

INTERNATIONAL DOCTORAL
SCHOOL OF THE USC

Sara
Domínguez Alonso

PhD Thesis

Deciphering the non-coding genome in Autism Spectrum Disorders (ASD): in silico characterization of common, splicing and regulatory variants

Santiago de Compostela, 2024

Doctoral Programme in Molecular Medicine

INTERNATIONAL DOCTORAL
SCHOOL OF THE USC

DOCTORAL THESIS

DECIPHERING THE NON-CODING GENOME IN AUTISM SPECTRUM DISORDERS (ASD): *IN SILICO* CHARACTERIZATION OF COMMON, SPLICING AND REGULATORY VARIANTS

Sara Domínguez Alonso

Directors: Ángel Carracedo Álvarez, María Cristina Rodríguez Fontenla

Tutor: Ángel Carracedo Álvarez

PHD PROGRAMME IN MOLECULAR MEDICINE

SANTIAGO DE COMPOSTELA

2024

DECLARACIÓN CONFLICTOS DE INTERÉS

Yo, Sara Domínguez Alonso, con DNI 53198091B, declaro no tener ningún conflicto de interés en relación con la tesis de doctorado.

Por la presente, la doctoranda declara la autoría completa de todas las figuras mostradas en la tesis. Aquellas figuras que se han derivado de otras fuentes están debidamente atribuidas en el pie de la figura, y se cuenta con los permisos de reproducción correspondientes para su adaptación.

En Santiago de Compostela,
de marzo de 2024

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

DECLARACIÓN DE AUTORÍA DE LOS DATOS PRESENTADOS

La doctoranda cuenta con los permisos correspondientes de la editorial para la reproducción parcial de los contenidos presentes en los siguientes artículos, de los cuales Sara Domínguez Alonso es coautora, y que derivan del trabajo realizado en su tesis doctoral.

Dominguez-Alonso, S., Carracedo, A., & Rodriguez-Fontenla, C. The non-coding genome in Autism Spectrum Disorders. European journal of medical genetics. 2023;66(6), 104752. https://doi.org/10.1016/j.ejmg.2023.104752



The non-coding genome in Autism Spectrum Disorders

Author: S. Dominguez-Alonso,A. Carracedo,C. Rodriguez-Fontenla

Publication: European Journal of Medical Genetics

Publisher: Elsevier

Date: June 2023

© 2023 The Authors. Published by Elsevier Masson SAS.

Journal Author Rights

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights

Dominguez-Alonso, S., Carracedo, A. & Rodriguez-Fontenla, C. eQTL colocalization analysis highlights novel susceptibility genes in Autism Spectrum Disorders (ASD). Transl Psychiatry. 2023;13, 336. https://doi.org/10.1038/s41398-023-02621-0

eQTL colocalization analysis highlights novel susceptibility genes in Autism Spectrum Disorders (ASD)

Author: S. Dominguez-Alonso et al

Publication: Translational Psychiatry

Publisher: Springer Nature

Date: Oct 31, 2023

Copyright © 2023, The Author(s)

Creative Commons

This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

To request permission for a type of use not listed, please contact Springer Nature

En Santiago de Compostela,
de marzo de 2024

Fdo. Sara Domínguez Alonso

DECLARACIÓN SOBRE LA APROBACIÓN POR EL COMITÉ DE ÉTICA DE LA INVESTIGACIÓN DE LA QUE DERIVA LA PRESENTE TESIS

Por la presente, la doctoranda declara la resolución favorable por parte del comité de ética para los estudios identificados con los códigos 2012/098 y 2020/400 de los cuales se derivan los resultados expuestos en los capítulos 2 y 3. Para que así conste, se adjuntan dichos informes en el anexo “10.4 Favorable reports from the ethics committee”.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

AGRADECIMIENTOS

Inevitablemente, las primeras palabras son para vosotros. Gracias Cris. Gracias Ángel. Gracias por darme una de las oportunidades más grandes de mi vida, hace ya casi cuatro años. Gracias por el cariño y por cada una de las veces en las que me habéis tendido la mano. Gracias por hacer que esto hoy sea posible, y gracias por todo lo que vendrá después.

Gracias al grupo de Medicina Xenómica. Ojalá algún día pueda volver a tener un equipo de trabajo como este. Y, sobre todo, gracias Silvia. Por todos los agobios y todos los desahogos en la puerta del CIMUS. Por esta y por todas las etapas que hemos vivido juntas.

Gracias a todos, y no somos pocos, los integrantes de esa mesa eterna de la Resting Room. Gracias porque cualquier agobio se me quitaba a la una, sobre todo en las comidas de dos horas. Gracias porque sois el clarísimo ejemplo de que no importa el lugar, si no la compañía. Ahora que no voy a poder veros todos los días, me doy cuenta de que la lluvia no era tan mala, ni hacía tanto frío. Estos últimos años he tenido al sol sobrevalorado; a vosotros sí que os voy a echar de menos.

Gracias a mi familia. Sólo espero que esta nueva etapa me permita, de una forma u otra, devolveros sólo un poquito de lo que me habéis dado a mí. Gracias por haber estado ahí siempre, aunque no fuese juntos. Gracias por el mejor de los cariños que podría pedir. No sé qué haría sin vosotros.

Gracias Vigo, Valencia y Santiago, mi otra familia. Por entenderme siempre y por hacerme reír. Por todas las videollamadas y todos los reencuentros. Me habéis salvado de muchísimo más de lo que os podéis imaginar. Tengo a los mejores amigos del mundo.

Gracias mamá, gracias papá. No sólo por aguantarme en esta y en cada una de mis montañas rusas, que no es poco, ni han sido pocas. Tampoco por apoyarme en mis decisiones, por encima, en todo momento, de cualquier consecuencia. Ni siquiera por ayudarme siempre a que todo sea más fácil. Gracias, sobre todo, por hacerme ser lo que soy hoy; sin vuestra cara de orgullo cada vez que he dado un paso, por muy pequeño que fuese, no habría dado ni la mitad.

Y, para acabar, gracias Javi. Has sido mis ganas más grandes de dejarlo todo e irme, y aun así, me diste fuerzas para quedarme. Nunca va a escribirse un agradecimiento más sincero y en deuda que este; gracias por ser el mejor equipo que existe.

No he podido tener más suerte.
Gracias a todos.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ABBREVIATIONS

AD	Alzheimer's disease
ADHD	Attention-deficit/hyperactivity disorder
AG	Acceptor gain
AL	Acceptor loss
APA	Alternative cleavage and polyadenylation
AS	Alternative splicing
ASD	Autism spectrum disorders
ATAC-Seq	Assay for transposase-accessible chromatin with sequencing
BD	Bipolar disorders
BED	Browser extensible data
bp	Base pairs
BP	Biological process
CC	Cellular component
cCRE	Candidate CRE
CGH	Comparative genomic hybridization
ChIP-Seq	Chromatin immunoprecipitation with sequencing
CNV	Copy number variant
CRE	Cis-regulatory element
CWAS	Category-wide association study
dbGaP	Database of genotypes and phenotypes
dELS	Distal enhancer-like signature
DG	Donor gain
DHS	DNase hypersensitivity site
DL	Donor loss
DNase-Seq	DNase I hypersensitive sites sequencing
DNV	*De novo* variation
DSM	Diagnostic and statistical manual of mental disorders
DZ	Dizygotic
ELS	Enhancer-like motif
Encode	Encyclopedia of DNA elements
eQTL	Expression quantitative trait loci
eRNA	Enhancer RNA
FDR	False discovery rate
FXS	Fragile X syndrome
GBA	Gene-based association analysis
GERP	Genomic evolutionary rate profiling
GI	Gastrointestinal
gnomAD	Genome aggregation database
GO	Gene ontology
GQ	Genotype quality

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRCh38/37	Genome reference consortium human build 38/37
GSEA	Gene set enrichment analysis
GTEx	Genotype-tissue expression project
gVCF	Genomic variant call format
GWAS	Genome-wide association study
ID	Intellectual disability
Indel	Insertion-deletion
iPSYCH	Lundbeck foundation initiative for integrative psychiatric research
KdVS	Koolen-de Vries syndrome
kb	Kilobase
LD	Linkage disequilibrium
LGD	Likely gene disrupting
lncRNA	Long non-coding RNA
LoF	Loss-of-function
LOI	Locus of interest
MAF	Minor allele frequency
Mb	Megabase
MD	Major depression
MF	Molecular function
miRNA	MicroRNA
MPC	Missense badness, PolyPhen-2, and constraint
mRNA	Messenger RNA
MZ	Monozygotic
ncRNA	Non-coding RNA
NDD	Neurodevelopmental disorder
NGS	Next-generation sequencing
OCD	Obsessive-compulsive disorder
OR	Odds ratio
PASCAL	Pathway scoring algorithm
pELS	Proximal enhancer-like signature
PGC	Psychiatric genomics consortium
PLAC-Seq	Proximity ligation-assisted ChIP-Seq
PLS	Promoter-like signature
PPI	Protein-protein interaction
PTVs	Protein-truncating variant
QC	Quality control
QTL	Quantitative trait locus
RBP	RNA-binding protein
RNA-Seq	RNA sequencing
RR	Recurrence risk
RVIS	Residual variation intolerance score
SNP	Single-nucleotide polymorphism

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

SNV	Single-nucleotide variant
STR	Short tandem repeat
SV	Structural variation
SZP	Schizophrenia
TAD	Topologically associated domain
TADA	Transmission and *de novo* association test
TF	Transcription factors
Ti/Tv ratio	Transitions to transversions ratio
TPM	Transcripts per million
TR	Tandem repeat sequence
TSS	Transcription start site
TWAS	Transcriptome wide association studies
UTR	Untranslated region
VOUS	Variant of uncertain significance
VQSLOD	Variant quality score log odds
VQSR	Variant quality score recalibration
WES	Whole-exome sequencing analysis
WGS	Whole genome sequencing

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

TABLE OF CONTENTS

1 EXTENDED SUMMARY IN GALICIAN 2
2 ABSTRACT 11
3 INTRODUCTION 13
3.1 Autism spectrum disorders (ASD) 13
3.1.1 Prevalence 15
3.1.2 Comorbidities 16
3.1.3 Genetic vs. non-genetic factors 17
3.1.3.1 Genetic factors: Early insights from twin and family studies 18
3.1.3.2 Non-genetic factors: Environmental factors and epigenetic modifications 19
3.2 Genetic basis of ASD: liability and genetic studies 20
3.2.1 Linkage and candidate gene studies 24
3.2.2 Whole exome sequencing analysis (WES) 25
3.2.2.1 *De novo* variants (DNVs) 25
3.2.2.2 Inherited variants 25
3.2.2.3 Gene discovery in WES 26
3.2.3 Copy number variant (CNV) analysis 30
3.2.4 Genome-Wide Association Studies (GWAS) 31
3.2.4.1 Gene discovery in GWAS 33
3.2.4.1.1 Gene-based association analysis (GBA) 34
3.2.4.1.2 Transcriptome wide analysis (TWAS) 34
3.2.5 The “missing heritability” in ASD 35
3.3 The non-coding genome in ASD 36
3.3.1 Biased studies 38
3.3.1.1 Genetic variants in 5' and 3' untranslated regions (UTRs) 38
3.3.1.2 Promoters 39
3.3.1.3 Cis-regulatory elements (CREs) 40
3.3.1.4 Non-coding RNAs (ncRNAs) 41
3.3.1.4.1 Long non-coding RNAs (lncRNAs) 42
3.3.1.4.2 microRNAs (miRNAs) 43
3.3.1.5 Repetitive DNA 44
3.3.1.5.1 Tandem repeats (TRs) 44
3.3.1.5.2 Retrotransposons 45
3.3.1.6 Splicing variants 46
3.3.1.7 Gene discovery in the non-coding genome 47

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3.3.2 Unbiased: WGS (whole genome sequencing) analysis 49
3.3.2.1 WGS limitations 49
3.3.2.2 Characterization of the non-coding genome in WGS 50
3.3.2.3 Key findings in WGS 51
3.4 Genes and molecular pathways 54
3.4.1 Aberrant mRNA translation of synaptic proteins 54
3.4.2 Synaptic function 54
3.4.3 Chromatin remodeling 55
3.4.4 Concluding remarks 56
4 JUSTIFICATION AND OBJECTIVES 59
5 CHAPTER 1 62
5.1 Objective 64
5.2 Materials and methods 65
5.2.1 Samples and collection 65
5.2.1.1 Discovery set 65
5.2.1.2 Neurodevelopmental Disorders (NDDs) sets 66
5.2.1.3 Data harmonization 66
5.2.2 eQTL (expression quantitative trait locus) analyses 67
5.2.2.1 Selecting ASD associated genes from GWAS, GBA and TWAS 67
5.2.3 eQTpLot: visualization of colocalization between eQTL and GWAS data 68
5.2.3.1 Input data and significance thresholds 69
5.2.4 eQTL-GWAS colocalization plots 69
5.2.4.1 Multi and pan-tissue analysis 69
5.2.4.2 Congruous and incongruous variants 70
5.2.4.3 eQTL enrichment plots 70
5.2.4.4 p-p correlation plots 70
5.2.5 Gene expression, gene network and related human disease-associated genes analysis 71
5.2.5.1 Gene expression heatmap 71
5.2.5.2 Gene network analysis 71
5.3 Results 72
5.3.1 Genes with ASD colocalization in the multi-tissue analysis 73
5.3.2 Genes with ASD colocalization in the pan-tissue analysis 75
5.3.2.1 *CRHR1* (Corticotropin Releasing Hormone Receptor 1) 76
5.3.2.2 *KANSL1* (KAT8 Regulatory NSL Complex Subunit 1) 76
5.3.2.1 *MANBA* (Mannosidase Beta A) 76

