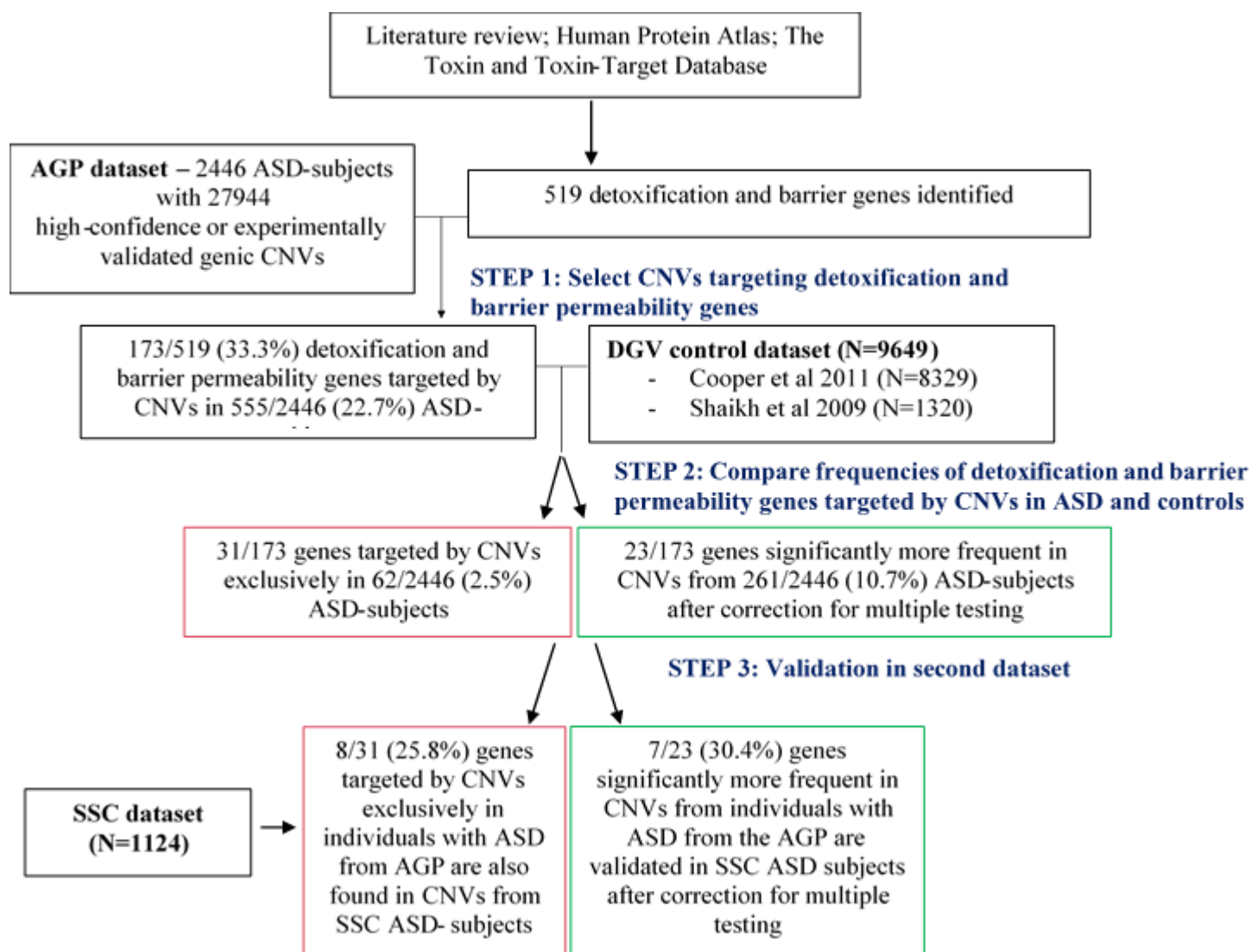


Figure S2: Main results regarding the numbers and percentages of detoxification and barrier genes targeted by CNVs in individuals from the AGP dataset and the validation results using the SSC dataset.