5.3.2.2 *MMP12* (Matrix Metallopeptidase 12) 76
5.3.2.3 *WNT3* (Wnt Family Member 3) 76
5.3.3 Gene expression, gene networks and related human disease-associated genes 77
5.4 Discussion 78
5.4.1 Colocalization in brain tissues 78
5.4.2 Colocalization in the pan-tissue analysis 81
5.4.2.1 *CRHR1, KANSL1* and *WNT3* in adjacent loci 81
5.4.2.2 Additional implications 82
5.4.3 Implications of the eQTL colocalization analysis 83
6 CHAPTER 2 85
6.1 Objective 85
6.2 Methods 86
6.2.1 Subjects 86
6.2.2 *In silico* prediction and prioritization of splicing variants 86
6.2.3 *In silico* variant validation 88
6.2.3.1 SpliceVault 88
6.2.3.2 Tissue specificity: ABSplice 88
6.2.4 Complementary datasets 89
6.2.5 Gene ontology analysis 90
6.2.5.1 Gene network analysis 90
6.2.5.2 Enrichment analysis 90
6.2.5.2.1 Gene ontology (GO) enrichment analysis 90
6.2.5.2.2 Cluster comparer 91
6.2.6 Gene expression analysis 91
6.3 Results 91
6.3.1 *In silico* variant prediction and validation 91
6.3.2 Tissue specificity of predicted splice variants 97
6.3.3 Cluster enrichment 99
6.4 Discussion 103
6.4.1 Tissue-specificity in alternative splicing (AS) 104
6.4.2 Biological underpinnings of AS 104
6.4.3 Future perspectives 106
7 CHAPTER 3 108
7.1 Objective 108
7.2 Methods 109
7.2.1 Targeted sequencing and data processing 109

7.2.1.1 Cohort description 109
7.2.1.2 Sample selection 109
7.2.1.3 Selection of the regulatory regions of interest 109
7.2.1.4 Targeted sequencing 110
7.2.1.4.1 Joint genotyping 111
7.2.1.4.2 Variant quality score recalibration (VQSR) 111
7.2.2 Analysis 112
7.2.2.1 Dataset Quality Control 114
7.2.2.2 Variant-level quality control 114
7.2.2.3 Genotype-level quality control 114
7.2.3 *De novo* variant detection 115
7.2.4 Rare inherited variation 116
7.2.5 Regulatory impact prediction 116
7.2.6 Candidate gene elucidation 119
7.2.7 Variant prioritization 120
7.2.8 Protein-protein interaction (PPI) network 121
7.2.9 Cluster profiler 121
7.2.10 Transcription factors (TF) enrichment 121
7.2.11 Presence in topologically associated domains (TADs) 121
7.2.12 Statistical significance tests 122
7.3 Results 122
7.3.1 Dataset 122
7.3.1.1 *De novo* variants 122
7.3.1.2 Inherited variants 122
7.3.1.3 Variant enrichment in cCREs (candidate cis-Regulatory Elements) 123
7.3.2 Effect of genetic variants in gene regulation 124
7.3.2.1 *De novo* variants 124
7.3.2.2 Inherited variants 124
7.3.2.3 Differences in gene regulation for inherited and *de novo* variants 125
7.3.3 Candidate gene association 126
7.3.3.1 *De novo* 126
7.3.3.2 Inherited 128
7.3.4 Gene function 130
7.3.5 TF enrichment 131
7.3.6 Variant prioritization and diagnostic yield 131
7.4 Discussion 132

7.4.1 Exploring the sex bias in regulatory variation within ASD 132
7.4.2 Sequence class dysregulation 133
7.4.3 Biological, molecular, and cellular consequences of regulatory variants 135
7.4.4 Preliminary genotype-phenotype associations 135
7.4.5 Limitations and future perspectives 136
8 CONCLUDING REMARKS 139
9 BIBLIOGRAPHY 142
10 APPENDIX 168
10.1 Supplementary Figures 168
10.1.1 Supplementary Figure 1. eQTpLot for ASD and *NKX2-2* multi-tissue analysis. 168
10.1.2 Supplementary Figure 2. eQTpLot eQTpLot for ASD and *CRHR1* pan-tissue analysis. 169
10.1.3 Supplementary Figure 3. eQTpLot for ASD and *KANSL1* pan-tissue analysis. .. 169
10.1.4 Supplementary Figure 4. eQTpLot for ASD and *MANBA* pan-tissue analysis. 170
10.1.5 Supplementary Figure 5. eQTpLot for ASD and *MMP12* pan-tissue analysis. 170
10.1.6 Supplementary Figure 6. eQTpLot for ASD and *WNT3* pan-tissue analysis. 171
10.1.7 Supplementary Figure 7. Validated variants with the highest impact restricted to testis. 171
10.2 Supplementary tables 172
10.2.1 Supplementary Table 1. *MAPT, NKX2-2* and *PTPRE* functionally related genes in brain tissue 172
10.2.2 Supplementary Table 2. *De novo* variants with high confidence predicted splice altering scores (SpliceAI ≥ 0.8) and *in silico* validation results 173
10.2.3 Supplementary Table 3. *CACNA1I*, *CBLB*, *CLTB*, *DLGAP1*, *DVL3*, *KIAA0513*, *OFD1*, *PKD1*, *SLC13A3* and *SCN2A* functionally related genes in brain tissue (splicing gene list). 173
10.2.4 Supplementary Table 4. Final list of *de novo* high regulatory impact variants. .. 174
10.2.5 Supplementary Table 5. Enrichments in the number of sequence classes assigned to ultra-rare inherited variants in each of ENCODE's cCREs categories. 174
10.2.6 Supplementary Table 6. Final list of inherited high regulatory impact variants .. 174
10.2.7 Supplementary Table 7. Enrichment of TFs among high regulatory impact *de novo* variants 174
10.2.8 Supplementary Table 8. Enrichment of TFs among high regulatory impact inherited variants 175
10.2.9 Supplementary Table 9. ASD risk genes custom list 175
10.2.10 Supplementary Table 10. Inherited variants prioritized by the presence of the assigned PLARV gene into our custom list of ASD risk genes 181
10.3 VQSR in detail 181

10.3.1 Resources to build the recalibration model 181
10.3.2 Covariates employed for variant recalibration 182
10.4 Favorable reports from the ethics committee 184
11 LIST OF TABLES 188
12 LIST OF FIGURES 190
13 PUBLICATIONS AND CONTRIBUTIONS 192

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

EXTENDED SUMMARY IN GALICIAN

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1 EXTENDED SUMMARY IN GALICIAN

A definición do autismo sufriu unha transformación significativa dentro do Manual Diagnóstico e Estatístico de Trastornos Mentais (DSM), a guía recoñecida mundialmente polos profesionais sanitarios para diagnosticar os trastornos mentais. Esta evolución reflicte cambios de perspectivas e avances no coñecemento científico.

Caracterizado inicialmente por Leo Kanner en 1943 como unha forma de esquizofrenia infantil, o autismo foi recoñecido máis tarde como un conxunto de trastornos do desenvolvemento interconectados. Xa en 2013, o DSM-5 adoptou o concepto dun espectro continuo, introducindo o termo "Trastornos do Espectro do Autismo" (TEA), para abordar as anteriores inconsistencias diagnósticas. Os criterios revisados fixeron fincapé en dúas características fundamentais: o deterioro persistente da comunicación e da interacción social, e patróns de comportamento restrinxidos e repetitivos, ambos manifestados nas primeiras etapas da infancia.

A etioloxía dos TEA é multifactorial e implica factores xenéticos e non xenéticos (ambientais) (**Figura 2**). Estudos recentes a gran escala informaron de estimas de herdabilidade de ata o 82.7%, establecendo os TEA como unha das condicións definidas por comportamento máis herdables. Non obstante, descifrar a arquitectura xenética dos TEA supón un gran desafío debido á implicación de múltiples factores xenéticos (**Figura 3**).

Unha fracción importante da herdabilidade xenética dos TEA pode ser explicada pola variación común (frecuencia do alelo menor: MAF > 5%), contribuíndo nun 50% á etioloxía. Sen embargo, a variación xenética rara (MAF < 1%) confire un maior risco individual (**Figura 4**).

A secuenciación de exoma completo e os estudos de hibridación xenómica ou *arrays* de polimorfismos dun só nucleótido (SNPs) son o tipo de análise máis comunmente empregados en TEA. Grazas a estes estudos, demostrouse a implicación, na etioloxía dos TEA, de variacións *de novo* (DNVs) e variacións no número de copias (CNVs), salientando a asociación de centos de xenes que están implicados na etioloxía dos TEA e converxendo en rutas biolóxicas implicadas na regulación transcripcional, funcións sinápticas e remodelación da cromatina (**Figura 10**).

Doutra banda, os estudos de secuenciación de exoma completo cobren só a parte codificante do xenoma (1% de todo o xenoma). Por iso, nos últimos anos, comezáronse a estudar as variantes presentes nas rexións non codificantes grazas aos estudos de xenoma completo. Recentemente, a secuenciación dirixida de elementos reguladores, presentes nas rexións non codificantes, mostrou un enriquecemento de DNMs en elementos reguladores activos do cerebro fetal. Ademais, demostrouse que a variación estrutural herdada rara nos elementos reguladores predispón aos TEA.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En conxunto, estes resultados, entre outros, parecen apuntar a que as variantes xenéticas dentro das rexións reguladoras deben ser exploradas en maior profundidade. O principal problema é que o número de estudos de xenoma completo realizados en TEA está moi por debaixo dos realizados para exoma (**Táboa 5** *versus* **Táboa 8**), e as metodoloxías para anotar o seu impacto funcional apenas están comezando a desenvolverse (**Figura 9**).

Non obstante, as altas estimas de herdabilidade seguen sen poder confirmarse cos datos obtidos das análises xenéticas actuais, dando lugar ao que comunmente se denomina "fenda de herdabilidade" ou "herdabilidade perdida". A herdabilidade perdida nos TEA consiste naquela parte da herdabilidade que segue sen ser explicada, o que pode deberse a unha sobreestima da herdabilidade, pero tamén a variantes xenéticas aínda sen descubrir e/ou á interacción entre elas.

A hipótese xeral desta tese, por tanto, consiste en afirmar que o problema da herdabilidade perdida podería resolverse mediante un estudo xenético que permita desentrañar a arquitectura polixénica dos TEA a través da identificación de novas variantes non codificantes implicadas no trastorno, axudando a esclarecer o seu papel na patoxénese dos TEA e a súa relación con novos xenes de risco.

Así, os principais obxectivos son: (i) investigar como a variación común asociada cos TEA, predominantemente situada na porción non codificante do xenoma, inflúe na aparición da enfermidade, (ii) avaliar as posibles repercusións da variación de *splicing* alternativo utilizando datos do exoma nunha cohorte de TEA de 360 tríos mediante o uso de algoritmos de *deep learning*, e (iii) caracterizar o impacto da variación xenética (tanto *de novo* como herdada) no panorama regulador en TEA.

En canto ao primeiro obxectivo (**Capítulo 1**), a investigación céntrase en desentrañar as bases xenéticas dos TEA mediante un estudo combinado de datos de GWAS (estudos de asociación a nivel de xenoma completo) e a análise de eQTLs (*expression quantitative trait loci*), variantes que están asociadas coa expresión dun xen.

Mediante estudos de colocalización, isto é, a correlación entre os sinais obtidos nos GWAS e os datos de eQTL, analizouse como a expresión xénica en distintos tecidos pode ser influenciada polas variantes asociadas (**Figura 13**). Estes estudos realizáronse a nivel de cerebro, onde esperabamos atopar sinais de colocalización significativos, por ser un tecido de relevancia no desenvolvemento do trastorno. Por outro lado, debido a que, en gran parte, o efecto dos eQTL se pode xeralizar ao conxunto total de tecidos, tamén se realizaron estudos de colocalización a nivel global, para aqueles xenes con insuficientes datos de eQTL en cerebro.

Ademais, mediante a integración de resultados de GWAS de trastornos xeneticamente correlacionados, melloramos o tamaño da mostra para poder compensar, parcialmente, a baixa potencia estatística que mostraron ata o momento os estudos de GWAS en TEA (**Figura 11**). Isto foi posible dado que un gran número de *loci* son compartidos entre distintos trastornos do neurodesenvolvemento (TND): na maior meta-análise de datos cruzados de GWAS en TND, que incluíu máis de 725,000 casos e controis con diversos diagnósticos, atopáronse 109 loci xenéticos que afectan ao mesmo tempo a múltiples TND. Así, incluímos na análise o conxunto de TND dos que puidemos obter datos de acceso público, que comprende TDAH (trastornos do déficit de atención e hiperactividade), depresión e esquizofrenia (**Táboa 9**).

UDC

Para a análise da colocalización empregouse a ferramenta eQTpLot, que ofrece un enfoque integral para comprender os fundamentos xenéticos dos TEA.

O estudo utilizou estatísticos resumo do GWAS da cohorte de TEA máis grande analizada ata a data, que incluíu 18,381 casos e 27,969 controis (**Figura 12**). Os datos de eQTL foron obtidos da versión 8 de GTEx, que inclúe 15,201 mostras de secuenciación de ARN de 49 tecidos. Ademais, asignáronse identificadores únicos a cada variante mediante a comparación de coordenadas xenómicas con datos derivados do estudo de 1000 *Genomes,* un catálogo de variación común en humanos. Os recontos alélicos da poboación europea obtivéronse do mesmo proxecto e empregáronse para calcular os valores de desequilibrio de ligamento entre as distintas variantes.

O proceso de selección de xenes candidatos a incluír na análise consistiu nunha búsqueda bibliográfica de estudos previos sobre TEA nos que se estableceron asociacións significativas de xenes mediante diferentes metodoloxías. Así, a selección dos primeiros 12 xenes de interese baseouse na proximidade aos SNPs que acadaron o limiar de significación no GWAS (definida como valores de desequilibrio de ligamento ≥ 0.6 con respecto aos SNPs significativamente asociados). A maiores, incluíronse xenes adicionais asociados en estudos de integración de múltiples ómicas, sumando un total de 27 xenes de interese (**Táboa 10**).

A ferramenta de eQTpLot aplicouse cun limiar de significación para os datos de eQTL de p-valor ≤ 0.05, mentres que para os datos GWAS se fixou en p-valor $< 5 \times 10^{-8}$. Estes límites son os empregados habitualmente para detectar asociación en análises de asociación, coa expresión xénica e de GWAS, respectivamente.

Para unha mellor interpretación dos datos, xeráronse gráficos de colocalización, visualizando a correlación entre os sinais (mediante a estratexia de *best fit linear regression*) e o enriquecemento de sinais significativos de eQTL entre os sinais significativos do GWAS (isto é, que porcentaxe do total dos sinais obtidos no GWAS están actuando como eQTLs). Así, definiuse unha colocalización significativa como unha correlación positiva ($r \geq 0.7$) e significativa (p-valor ≤ 0.05).

Para obter información funcional, utilizáronse representacións de *heatmaps* e construción de redes de expresión xénica específicas de cerebro, que se utilizaron para comprender as relacións funcionais dos xenes que mostraban unha colocalización significativa no tecido cerebral. Seguidamente, realizáronse análises de enriquecemento en determinados procesos biolóxicos de Gene Ontology (GO) e enfermidades (DIsGeNET), que proporcionaron información adicional sobre a relevancia biolóxica dos xenes identificados.

Así, o estudo centrouse en analizar a colocalización dos sinais de eQTL e de GWAS en TEA, incluíndo 27 loci de interese, revelando colocalización significativas en oito xenes: *CRHR1*, *KANSL1*, *MANBA*, *MAPT*, *MMP12*, *NKX2-2*, *PTPRE* e *WNT3* (**Táboa 12**). Ademais, *SRPK2* mostrou unha correlación positiva sen acadar o límite de significación. A análise de tecidos cerebrais identificou fortes correlacións para *MAPT*, *NKX2-2* e *PTPRE*. Estas asociacións foron replicadas no estudo de TND relacionados para *MAPT* (**Figura 14**) e *PTPRE* (**Figura 15**). *NKX2-2*, aínda que demostrou o maior sinal de correlación no tecido cerebral, careceu de replicación noutros TND (**Figura suplementaria 1**).

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Todos os xenes, excepto *NKX2-2*, mostraron colocalizacións significativas nos estudos de seguimento de TND correlacionados. En particular, *CRHR1*, *KANSL1*, *MAPT*, *PTPRE* e *WNT3* mostraron correlacións significativas cando se restrinxiron as análises ao tecido cerebral.

A análise de expresión utilizando os datos RNA-Seq de BrainSpan revelou unha sobreexpresión de *MAPT* en todas as etapas do desenvolvemento, e *KANSL1* mostrou unha expresión restrinxida durante os períodos relevantes de risco de TEA (etapas de desenvolvemento temperás) (**Figura 16 A, B**).

A análise de enriquecemento funcional indicou que *MAPT*, *NKX2-2* e *PTPRE*, xunto cos seus interactores no tecido cerebral, estaban asociados a procesos relacionados coa neuroxénese, a diferenciación neuronal e o desenvolvemento do sistema nervioso. Estes xenes tamén mostraron asociacións con enfermidades humanas relacionadas cos TEA, como son os trastornos do carácter e a intelixencia (**Figura 16, C**), e enriquecementos en procesos biolóxicos relacionados coa neuroxénese, a diferenciación neuronal e o correcto funcionamiento do sistema nervioso (**Táboa 13**).

Un dos achados máis relevantes na análise é a asociación de *MAPT* en tecido cerebral. Este xen, anteriormente asociado a outros TNDs (asociación replicada na nosa análise a través dos estudos de seguimento), non conta coa suficiente evidencia, ata a data, en canto á asociación cos TEA. Mediante o estudo realizado proporcionamos unha maior evidencia desta asociación, e esperamos que futuros estudos funcionais poidan esclarecer a súa función de xeito robusto.

Por outra banda, *CRHR1* demostrou ter unha alta correlación en tecido cerebral e está asociado con trastornos de depresión, cuxa relación foi tamén confirmada no estudo de seguimento para a cohorte de depresión analizada.

Ámbolos xenes presentan claras sinais de colocalización e, ademais, o feito de que ambos xenes estean implicados nunha mesma rexión xenómica parece indicar que esta acadará significación en posteriores estudos de xenoma completo nos TEA, asociándoos de xeito definitivo ao risco de desenvolver o trastorno.

Un dos principais retos dos GWAS é determinar con precisión os xenes de susceptibilidade á enfermidade, sobre todo tendo en conta que a maioría das asociacións atópanse en rexións non codificantes do xenoma. Así, o estudo demostra que a colocalización de eQTL pode axudar a superar este reto conectando os loci asociados á enfermidade cos mecanismos subxacentes, e proporciona unha visión completa da colocalización dos sinais eQTL e GWAS nos TEA, destacando xenes específicos e a súa potencial relevancia funcional no neurodesenvolvemento e o risco de enfermidade.

Con todo, recoñécense tres limitacións principais na exploración da variación xenética común nos TEA: (i) limitacións no poder estatístico debido a menores colección de pacientes en relación a outros TNDs, (ii) a necesidade de obter datos de eQTL máis detallados en tipos celulares específicos e etapas de desenvolvemento relevantes para a enfermidade, e (iii) a posibilidade de que as variantes estean regulando a expresión xénica mediante outros mecanismos, como por exemplo, o de *splicing* alternativo, discutido a continuación.

En canto ao segundo obxectivo (**Capítulo 2**), a investigación comprende a exploración de fundamentos xenéticos adicionais nos TEA, facendo fincapé no papel das DNVs no *splicing*

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

alternativo, un mecanismo que diversifica as isoformas proteicas dun único xen, con especial relevancia no desenvolvemento neuronal.

Centrándonos nunha cohorte española de 360 tríos (paciente afecto con nai e pai non afectos) de TEA, e empregando ferramentas de *deep learning* (como son SpliceAI, SpliceVault e ABSsplice; **Figura 17**) para predicir e analizar os sitios de *splicing* e o impacto que as variantes detectadas teñen sobre os mesmos, identificáronse variantes con alto poder disruptivo e deliñáronse os procesos moleculares e biolóxicos implicados no trastorno en relación cos eventos de *splicing* aberrante.

Con isto, identificáronse DNVs e filtráronse en función de determinados criterios de calidade empregados recorrentemente na bibliografía. O conxunto total de DNVs foi puntuado polo seu potencial efecto no *splicing*, seleccionando aquelas variantes, que con alta seguridade, provocan unha disrupción no mesmo, obténdose 43 variantes de *splicing* de alta confianza (puntuación de SpliceAI ≥ 0.8 (**Figura 18**)).

Para a validación *in silico* das mesmas, utilizouse SpliceVault, unha ferramenta que proporciona datos para os eventos de *splicing* máis comúns en cada unha das posicións nas que acontencen as variantes. Así, se a predicción de splicing anotada por SpliceAI coincide cun dos 4 eventos máis comúns, pódese confirmar, con alta probabilidade, que a variante xerará o efecto predito por SpliceAI.

A especificidade a nivel de tecido para cada variante investigouse utilizando a ferramenta ABSplice en 49 tecidos humanos, e considerouse unha variante validada se tiña unha puntuación superior a 0.2 en alomenos un tecido (puntuación con porcentaxes de validación similares á escollida para SpliceAI).

Con isto, atopáronse 31 variantes en 31 xenes distintos. De seguido, empregáronse conxuntos de datos complementarios, incluíndo, entre outros, xenes relacionados cos TEA e implicados no proceso de *splicing*, para poder concluír a posible implicación nos TEA dos xenes seleccionados mediante a análise. Así, atopáronse un total de 10 xenes con potenciais implicacións no proceso de *splicing* (*CACNA1I*, *CBLB*, *CLTB*, *DLGAP1*, *DVL3*, *KIAA0513*, *OFD1*, *PKD1*, *SLC13A3* e *SCN2A*) (**Figura 18**).

Por outro lado, empregáronse datos doutro estudo que integra variantes codificantes *de novo* e herdadas, sen efectos aparentes no proceso de *splicing*, para poder realizar análises comparativas (**Táboa 14**). Ademais, construíuse unha rede de xenes asociados aos xenes validados en cerebro, empregando HumanBase (**Táboa suplementaria 3**). Así, realizáronse análises de GO, mediante análises de enriquecemento e comparación de *clusters* de termos enriquecidos grazas á ferramenta ClusterProfiler.

En resumo, os datos preliminares suxiren que o tecido adiposo pode desempeñar un papel nos TEA, malia que moitas das variantes puntuadas polo seu maior efecto no proceso de *splicing* en tecido adiposo amosan efectos similares noutros tecidos e patróns de expresión limitados ao mesmo, o que obriga a seguir explorando a súa relevancia. Tamén salienta a posible relevancia de tecido testicular, aínda que istos achados tamén deben ser interpretados con cautela, xa que este tecido é, xunto con cerebro, no que con máis relevancia acontece o fenómeno de *splicing* alternativo. En resumo, isto subliña a complexidade da xenética dos TEA,

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

na que múltiples tipos de variantes, con efectos en múltiples tecidos, están implicados (**Táboa 17**, **Figura 19**).

Ademais, o estudo identificou xenes con DNVs con alto efecto no *splicing* e explorou os seus papeis funcionais mediante análises GO. Os achados semellan apuntar a unha división dos enriquecementos en categorías GO para xenes con variantes de *splicing* e xenes con variación codificante, converxendo en rutas biolóxicas previamente implicadas nos TEA (**Figura 20, 21**, **Táboa 18**).

Se ben é certo, a investigación adopta a predición e validación *in silico* de variantes de *splicing* empregando un limiar conservador (SpliceAI $\Delta \geq 0.8$) con potenciais falsos negativos. Sen embargo, a validación *in vitro* foi inalcanzable debido a problemas de dispoñibilidade de mostras, o que levou a establecer un *framework* que posibilitou a súa validación doutros xeitos. Polo tanto, futuras direccións inclúen a validación funcional, a integración multi-ómica, e a exploración da interacción entre o *splicing* e outros mecanismos reguladores, para desentrañar máis a fondo a complexa rede molecular que contribúe ós TEA.

En canto ao terceiro obxectivo (**Capítulo 3**), o estudo aborda o reto de comprender o papel das variacións non codificantes nos TEA, sinalando as limitacións históricas no estudo de tales variacións.

Os esforzos recentes de secuenciación do xenoma completo ampliáronse a cohortes máis grandes, facendo fincapé na complexidade da detección de asociacións debido á dificultade na anotación funcional das mutacións non codificantes ademais da heteroxeneidade, xenética e fenotípica, salientable nos TEA.

A investigación adoptou por tanto, unha metodoloxía imparcial, seleccionando elementos non codificantes candidatos en función da súa actividade en tecidos relevantes para os TEA, aproveitando iniciativas como o proxecto ENCODE (**Figura 22**). A secuenciación dirixida en base ao tecido no que o elemento regulador ten actividade detectada funcionalmente, e a análise da carga mutacional, dan prioridade ás mutacións con impacto funcional, evitando os problemas de potencia estatística que xorden ao testar asociacións en base ao amplio número de categorías non codificantes (a través da intersección de factores tales como a conformación da cromatina, modificación epixenéticas, conservación, *etc*.).

Unha vez seleccionadas as rexións candidatas, estas foron analizadas en canto á presencia de mutacións *de novo* e variantes herdadas ultrararas (isto é, detectadas nunha única familia da cohorte), con motivo da súa relevancia no desenvolvemento dos TEA. Estas variantes foron obxecto de múltiples filtros de control de calidade (**Figura 24**).

Unha vez designada a lista final de variantes, estas foron puntuadas en canto ao seu efecto regulador grazas ao algoritmo Sei (**Figura 25**). Así, empregouse o marco Sei para avaliar as consecuencias funcionais destas mutacións a través dunha asignación, a nivel de xenoma completo, de calquera secuencia dada ás actividades reguladoras, clasificadas en 40 clases distintas (**Táboa 19**). Unha vez puntuadas as distintas variantes, estas foron asignadas á clase reguladora correspondente.

Logo desta puntuación, as variacións de alto impacto, isto é, aquelas cunha diferencia no efecto regulador > 1.1 (puntuación obtida a través de Sei), foron asignadas a xenes candidatos mediante a integración de datos de PLAC-Seq. Estes mapas interactómicos específicos de cada

tipo de célula cerebral foron empregados para confirmar rexións reguladoras non codificantes e identificar xenes probablemente regulados por estas variantes. Isto permitiu a creación dos conceptos: (i) xenes PARV (xenes que inclúen nos seus promotores variantes reguladoras) e (iii) xenes PLARV (xenes que non inclúen estas variantes, pero que, do contrario, están ligados ás mesmas por contactos de cromatina) (**Figura 26**).

A cohorte seleccionouse estratexicamente, incluíndo 200 tríos tras aplicar criterios de exclusión relacionados con diagnósticos negativos no estudo de exoma e CNVs (ningún achado puido concluintemente ser relacionado coa presenza do trastorno). As rexións reguladoras de interese elixíronse baseándose no rexistro de elementos cis-reguladores candidatos (cCREs) de ENCODE, centrándose nos tecidos cerebrais e gastrointestinais. A secuenciación dirixida destas rexións identificou variantes, posteriormente filtradas e sometidas a procedementos de control de calidade.

Ademais, a investigación profundou na priorización de variantes, considerando listas de consenso de xenes de risco de TEA coñecidos e seleccionando aquelas variantes con interaccións confirmadas, caracterizadas a nivel celular, cos distintos xenes de risco nos TEA (**Táboa Suplementaria 9)**. Realizáronse tamén análises de enriquecemento de factores de transcrición para comprender mellor as implicacións biolóxicas das variantes identificadas e validar os resultados obtidos (**Táboas suplementarias 7, 8**). O estudo tamén explorou a presenza de variantes en dominios topoloxicamente asociados (TADs), ofrecendo un contexto espacial á organización xenómica.

En resumo, o estudo implementou unha metodoloxía que incluía unha secuenciación dirixida, que abarcou 85,394 cCRES activos en tecidos cerebrais e gastrointestinais, a predición do impacto regulador, e análises integradoras para dilucidar a contribución das variantes non codificantes aos TEA.

Para as variantes *de novo*, identificáronse 164 ocorrencias dentro das rexións reguladoras candidatas. As variantes herdadas, un total de 53,023, mostraron un sesgo de transmisión materna, en consonancia con observacións anteriores para CNVs e SNVs privadas herdadas.

Unha vez priorizadas polo seu impacto en funcións reguladoras, destacou a prevalencia de variantes asignadas á categoría CTCF. En concreto, con alta potencia estatística, as variantes herdadas supuxeron unha perda na afinidade por CTCF. Aínda que as variantes *de novo* non mostraron diferenzas significativas nesta clase reguladora, seguiron a mesma tendencia. Debido a que CTCF é un elemento que demarca límites na conformación cromatínica, amósase deste xeito a implicación dunha desregulación a nivel de estrutura 3D nos TEA, un mecanismo implicado recentemente (**Figuras 29, 30**).

Mediante a análise de interaccións PLAC para vincular variantes *de novo* e herdadas aos xenes PARV e PLARV puidemos concluír que, nunha proporción considerable de casos, as variantes reguladoras actúan sobre xenes a longa distancia, mediante contactos de cromatina, e non sobre o xen en maior proximidade xenómica (**Figuras 31-33**). Ademais, a maioría de xenes non foron identificados previamente como xenes de risco nos TEA. Aínda así, en termos de función xenética, identificáronse enriquecementos significativos para as categorías de funcións moleculares e procesos biolóxicos, incluídos aqueles relevantes para os TEA, coma o proceso

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

de *splicing*, organización da sinapse, o desenvolvemento do axón, e a regulación da estabilidade proteica (**Figura 34**).

O estudo concluíu centrándose na priorización de variantes e unha posible conexión xenotipo-fenotipo, facendo fincapé nas variantes dentro de rexións non codificantes con conexións validadas con xenes diana e asociacións coñecidas co risco de TEA. Destacan casos concretos, que mostran pacientes con múltiples variantes e características clínicas asociadas. En particular, o paciente co maior número de mutacións de alto impacto regulador, ten outro irmán afecto. Isto é relevante en canto a que, debido á inexistencia dun diagnóstico definitivo a nivel de exoma, o feito de que este paciente albergue un salientable número de variantes de alto impacto no seu xenoma regulador, pode estar vinculado co desenvolvemento do trastorno, e probablemente, co risco de recorrencia que amosa a familia. Por outro lado, establécense asociacións preliminares cos xenes *RNF152* e *MIDN*, aportando evidencia adicional á súa recente implicación nos TEA.

En xeral, os descubrimentos chave inclúen a identificación dun alto potencial regulador no 28% das variantes *de novo* e o 25% das variantes herdadas en pacientes con resultados negativos de análises previas. Os datos de PLAC-Seq permiten a asignación funcional de xenes para aproximadamente o 80% e o 85% das variantes *de novo* e herdadas, respectivamente (**Táboas suplementarias 4, 6**).

En resumo, os resultados obtidos na presente tese doutoral salientan a complexidade da anotación funcional das rexións reguladoras, e a necesidade imperante de establecer algoritmos para esta labor, que permitan discernir mutación probablemente causais, e así poder abordar de xeito concluínte o tema da herdabilidade perdida.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ABSTRACT

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

2 ABSTRACT

Despite the high heritability estimates in autism spectrum disorders (ASD), there is a lack of direct confirmation from current genetic studies, leading to the term "missing heritability", which signifies the incomplete understanding of the genetic factors contributing to ASD etiology. In recent years, speculation surrounding this named "missing heritability" has shifted towards the non-coding genome, a portion nearly 99 times larger in genomic size than protein-coding genes. However, elucidating the role of the non-coding genome in genetic studies is challenging because, unlike coding regions that directly produce proteins, the functions of many non-coding regions are not as well annotated or understood. Thus, identifying and characterizing these functional elements requires specific bioinformatic analysis.

In this PhD thesis, we have applied specific bioinformatic workflows aiming to decipher part of the missing heritability lying within the non-coding genome, elucidating novel genes, biological pathways, and molecular functions. This was achieved by (i) conducting an eQTL colocalization analysis that integrates expression data from GTEX with summary statistics from the largest ASD GWAS to date, (ii) assessing the role of splicing variation in ASD by mining exome data using deep learning methods, and (iii) conducting a targeted sequencing study of a selection of regulatory elements (cCREs) from ENCODE in ASD trios to detect *de novo* and inherited mutations and their functional implications.

Through a combination of bioinformatics methods and diverse genetic data collections, we have uncovered insights into a broad spectrum of genetic variation. In summary, the findings from this *in silico* characterization of common, splicing, and regulatory variants include the identification of specific genes most likely to be affected by non-coding common variation, a deeper understanding of how splicing variation may mediate disease risk, and the establishment of a crucial role for chromatin conformation in ASD.

This thesis aims to further investigate the role of missing heritability in ASD genetics using different genetic and bioinformatic study strategies. In this way, novel potential ASD risk genes were pinpointed, and the role of mutations within regulatory elements and chromatin structure has been unveiled. Future functional validation of these targets will open new avenues for improving ASD diagnosis and, subsequently, treatment.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

INTRODUCTION

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

3 INTRODUCTION

Part of the content of this chapter has been published as: Dominguez-Alonso S[1], Carracedo A[2], Rodriguez-Fontenla C[3]. The non-coding genome in Autism Spectrum Disorders. Eur J Med Genet. 2023;66(6):104752. doi: 10.1016/j.ejmg.2023.104752. Epub 2023 Apr 5. PMID: 37023975.

1. Grupo de Medicina Xenómica, Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidad de Santiago de Compostela, Santiago de Compostela, Spain.

2. Grupo de Medicina Xenómica, Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidad de Santiago de Compostela, Santiago de Compostela, Spain; Grupo de Medicina Xenómica, Fundación Instituto de Investigación Sanitaria de Santiago de Compostela (FIDIS), Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidad de Santiago de Compostela, Santiago de Compostela, Spain.

3. Grupo de Medicina Xenómica, Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidad de Santiago de Compostela, Santiago de Compostela, Spain.

3.1 AUTISM SPECTRUM DISORDERS (ASD)

The understanding of autism has experienced substantial transformation within the Diagnostic and Statistical Manual of Mental Disorders (DSM), the globally recognized handbook used by healthcare professionals as the definitive guide for diagnosing mental disorders. This evolution mirrors changes in perspectives and advancements in scientific knowledge.

Initially characterized by Leo Kanner in 1943[1] as a form of childhood schizophrenia attributed with cold parenting, autism was later recognized as a set of interconnected developmental disorders.

Consistent with Kanner's viewpoint, the second edition of the DSM, the DSM-2, released in 1968, characterized autism as a psychiatric condition; a manifestation of childhood schizophrenia distinguished by a detachment from reality.

The DSM-3 in 1980 marked a pivotal shift, establishing autism as a distinct “pervasive developmental disorder”, with specific diagnostic criteria emphasizing communication impairments and behavioral anomalies within the first 30 months of life.

The subsequent DSM-4 in 1994 categorized autism as a spectrum disorder, acknowledging various conditions like Asperger’s disorder and childhood disintegrative disorder and eliminating the prerequisite for symptoms to manifest before the age of 30 months.

In the 1990s, genetic research attempted to pinpoint autism-related genes, but the elusive nature of a single genetic basis became apparent. Following the completion of the Human Genome Project in 2003, numerous studies sought to identify “autism genes”, discovering hundreds without exclusive links to autism.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

As a result, the DSM-5[2] in 2013 embraced the concept of a continuous spectrum, introducing the term "autism spectrum disorders" (ASD) to address diagnostic inconsistencies and eliminate specific subcategories like Asperger's syndrome and classic autism. The revised criteria emphasized two core features, elaborated further in Table 1: persistent impairment in social communication and interaction, as well as restricted and repetitive behavioral patterns, both manifesting early in childhood.

Table 1. Diagnostic criteria for ASD according to the DSM-5.

A. Persistent deficits in social communication and social interaction across multiple contexts, as manifested by the following, currently (observed at the present moment) or by history (documented or reported in the individual's past): • Deficits in social-emotional reciprocity. • Deficits in nonverbal communicative behaviors used for social interaction. • Deficits in developing, maintaining, and understanding relationships.
B. Limited and repetitive behavior patterns, interests, or activities, exhibited through a minimum of two of the following, currently or by history: • Repetitive motor movements, utilization of objects, or speech that is stereotyped in nature. • A strong preference for consistency, rigid adherence to routines, or ritualized patterns in both verbal and nonverbal behavior. • Intensely focused and highly restricted interests that deviate from the norm in terms of intensity or concentration. • Over- or under- reactivity to sensory stimuli or atypical fascination with sensory aspects of the environment.
C. Symptoms must be evident in the early stages of development, although they may not become fully apparent until social demands surpass limited capacities or are concealed by learned strategies in later life.
D. Symptoms result in significant impairment in social, occupational, or other crucial aspects of the individual's current functioning.
E. These disruptions cannot be more effectively accounted for by intellectual disability (intellectual developmental disorder) or global developmental delay. While intellectual disability and ASD often coincide, for concurrent diagnoses, social communication should fall below the expected level for general developmental ability.

Furthermore, the DSM-5 delineates severity based on the impairments in social communication and the presence of restricted, repetitive patterns of behavior, distinguishing between three levels (Table 2).

Table 2. Severity levels for ASD according to the DSM-5.

Severity level	Social communication	Restricted, repetitive behaviors
Level 3 "Requiring very substantial support"	Profound deficiencies in both verbal and nonverbal social communication skills result in significant impairments in functioning, characterized by extremely limited initiation of social interactions and minimal responsiveness to social overtures from others.	Rigidity in behavior, extreme challenges in adapting to change, or other prominently restrictive and repetitive behaviors significantly impede functioning across various aspects. The individual experiences significant distress or difficulty when attempting to shift focus or alter actions.
Level 2 "Requiring	Significant shortcomings in both verbal and nonverbal social communication skills are	The inflexibility of behavior struggles with coping with change, or other evident

substantial support"	observed, with noticeable social impairments persisting even with support. The affected individual demonstrates limited initiation of social interactions and exhibits reduced or abnormal responses to social overtures from others.	restricted and repetitive behaviors occur with enough frequency to be noticeable to casual observers and disrupt functioning in various contexts. The individual experiences distress and/or encounters difficulty when attempting to change focus or actions.
Level 1 "Requiring support"	In the absence of support, deficits in social communication result in noticeable impairments. There is difficulty initiating social interactions, and evident instances of atypical or unsuccessful responses to social overtures from others. The individual may seem to have decreased interest in social interactions.	The inflexibility of behavior results in significant interference with functioning in one or more contexts. There is difficulty transitioning between activities, and challenges in organization and planning impede independence.

3.1.1 Prevalence

According to a systematic review conducted in 2012, the median global prevalence of ASD was found to be 62 per 10,000 children and boys consistently exhibited a higher prevalence compared to girls[3]. Over the past decade, prevalence estimates of ASD have demonstrated a notable upward trajectory, and significant variations exist both within and across regions (Figure 1).

Indeed, ten years later, the global median prevalence of ASD stands at 100 per 10,000 children (*i.e.*, global prevalence of 1%), with a range spanning from 1.09 per 10,000 to 436.0 per 10,000. Furthermore, the current estimate for the median male-to-female ratio is 4.2[4].

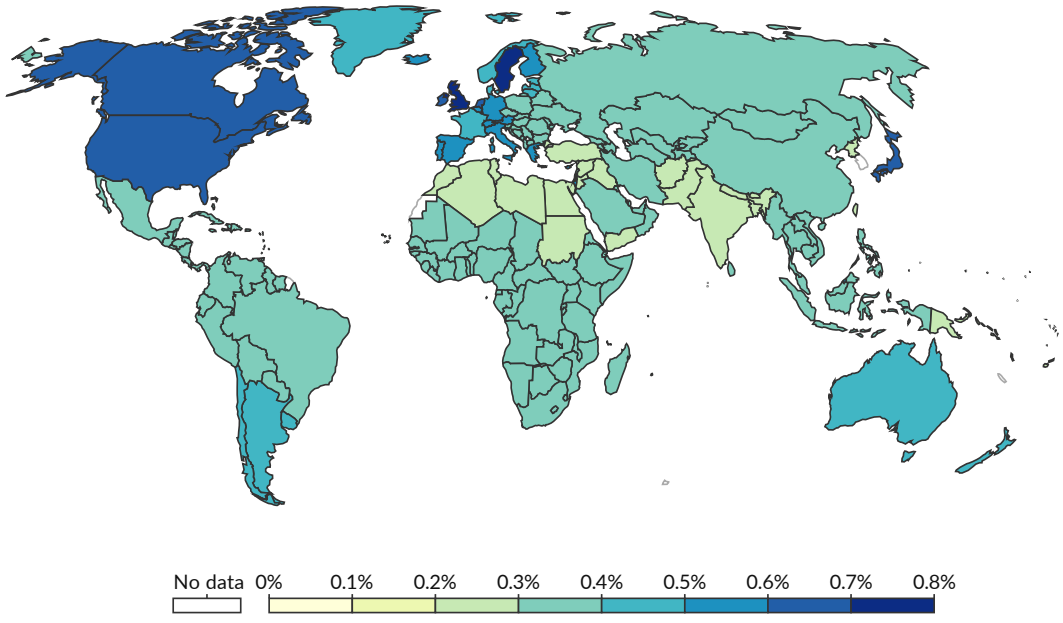


Figure 1. Global ASD prevalence in 2019. The worldwide prevalence of ASD is slightly below 1%, although estimates are higher in high-income countries[5]. The legend illustrates the color code for prevalences, with countries lacking data represented in white and grey dashed lines. The figure represents open-access data that have been previously age-standardized and obtained from the Institute for Health Metrics and Evaluation (Global Burden of Disease)[6].

3.1.2 Comorbidities

As previously discussed, ASD are characterized by core symptoms encompassing social communication difficulties, restricted interests, and repetitive behaviors. Beyond these fundamental diagnostic features, individuals with ASD bear a heightened burden of comorbid conditions in contrast to the general population.

Comorbidity, within the realm of health, refers to the simultaneous occurrence of one or more additional diseases or disorders alongside a primary condition. In this context, a comorbid condition constitutes a secondary diagnosis featuring distinct core symptoms separate from those associated with the primary disorder, ASD. For instance, individuals diagnosed with ASD are more prone to being overweight or obese and are more likely to have attention-deficit/hyperactivity disorder (ADHD), intellectual disability (ID), sleep disorders, and gastrointestinal (GI) diseases. Prevalence estimates for various other genetic disorders, including but not limited to fragile X syndrome (FXS), Rett syndrome and Angelman's syndrome, are also considerably higher in individuals with ASD compared to the general population[7]. Furthermore, ASD patients are more likely to experience co-occurring psychiatric conditions such as anxiety and depression, compared to those without ASD (Table 3).

Table 3. Most common comorbidities in ASD.

	Comorbidity	Prevalence	Further consequences and comorbidity interplay
Neurodevelopmental	ADHD	~28%	Profound deficits in adaptive and executive functioning manifest as challenges in areas such as attention, hyperactivity/impulsivity, social interaction, and planning impairments[8].
	ID*	<70%	ID may impede responsiveness to interventions commonly employed during psychiatric hospitalization to address severely disruptive behavior and could heighten the likelihood of experiencing restraint and/or seclusion[9].
	Tics and Tourette syndrome	2.6-36%	Motor tics, stereotypies, and semi-voluntary movements are frequently observed in ASD individuals[8].
	Motor abnormalities	50-85%	The prevalence of motor impairments in ASD poses a notable obstacle to a child's ongoing social communication development and adaptive functioning[10].
Psychiatric	Anxiety	1.47-54%	Anxiety significantly influences the trajectory and fundamental aspects of ASD, intensifying social withdrawal and repetitive behaviors. Untreated comorbid anxiety in ASD has been linked to the emergence of depression, aggression, and self-injury[11].
	Mood disorders	11% (MD) 5% (BP) 12.6%(PD)	Depressive symptoms involve changes in stereotyped and compulsive behaviors, loss of language, social withdrawal, mood swings, obsession, and occasional self-injurious behaviors. Other symptoms may include sleep disturbances or decreased communication[7,8].
	OCD	9-22%	Individuals with OCD often exhibit heightened rates of obsession and compulsion, coupled with greater severity of symptoms. However, they are less likely to experience thoughts related to aggression: while repetitive behaviors may not necessarily cause distress, they tend to be time-consuming.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

			Interrupting these repetitive behaviors can trigger sudden bursts of anger[8].
	Aggressive behavior (including self-harm)	35–50%	Aggressive behavior is significantly linked to psychiatric hospitalization. Beyond its impact on treatment, aggression also hampers independence, community engagement, and the ability to build relationships[11].
	FED	1.4-7.9%	Anorexia nervosa stands as the most prevalent FED, followed by bulimia nervosa and binge-eating disorder. In individuals with ASD, atypical eating and feeding behaviors, such as food refusal, food selectivity, obsessive routines for meals, and preferences for specific food colors and textures, are common. These behaviors, coupled with GI symptoms, have the potential to exacerbate ASD-specific symptomatology[8].
Medical conditions	Epilepsy	10-30%	A significant proportion of children with ASD, up to 60%, exhibit abnormal electroencephalogram results, in contrast to the 6%-7% prevalence observed in typically developing children[7].
	GI symptoms	46-84%	Children with ASD commonly experience a range of GI issues (*e.g.,* chronic constipation and/or diarrhea, nausea and/or vomiting, ulcers, food intolerance, *etc*.). Children exhibiting these symptoms are prone to experiencing irritability, agitation, social withdrawal, lethargy, and hyperactivity[8].
	Sleep disorders	13-80%	Sleep disturbances are not only linked to issues like daytime fatigue but can also contribute to difficulties in learning, increased hyperactivity, inattention, heightened anxiety, aggression, and various health problems. Moreover, there is an observed association between sleep disorders and GI dysfunction in children with ASD (24.5% of children)[7,8].
	Obesity**	17%	The core and associated symptoms of ASD, which can lead to restricted food preferences, decreased physical activity, and the use of food as a reward in certain behavioral therapies, are likely contributing factors to excess body weight in individuals with ASD[12].
	Immune deficiency and dysfunction	25%	Children with GI disorders are more likely to have immunodeficiency[7].
Genetic disorders	FXS	2-3%	Children diagnosed with both FXS and ASD exhibit higher rates of social anxiety, ID, and repetitive behaviors[7].
	Rett syndrome	1–8%	Children diagnosed with Rett syndrome show partial or complete loss of acquired spoken language and stereotypic hand movements[13,14].

*The repeated exclusion of individuals with intellectual disabilities from ASD-focused research may affect the accuracy of its prevalence estimates within the ASD population. **Studies on comorbidities like obesity, primarily conducted in North America, limit generalizability to other countries. *BP, bipolar disorders; FED, feeding and Eating Disorders; MD, major depression; OCD, Obsessive-compulsive disorder; PD, personality disorders.*

3.1.3 Genetic vs. non-genetic factors

The etiology of ASD is multifactorial and involves a combination of genetic and non-genetic (environmental) factors. The interplay between these factors is an active area of research, and our understanding of the contributions of each factor continues to evolve. As

USC DE COMPOSTELA

thoroughly discussed in this and following chapter sections, it is important to recognize the heterogeneity within the ASD population, as different individuals may have different combinations of genetic and environmental risk factors.

On the one hand, studies of twins have shown a higher concordance rate for ASD in identical twins (monozygotic; MZ) compared to fraternal twins (dizygotic; DZ), indicating a genetic influence. On the other hand, various environmental factors during prenatal and perinatal development have been investigated for their potential role in ASD[15].

It is also crucial to highlight that the relationship between genetic and non-genetic factors is not always straightforward. There is growing recognition of gene-environment interactions as part of the complex interplay between genetic and non-genetic factors, where genetic predispositions may interact with specific environmental factors to increase the risk of ASD[15,16] (Figure 2, section "3.1.3.2 Non-genetic factors: Environmental factors and epigenetic modifications").

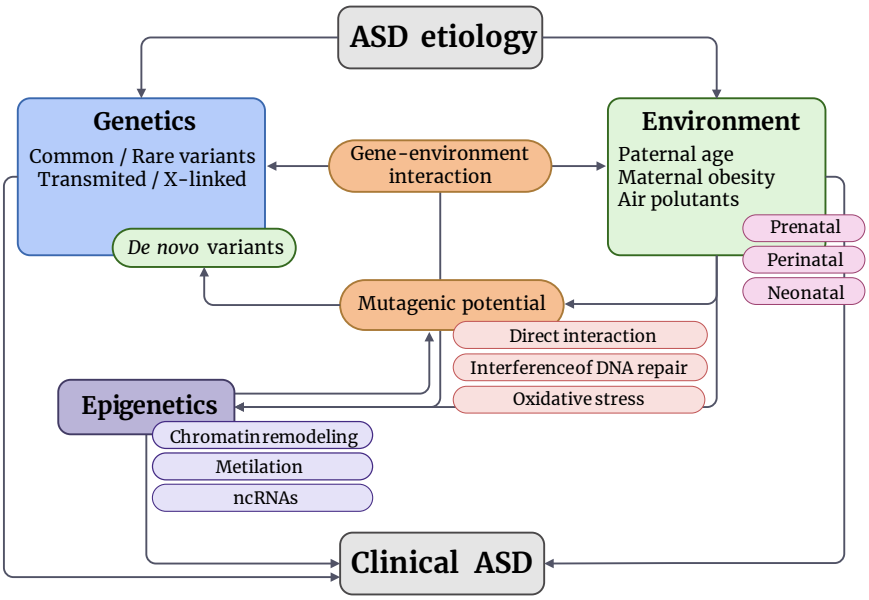


Figure 2. Diagram of the interaction between genetic and environmental risk factors in the development of ASD. While *de novo* variants represent genetic causes of ASD, are considered as environmental causes (in green) as they do not contribute to heritability. This figure was adapted from Pugsley *et al.*[16] and created with BioRender.com.

3.1.3.1 Genetic factors: Early insights from twin and family studies

In its early definitions around 1943, there was a damaging belief that ASD was caused by a cold parenting style. However, thirty years later, twin studies revolutionized the field of ASD research by demonstrating that ASD have a high heritability (*i.e.*, the proportion of phenotypic variation that is attributable to inherited genetic factors, in contrast to environmental ones) and that increasing relatedness to an affected individual increases risk (for a review, see Ronald & Hoekstra[17]).

As previously discussed, the phenotypic variance in ASD can be accounted for by both genetic and environmental factors. Consequently, heritability estimates are classified into four categories: additive genetic influences (A, determined by the cumulative effect of alleles at all

loci influencing the phenotype), shared environmental factors (C), non-additive genetic influences (D, genetic influences encompassing interactions between alleles at the same locus (dominance) or different loci (epistasis)), and unique environmental influences (E)[18].

In the literature, two distinct definitions of heritability are commonly used. Narrow-sense heritability (h^2) exclusively considers the A component and signifies the extent to which genes passed down by parents contribute to their children's phenotype. Conversely, broad-sense heritability (H^2) is the ratio of total genetic variance (A and D) to total phenotypic variance. It elucidates the proportion of phenotypic variation attributed to genetics, encompassing gene-gene interactions and dominant effects[19].

Twin studies serve as a commonly employed approach to ascertain heritability by leveraging the nearly 100% shared genetic information in MZ pairs. Since MZ pairs are genetically identical (A and D) and share the same environmental factors (C), any variations in phenotypic traits can be attributed to unique environmental influences (E). In contrast, DZ twins, like other sibling pairs, only share 50% of their genetic makeup and are presumed to share the same environment (C). Assuming that the unique environment (E) contributes equally to the development of the phenotype of interest in both MZ and DZ pairs, it becomes possible to estimate the variance contributed by both additive and non-additive genetic effects (H^2) by comparing the phenotypic correlation in MZ and DZ pairs[18,19].

In 1977, the initial estimate of H^2 in twins was reported, relying on an examination of 10 DZ and 11 MZ pairs. Among the 11 MZ pairs, four (36%) showed concordance for ASD, while none of the DZ pairs exhibited concordance[20].

Since then, numerous extensive twin studies, along with subsequent meta-analyses, have been published, consistently reinforcing the notion of the substantial heritability of ASD. Moreover, H^2 estimates are largely comparable among the studies that have been published[17,21].

In a recent meta-analysis encompassing seven primary twin studies with a clinical diagnosis of ASD, it was reported that H^2 estimates ranged from 64% to 93%. The correlations for MZ twins were notably high at 0.98, while DZ twins exhibited correlations at 0.67. These findings have resulted in heritability estimates exceeding 90%, thus establishing ASD as one of the most heritable behaviorally-defined conditions[22].

Furthermore, family studies (*i.e.*, studies using twins along with full siblings, cousins and maternal and paternal half-siblings, enhancing both generalizability and precision in findings) have demonstrated that the relative recurrence risk (RR) (*i.e.*, the probability that a condition present in one or more family members will recur in another relative in the same or future generations) of a child having ASD is linked to the degree of shared genetic material with affected relatives[23]. In a population-based sample of 14,516 children diagnosed with ASD, the recurrence risk for ASD, compared to the general population, was estimated to be 153.0 for MZ twins, 8.2 for DZ twins, 10.3 for full siblings, 3.3 for maternal half-siblings, 2.9 for paternal half-siblings, and 2.0 for cousins. Of note, this is crucial for early diagnosis in ASD relatives.

3.1.3.2 Non-genetic factors: Environmental factors and epigenetic modifications

As already discussed in the previous section, twin studies have confirmed a strong genetic heritability for ASD but have also supported an important contribution of environmental

factors. Relying on concordances between DZ pairs, twin studies have additionally indicated that environmental factors may account for 7–35% of the variance in ASD predisposition[21,22].

Shared environmental factors (*i.e.,* environmental influences that are common to both twins and contribute to the similarity of children growing up in the same family) have not shown significant associations with disease onset in the latest, largest studies[22–25].

In contrast, the non-shared environment (*i.e.,* unique environmental influences that contribute to the differences among children growing up in the same family) has been identified as influencing ASD, albeit to a much lesser extent than genetic factors[26]. Nonetheless, the diverse clinical and biological manifestations seen in ASD strongly indicate that, in individuals with a genetic predisposition, environmental risk factors likely interact or synergize to establish a "threshold point" that could contribute to dysfunction[15,16]. This influence may be guided through (i) direct interaction with genetic material, (ii) interference with endogenous DNA repair mechanisms, and (iii) the induction of indirect DNA damage through exposure-induced oxidative stress[16].

As of now, three biologically relevant windows for brain development (namely, the conception, prenatal and early postnatal periods) have been hypothesized as critical in terms of exposure to environmental risk factors (for a detailed summary see Gialloreti *et al.*[15]).

Commonly cited environmental risk factors include preterm birth, exposure to toxins and heavy metals in utero, the proximity of birth and maternal residence to sources of ambient air pollution, the use of drugs during pregnancy (both prescribed and illicit), and various parental health factors, such as age, intelligence quotient, maternal obesity and diabetes, as well as vitamin status before and during gestation[15,25].

Moreover, neurological factors, such as differences in brain structure and function have also been observed in individuals with ASD[16].

Furthermore, epigenetic modifications, which lead to changes in gene expression without alterations in the underlying DNA sequence, may also play a role[27]. As an example, DNA methylation, which affects gene expression, chromatin structure, and the proper function of regulatory elements, has been proposed as a mechanism modulating the neurodevelopmental origins of ASD[27–29]. Moreover, these modifications can be influenced by environmental factors and could contribute to the development of ASD: levels of maternal plasma folate during pregnancy have been reported to impact the methylome landscape in newborns, with influences in neural tube defects and neurological functions[30].

3.2 GENETIC BASIS OF ASD: LIABILITY AND GENETIC STUDIES

As introduced in the preceding section, twin and family studies have supplied essential insights into the heritability of ASD and their overwhelming heterogeneity and complexity[31]. Nonetheless, they do not offer details about the specific genes implicated or the quantity and prevalence of their variations.

The genetic complexity of ASD encompasses a wide array of variants with differences in their frequency in the general population (*e.g.*, rare *versus* common), their nature (*e.g.,* large chromosomal rearrangements, small insertions or deletions, and single-nucleotide variants), their mode of inheritance, *etc*.

Regarding their frequency in the general population, variants can be classified in terms of minor allele frequencies (MAF). MAF refers to the frequency at which the second most common allele occurs in a given population at a specific locus (*i.e.*, its location in the genome). With this regard, variants can be divided into:

i. Common variation, which refers to any variation from the reference genome that is common in the population (*i.e.*, MAF > 1%-5%, depending on the study[32,33]).
ii. Low-frequency variation (*i.e.*, less common variants (0.01 < MAF < 0.05)).
iii. Rare variation (*i.e.*, any variation from the reference genome that is present in the population at a frequency ≤ 1%).
iv. Very rare variants (MAF < 0.01%).

In terms of the inheritance pattern, variants may be:

i. Transmitted (*i.e.*, genetic variation from the reference genome that exists in every cell of a parent (or parents) and is subsequently passed on to their offspring).
ii. *De novo* variation (DNV) (*i.e.*, genetic alterations that are present in the offspring but are either absent in the parents or present only in the parental germ cells).
iii. X-linked (*i.e.*, variants that are present in genes on the X chromosome).

In terms of their structure, variants may be divided into:

i. Single-nucleotide variants (SNV), which comprise substitutions of a single nucleotide in the DNA sequence, whether rare or common in the population. In addition, single-nucleotide polymorphisms (SNP) correspond to the substitution of a single nucleotide in the DNA sequence that is common in the population.
ii. Insertion-deletions (indels) (*i.e.*, variations in the DNA that involve the insertion or deletion of a small number of nucleotides).
iii. Structural variation (SV), which corresponds to regions of DNA approximately ≥ 1 kilobase (kb) in size[34] (commonly categorized as deletions, duplications, insertions, inversions, and translocations), involving various configurations of DNA alterations (namely, gains, losses, or rearrangements). Moreover, copy number variants (CNV) are a class of SV mainly represented by deletions and duplications in the number of copies in regard to the reference genome (of note, the boundary between a small CNV and an indel is arbitrary).

Regarding their genomic localization variants may be:

i. Coding variants (*i.e.*, a variation from the reference genome that is present in the portion of the genome that codes for the transcription of proteins (1%)[35]).
ii. Non-coding variants; variation present in the remaining 99% of the genome, including introns, regulatory elements, and intergenic segments.
iii. Splicing variants. As we will further discuss in section "3.3.1.6 Splicing variants", splicing is the process by which introns are excised and exons are concatenated together. Splice variants are thus, variants present either in (or

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

near) the exon-intron boundary (*i.e.*, acceptor/donor splice sites) or in the splicing branchpoint (a cis-acting intronic motif required for messenger RNA (mRNA) splicing[36]).

Regarding the coding consequence impact, variants may be:

i. Synonymous (*i.e.*, variants that result in no change in the protein sequence (also known as silent mutations)).
ii. Missense variants (*i.e.*, point mutations where a single nucleotide change results in the change of the amino acid sequence).
iii. Nonsense variants (*i.e.*, point mutation that results in the production of a stop codon, thereby terminating protein synthesis prematurely).
iv. Frameshift (*i.e.*, insertion or deletion of several nucleotides in a number not divisible by three, causing a shift in the reading frame during translation).

De novo frameshift, canonical splice acceptor/splice donor, and nonsense variants are collectively known as protein-truncating variants (PTVs), genetic variants predicted to shorten the coding sequence of genes.

These variants collectively contribute to an increased individual risk of ASD. However, generalizing the relative contribution of each of these mutations to the etiology of the disorder proves challenging, as estimates vary based on study design. The variability in estimates can be predominantly attributed to (i) sample heterogeneity (*e.g.*, autistic individuals without ID exhibit a higher proportion of the additive variance explained by common genetic variants compared to individuals with ASD and ID)[37], (ii) study design disparities (investigations of families with a single affected child (simplex families), reveal a higher frequency of *de novo* variants but a depletion of inherited risk, in contrast to families with multiple affected children (multiplex families))[27,38], and (iii) differences in the methods employed to predict heritability estimates. Moreover, another source of bias includes ascertainment status or diagnostic system[25]. All these sources of bias will be further discussed in section "3.2.5 The "missing heritability" in ASD.

However, recent large-scale studies leveraging advancements in DNA sequencing technologies, extensive patient cohorts, and widespread data sharing (as further elaborated in this section), have produced significantly more reliable results compared to initial twin and family studies discussed in the preceding section.

The most recent assessment of h^2 (82.7%) comes from analyzing familial recurrence data in a Nordic pooled sample that includes five countries with similar health, ascertainment, and diagnostic systems. This analysis attributes the remaining 17.3% of the variance to unique environmental factors (E). Despite uncertainty about the contribution of the (C) component, the study concludes that it plays a minimal role in the risk of ASD, in contrast to (E)[25].

Concerning the (A) component, the SNP heritability, which represents the proportion of variance attributed to the additive effects of common genetic variants in ASD, varies from 50% to 12% according to the latest analysis by the Psychiatric Genomics Consortium (PGC)[37,38]. The most recent estimate for the heritability of rare inherited variants (nonsense and splicing) was 3%[38]. Consequently, the currently unexplained additive heritability is calculated to be 67.7% (total narrow-sense heritability minus the combined heritability of common and rare

inherited variants) (Figure 3). Of note, this is likely to be an upper limit, as we have considered the markedly lower recent h^2 estimate by Grove and colleagues[37], and current estimates for the general contribution of rare inherited variation are outdated. Nevertheless, earlier findings consistently converge on approximately 50% of the heritability being explained by common variation[38], a recurrent discrepancy in neurodevelopmental disorders (NDD) termed "missing heritability", which will be further discussed in section "3.2.5 The "missing heritability" in ASD".

Conversely, the term "environmental" pertains to factors that are neither additive nor inherited but contribute to the variability in ASD. Notably, *de novo* variants serve as genetic causes of ASD but do not contribute to heritability since they are only present in the affected individual (except for the relatively unusual occurrence of germinal mosaicism found in one of the parental germlines and passed on to multiple offspring)[39]. These *de novo* events are thus classified as "environmental causes" of ASD, despite their impact on the DNA molecule.

All exome *de novo* variants, including PTVs and missense mutations in autosomes, explain 1.92% of the ASD variance[40], while non-coding variations (limited to promoter regions) contribute to 0.54%[41]. It is important to note that this estimate is likely a conservative lower-bound, as ongoing assessments of other regulatory regions in the non-coding genome are revealing their influence on ASD. Furthermore, these estimates are confined to autosomes and exclude other types of variations, such as CNVs or tandem repeats (TRs, see section "3.3.1.5.1 Tandem repeats (TRs)"). Additionally, non-additive variation accounts for approximately ~4% (ranging from 3% to 5%) of the total variance[42]. Consequently, considering these estimates, ~11% of the total (E) component of the variance in ASD remains unaccounted for (Figure 3).

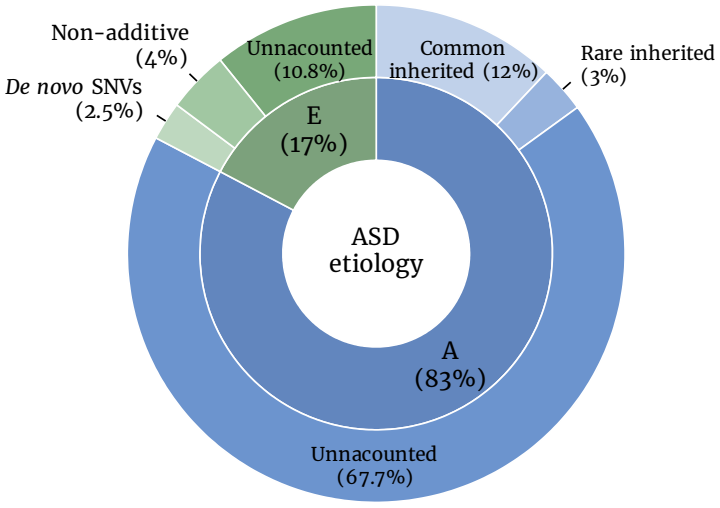


Figure 3. AE model for ASD variance. The genetic additive component (A) is represented in blue shades, in contrast to the unique environmental influence (E) in green. Of note, estimates are based on the most recent development in human genetics research in ASD, which might lead to highly conservative assessments for certain components (for further information, refer to the main text). This figure was adapted from Havdahl *et al.*[27] in R.

The upcoming sections will delve into a detailed discussion of genetic studies that have played a pivotal role in advancing our understanding of the genetic etiology of ASD.

3.2.1 Linkage and candidate gene studies

The initial studies on genetic factors in ASD involved genetic linkage studies in affected-relative pairs and multiplex families with the goal of identifying co-inherited chromosomal regions among affected individuals. Linkage studies involve the search for genetic markers (whose physical location is known) that co-segregate with the disease in families with several affected members[43].

The first genetic locus for ASD from such studies was published in 1998[44]. However, these and subsequent early studies faced challenges, including small sample sizes and scattered marker sets. As a result, just a few loci met genome-wide significance thresholds, and most of them were not consistently replicated in subsequent research.

Only two loci, one on chromosome 20p13[45] and one on chromosome 7q35[46] (providing strong statistical evidence for the involvement of common variation within *CNTNAP2*), have been successfully replicated.

The reported linkage regions, albeit of their statistical significance, serve as targets for rare/common variation screening. Candidate gene association studies leverage these regions as an alternative approach to identifying variants that may increase susceptibility to ASD. These studies involve the comparison of genetic variations, such as SNPs, between individuals with the condition (cases) and those without it (controls).

While candidate gene association studies investigated over 100 potential candidate genes for ASD associations, even the most frequently associated genes, such as *RELN* (reelin), have reported negative replications (as reviewed in Abrahams and Geschwind, 2008[47]).

The lack of consistent replication was attributed to phenotype's heterogeneity and limitations in study design including low statistical power, population diversity, and incomplete coverage of genetic variation within candidate genes.

Approaches such as the study of particular endophenotypes (*i.e.*, measurable traits that reflect underlying or component features of a disorder) have proven fruitful in revealing significant associations related to language delay across a 10 centimorgans (cM; a unit for measuring genetic linkage) region at 7q35[46]. This was achieved through the study of the association of quantitative trait loci (QTL), genetic regions that influence the phenotypic variation of a complex trait[48].

Furthermore, the inadequate statistical power is primarily attributed to sample sizes that are too small to produce findings that can be reliably replicated. Based on this assumption, numerous collaborative initiatives emerged so that sample sizes can be increased and individual data can be shared[49] (Table 4).

Table 4. Main ASD consortia.

Consortium	Description
Simons Foundation Powering Autism Research for Knowledge (SPARK)[50]	SPARK stands as the largest ASD cohort, currently including more than 100,000 individuals diagnosed with ASD and 175,000 of their family members.
Autism Genetics Resource Exchange (AGRE)[51,52]	The AGRE encompasses more than 2,000 multiplex families, with plans in place to expand the cohort further.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

MSSNG resource[53]	MSSNG provides WGS of over 11,500 individuals from families, including trios and multiplex, sourced from the AGRE repository and from other well phenotyped cohorts entering the study.
Autism Genome Project Consortium (AGP)[54]	The AGP's Phase 1, concluded in 2007, involved a sample exceeding 1,400 multiplex families. The ongoing Phase 2 involves the study of a new sample comprising over 3,000 trios.
Simons Simplex Collection (SSC)[55]	The SSC comprises 2,600 simplex families, but the active enrollment of participants concluded in 2011.
The Autism Simplex Collection[56]	This collection includes over 1,700 trios.

Moreover, the advent of genome-wide association studies (GWAS) and next-generation sequencing techniques (NGS) has significantly improved gene and variant discovery in ASD research, as discussed in the following sections.

3.2.2 Whole exome sequencing analysis (WES)

To identify additional ASD risk genes, researchers have employed whole-exome sequencing analysis (WES). WES is a widely used NGS method that involves sequencing the protein-coding regions of the genome, collectively known as the exome.

This approach aims to identify both *de novo* and inherited variants (mainly SNVs and indels), providing insights into the genetic factors contributing to ASD.

3.2.2.1 *De novo* variants (DNVs)

Recently, WES has confirmed the contribution of coding DNV, including indels and SNVs) in ASD. Main findings include:

1. Elevated rate of DNVs associated with advanced paternal age[57–61] (Pearson correlation = 0.4, $p = 4.2\times10^{-10}$)[61] with an increase of almost two mutations per year[62].
2. Notably, while the overall rate of genomic DNVs is roughly equivalent between affected and unaffected individuals (proband *versus* unaffected sibling, case *versus* control), the incidence of coding pathogenic mutations in crucial LoF (loss-of-function) intolerant genes (see next section) and gene pathways is significantly higher among ASD probands. This observation constitutes a critical feature of the disorder's etiology, as has been consistently replicated in the most relevant genetic studies focused on deciphering the coding genome in ASD (outlined in Table 5).
3. The association of hundreds of genes that are involved in the ASD etiology[63,64] (Table 5).

However, DNVs account for only a fraction of cases (contributing to ~5%-10% of ASD cases[58,65,66]) (see Figure 3), with estimates varying widely depending on the ascertainment strategy, and showing higher enrichment in females with ASD and those with ID[57,58,60,67–71].

3.2.2.2 Inherited variants

Initial advancements in gene discovery for ASD have been primarily drawn from investigations of simplex families[69,72], which are depleted for inherited risk compared to

USC UNIVERSIDADE DE SANTIAGO

multiplex families. Consequently, many ASD risk genes, particularly those associated with inherited risk, remain unidentified at first, raising important unanswered questions about whether rare *de novo* and inherited risk variants impact the same biological pathways.

Recent research focusing on families with two or more autistic children revealed a distinct genetic architecture in multiplex ASD families, highlighting a significant influence from rare, inherited variants:

1. Private, inherited SNVs that truncate proteins and indels are enriched in ASD probands (Table 5).
2. Rare inherited variations predominantly influence risk genes that were previously unidentified, but converge in protein-protein interaction (PPI) networks with known genes[73–76].

3.2.2.3 Gene discovery in WES

A crucial factor in gene association within WES studies has been the recurrence of DNVs, which are sufficiently rare to provide strong evidence for a gene's causal link to ASD. However, the complexity in interpreting the number of DNVs observed in a specific gene arises from variations in the background rate of DNVs, which can differ significantly between genes. As more individuals undergo sequencing, the likelihood of chance occurrences leading to multiple DNVs in the same gene increases, adding complexity to the determination of whether these variants are genuinely associated with a disorder or occur randomly[77].

To address this challenge, scientists utilize statistical techniques to evaluate whether the observed quantity of DNVs in a gene surpasses the expected number based on the background mutation rate. Some approaches followed in ASD WES studies in order to tackle this difficulty include the following (Table 5):

1. Prioritize PTVs, also known as likely gene disrupting (LGD) mutations that elevate the risk of ASD[57–59,69].
2. Include predictors of variant severity such as (i) the "missense badness, PolyPhen-2, and constraint" (MPC) score[69,78], which serves as a measure estimating the deleterious effects of missense variations; (ii) the Grantham matrix score[57,79], a prediction of the effect of substitutions between amino acids based on chemical properties, including polarity and molecular volume, characterized into classes of increasing chemical dissimilarity; or (iii) the genomic evolutionary rate profiling (GERP), which measures the degree of nucleotide level evolutionary conservation[80].
3. Focusing on rare variants, *i.e.*, those not found/having low frequencies in databases of genetic human variation[81], such as dbGaP (database of genotypes and phenotypes) and gnomAD (genome aggregation database). This is grounded on the premise that for severe, early onset NDDs, such as ASD, natural selection restricts the population frequency of variants, ensuring that those with larger effect sizes are exceptionally rare[49] (Figure 4).
4. Enhance the discovery of risk genes by incorporating the functional impacts of rare variation and metrics that assess negative selection for each gene. One such measure is

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

the 'loss-of-function observed/expected upper bound fraction' (LOEUF) score[82], a continuous measure indicating the selective pressure against PTVs in each gene. The residual variation intolerance score (RVIS)[72,83], or the pLI (for LoF mutations, represents an estimate of the "probability of being loss of function intolerant") and the z-score (for missense mutations), which are two scores collected in gnomAD, measure how constrained or intolerant a gene or transcript is to a specific type of variation[84].

5. Use of weighted, statistical models such as TADA (transmission and d*e novo* association test)[64] among others, to integrate functional annotations and differences in relative mutation rates between genes into joint analyses of *de novo* and transmitted coding variation, so as to prioritize candidate genes.
6. Rank genes based on their recurrence (increased presence of prioritized variants restricted to specific genes) or their presence in PPI networks (along with genes already associated with ASD).

Table 5. Key supporting evidence for the contribution of *de novo*/inherited variants (SNVs and indels) to the genetic etiology of ASD.

Study	Cohort description	Metho-dology	Relevant key findings	Candidate ASD genes
O Roak *et al.*, 2011[57]	20 ASD trios	WES	The rate of coding DNMs in ASD probands is slightly elevated compared to what is expected in the general population and corresponds to more highly conserved (GERP, $p<0.001$) and disruptive amino acid mutations (Grantham score, $p = 0.015$).	*FOXP1, GRIN2B, SCN1A* and *LAMC3.*
Sanders *et al.*, 2012[59]	343 ASD quads	WES	The frequency of nonsynonymous *de novo* SNVs is higher in affected *versus* unaffected siblings, with a significant increase in the proportion of gene-disrupting compared to silent mutations among probands in contrast to their siblings (OR = 1.93).	*SCN2A* is identified as a candidate gene by the occurrence of two or more *de novo* nonsense and/or splice site mutations "which is highly improbable by chance ($p = 0.008$)"[59].
Neale *et al.*, 2012[60]	175 ASD trios	WES	Genes carrying *de novo* missense/nonsense mutations show increased connectivity among the encoded proteins and exhibit excess connections to previously identified genes linked to ASD.	*CHD8* and *KATNAL2.*
He *et al.*, 2013[64]	~ 1000 ASD trios and 2000 case-control	-	*De novo* LoF mutations and "probably damaging" missense mutations show enrichment in ASD probands.	An estimation suggests the presence of approximately 1,000 ASD risk genes.
Jiang *et al.*, 2013[85]	32 ASD trios	WGS	*De novo* and rare inherited SNVs are detected in 19% of examined probands, affecting both known and novel ASD risk genes.	*ARID5A, DNMT3A, CACNA1C, CHD7,*

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

			The number of *de novo* mutations exhibits a significant correlation with paternal age (p<0.005).	*SCN2A*, *NRXN1*, among others.
De Rubeis *et al.*, 2014[69]	3,871 ASD cases 9,937 ancestry-matched or parental controls	WES	Individuals with ASD exhibit a notable surplus of *de novo* LoF mutations beyond the expected levels. Risk variation damages three crucial pathways for typical development: chromatin remodeling, transcription and splicing, and synaptic function.	TADA analysis implicates 22 autosomal genes (FDR <0.05), and a set of 107 autosomal genes (FDR <0.30) which show unusual evolutionary constraint.
Dong *et al.*, 2014[86]	787 ASD families (602 quads, 185 trios)	WES	ASD risk is influenced by *de novo* frameshift indels (OR = 1.6). These indels are more prevalent in female probands (p = 0.02), enriched in genes encoding *FMRP* targets (p < 0.001), and primarily originate from the paternal chromosome (p < 0.001).	*KMT2E* (*MLL5*) and *RIMS1*.
Iossifov *et al.*, 2014[58]	2517 families (2508 ASD probands, 1911 control siblings)	WES	LGD DNMs occur at significantly higher rates in affected *versus* unaffected siblings. LGD targets in affected females significantly overlap the targets in males of lower IQ. LGD targets are enriched for chromatin modifiers, *FMRP*-associated genes and embryonically expressed genes.	27 recurrent target genes (*CHD8*, *ARID1B*, *SCN2A*, and *KMT2E*, among others).
O Roak *et al.*, 2014[67]	3,486 ASD probands, 2,493 control siblings	WES	The resequencing of 64 candidate NDD risk genes yielded an elevated rate of PTVs (2.7-fold increase) among affected siblings relative to the unaffected ones. A model of recurrent mutation simulation identified novel genes in ASD risk, significantly more mutated in the probands. Mutation carriers generally have lower IQ.	*ADNP*, *CHD2*, *CHD8*, *GRIN2B*, *PAX5*, *PTEN*, *SYNGAP1*, *TBR1*, and *TRIP12*.
Samocha *et al.*, 2014[77]	1,078 ASD trios	WES	ASD cases show a significant excess of *de novo* LoF mutations ($p = 2.05 \times 10^{-7}$). No excess of *de novo* LoF mutations is observed in cases with high IQ.	*DYRK1A* and *SCN2A*.
Krumm *et al.*, 2015[72]	2377 families (2377 ASD probands, 1786 control siblings)	WES	A notable enrichment for transmission in probands is observed, particularly when focusing on LGD mutations in genes characterized by the lowest RVIS values (OR = 1.14, p = 0.0002). This enrichment exhibits a significant bias toward maternal transmission to sons.	By assessing SNVs and CNVs together, the three highest-highest ranked genes are *RIMS1*, *CUL7* and *CSMD1*.
Yuen *et al.*, 2016[61]	200 ASD trios	WGS	DNMs predominantly arise from the mother ($p = 7.7\times10^{-13}$), and their occurrence significantly rises with advancing paternal age only ($p = 4.2\times10^{-10}$). There is a notable	*DNMT3A* and *ADNP*.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

			enrichment of predicted damaging DNMs in ASD cases ($p = 8.0\times10^{-9}$; OR = 1.84).	
Yuen *et al.*, 2017[84]	2,066 multiplex ASD families	WGS	Candidate ASD-risk genes were selected based on intolerance to mutation (pLI and z-scores) and a mutation rate higher than expected (FDR < 15%) (Samocha et al 2014[77]). Participants with mutations in susceptibility genes demonstrate significantly lower adaptive ability ($p = 6 \times 10^{-4}$).	A total of 54 potential ASD-risk genes are identified, and an additional 7 are found on the X chromosome.
Feliciano *et al.*, 2019[87]	457 ASD simplex/ multiplex families	WES	*De novo* LGD variants in affected children is 1.76-fold higher than the baseline expectation[77].	Using the TADA framework, 26 ASD risk genes are identified at an FDR < 0.1.
Ruzzo *et al.*, 2019[75]	493 multiplex ASD families	WGS	Multiplex families exhibit a distinct genetic architecture, unveiling new ASD risk genes. Rare inherited variations impact specific biological processes like cytoskeletal organization and ion transport, distinct from those influenced by rare *de novo* variations. However, both *de novo* and inherited genes contribute to a shared PPI network. Excess of PTVs transmitted to all affected children but not to unaffected ones in constrained genes.	98 genes with high-risk inherited variants are identified.
Satterstrom *et al.*, 2020[40]	35,584 total samples, 11,986 with ASD (family-based, case control)	WES	Using an updated TADA model (He *et al.*[64]), which incorporates the pLI score as a continuous metric for PTVs and the MPC score as a two-tiered metric (≥ 2, 1–2) for missense variants, 30 novel risk genes are discovered.	The involvement of 102 genes in the risk for ASD is identified (FDR < 0.1).
Wilfert *et al.*, 2021[76]	3,474 simplex/ multiplex families	WGS	Candidate genes carrying inherited private LGD variants, which are preferentially transmitted to affected probands, show a convergence on the E3 ubiquitin–protein ligase complex (an established regulatory hub in ASD[88]), intracellular transport and Erb signaling protein networks.	*MED13*, *CPA6*, *FRA10AC1*, *MPHOSPH10*, and *RALGAPB*.
Fu *et al.*, 2022[74]	63,237 individuals (20,627 affected probands from family data and case-control cohorts)	WES	There is an enrichment of PTVs and damaging missense variants in individuals with ASD compared to unaffected individuals. The enrichment of PTVs is most pronounced in genes under selective constraint. Additionally, deleterious missense variants show a strong enrichment in ASD cases.	373 genes are associated with ASD at FDR ≤ 0.001.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Zhou *et al.*, 2022[73]	42,607 ASD cases (16,877 ASD trios)	WES+ WGS	Autistic individuals carrying *de novo*/rare inherited LoF variants in moderate-risk genes exhibit less cognitive impairment compared to those with LoF variants in high risk, well-established ASD genes.	A total of 60 genes are identified with exome-wide significance ($p<2.5\times10^{-6}$), among which five are newly discovered risk genes (*NAV3*, *ITSN1*, *MARK2*, *SCAF1*, and *HNRNPUL2*).
Cirnigliaro *et al.*, 2023[89]	1,004 families (1,836 autistic children and 418 non autistic children)	WGS	Seven previously unrecognized ASD risk genes are identified, and notably, five of them are affected by a significantly higher proportion of rare inherited PTVs compared to rare *de novo* variants.	74 genes significantly are associated with ASD at FDR <0.1.

FDR, false discovery rate; IQ, intelligence quotient; OR, odds ratio; WGS, Whole Genome Sequencing.

3.2.3 Copy number variant (CNV) analysis

CNVs refer to large deletions or duplications often involving several genes. In 2007, comparative genomic hybridization (CGH; a technique that permits the detection of chromosomal copy number changes[90]) was used to establish a significant association between *de novo* submicroscopic structural variation and ASD[91]. From that point onward, microarray-based studies have firmly established a clear etiological role for large, rare CNVs in ASD[71,92–94] (Table 6).

Advancements in technology now allow for the detection of CNVs through whole-genome sequencing (WGS), a DNA sequencing approach that assesses the entire genome, though this method has yet only been applied to relatively small cohorts of individuals with ASD[66,95,96]. However, relevant findings have been obtained, including:

1. Well-established presence of an excess of *de novo* CNVs in affected individuals (when compared to non-affected siblings)[72,91,92,94,97].
2. Recurrent CNVs constitute the primary established form of inherited risk variation for ASD[75,98].
3. The risk associated with deletion CNVs, as indicated by the odds ratio (OR), is comparable to that from LoF SNVs in cases *versus* controls or transmitted LoF variants from parents to offspring[69].
4. ASD probands exhibit a larger CNV average size and number of harbored genes compared to unaffected siblings[71,92].
5. Rare CNVs, including *de novo* events, were significantly more frequent in simplex *versus* multiplex families, indicating a higher burden of *de novo* variation in the genetic etiology of sporadic ASD[93,98].

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

6. Similar to observations of SNV mutations in conserved genes, genes within CNV intervals in probands have a lower average RVIS than those in siblings[72].

Table 6. Known CNVs in ASD curated from SFARI Gene.

CNV locus	CNV type	Spanning range (GRCh37/hg19)	Associated/candidate genes
1q21.1	Deletion-Duplication	chr1:145900678-147965543	-
2p16.3	Deletion-Duplication	chr2:47600165-53040270	*FBXO11, NRXN1*
2q11.2	Deletion-Duplication	chr2:97739057-98115695	-
3q29	Deletion-Duplication	chr3:195676676-197366632	*DLG1, PAK2, TM4SF19*
5q35	Duplication	chr5:178554060-179589550	-
7q11.23	Deletion-Duplication	chr7:72311894-74129587	*CLIP2, GTF2I, STX1A*
8p23.1	Deletion-Duplication	chr8:8123460-11384691	-
15q13.3	Deletion-Duplication	chr15:30938215-32914140	*ARHGAP11A, CHRNA7, FAN1, OTUD7A, TRPM1*
16p11.2	Deletion-Duplication	chr16:29692499-30792499	*CORO1A, KCTD13, MAPK3, SEZ6L2, SRCAP*
16p12.2	Deletion-Duplication	chr16:21356420-21577433	-
16p13.11	Deletion-Duplication	chr16:14972499-16522499	-
16p13.3	Deletion-Duplication	chr16:3392370-5752860	*CREBBP*
17p11.2	Deletion-Duplication	chr17:16532736-20464365	*RAI1*
17q11.2	Deletion-Duplication	chr17:29015932-29149664	-
17q12	Deletion-Duplication	chr17:37228545-39077997	*CACNB1, KRT26, NR1D1, THRA*
22q11.2	Duplication	chr22:21031117-21651381	*LZTR1*
22q13.3	Deletion	chr22:41122568-49565875	*EP300, TCF20, XRCC6*

Adapted from Zhang *et al.*[99].

3.2.4 Genome-Wide Association Studies (GWAS)

GWAS are research approaches used to uncover links between genetic variations and observable traits by examining variations in allele frequencies among individuals who share similar ancestry but exhibit phenotypic differences. In a GWAS, a population with the trait of interest (cases) is genotyped and compared to a population without it (controls). This comparison, often focusing on thousands or millions of SNPs, seeks to identify statistically significant changes in their allelic frequencies between cases and controls.

When the frequency of a SNP is higher in cases compared to controls, they are typically used to search for nearby variants in linkage disequilibrium (LD) regions that contribute directly to the disease or trait (genomic risk loci)[100]. LD is a property where certain alleles of genetic markers are inherited together. The existence of LD leads to the organization of the human genome into LD blocks that are characteristic of each human population. This structure was elucidated through the development of the HapMap project[101], where thousands of SNPs were genotyped in individuals from different populations. Thanks to this property, it is sufficient to select a SNP (tag SNP) with high LD with the other SNPs present in the haplotype block to infer the genetic variation of the region through a process called imputation, without the need to genotype all SNPs in a block. It is estimated that using this strategy, it is possible to detect

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

over 80% of the common variation present in the human genome using genotyping arrays containing between 500,000 and 1,000,000 representative SNPs.

In order to assess millions of associations between individual genetic variants and a phenotype, it's crucial to apply a stringent multiple-testing threshold to minimize false positives. Studies, including the International HapMap Project, indicate an average of approximately one million independent common genetic variants across the human genome[101]. This leads to a Bonferroni testing threshold of $p < 5 \times 10^{-8}$ (equivalent to a false discovery rate of $0.05/10^6$).

Up to this point, we have observed that rare variants distinctly contribute to ASD, with the role of *de novo* variation being particularly evident and easily characterized. Additionally, inherited variation plays a role in the overall liability for ASD. However, the contribution of inherited common variation to the heritability of ASD has been less substantiated[38]. Initially, it was believed that common polygenic variation, distributed across the genome, accounted for at least 20% (and up to 50%) of ASD liability[65], but recent estimates from the largest and latest GWAS have indicated a contribution of 12%[37].

Despite a few conducted GWAS, notable findings have been scarce[38,102–104], and the majority were not replicated in subsequent studies, once again, due to the disorder's phenotypic heterogeneity and the lack of sufficient sample sizes to reach the stringent statistical threshold of genome-wide significance[45,105,106]. However, this is to be expected with high genetic heterogeneity, where the contribution of many low effect common variants act additively through polygenic risk (Figure 4)[89,107].

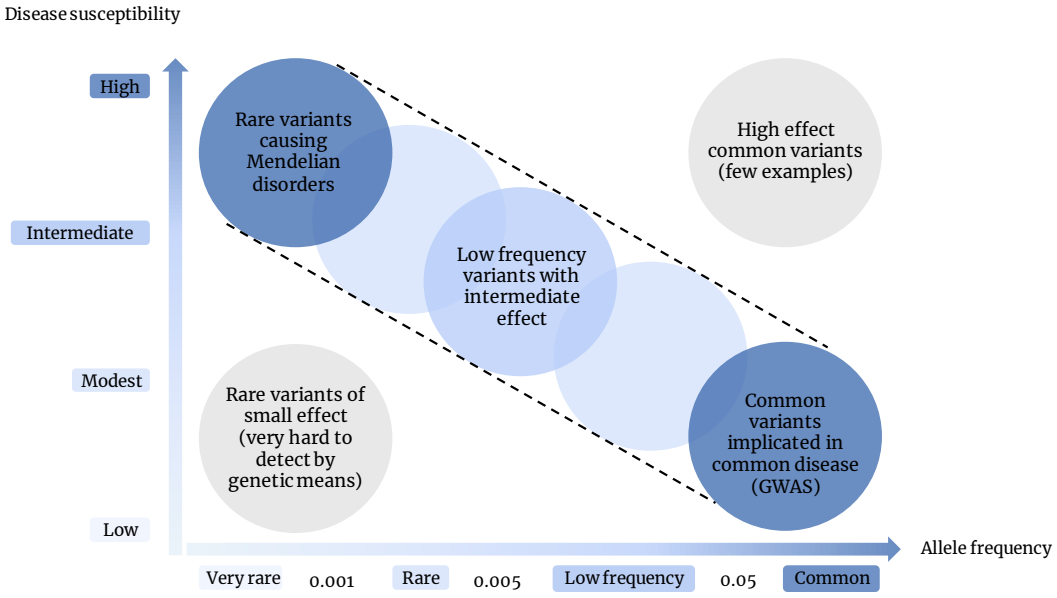


Figure 4. Frequency of genetic variants and disease susceptibility. From an evolutionary perspective, common variation undergoes significant selection pressure, reducing its likelihood of exerting substantial negative effects on early survival or reproductive fitness. This supports the hypothesis known as the "common disease, common variant", suggesting that prevalent diseases can be partially attributed to commonly occurring genetic variations identified through GWAS. On the contrary, rare genetic variations, including *de novo* mutations, may have larger biological impacts. The hypothesis implies that transmitted variations posing a substantial risk are kept at low allele frequencies due to purifying (negative) selection mechanisms. Thus, *de novo* mutations, which are typically infrequent or unique, as they have not had time to spread widely in the population, may play a much more relevant

role into disease susceptibility. Similarly, mendelian disorders, whether dominant or recessive, result from rare mutations with high penetrance (*i.e.*, the proportion of individuals with the mutation who are also either diagnosed with the disease or exhibit a specific phenotype), aligning with the observed low frequency of these conditions in most populations. Only a few examples of common variants of large effect exist, and the contribution of rare variants of small effects is hard to detect by genetic means. Therefore, most of the interest lies in uncovering associations with traits represented by the characteristics delineated by diagonal dotted lines[32,108]. This figure was re-created from Manolio *et al*.[32].

Even with substantial sized cohorts (greater than 18,000 cases and 28,000 controls), GWAS in ASD have to date only identified three loci reaching genome-wide significance. When analyzing the top 88 loci (p-values $<1\times10^{-5}$) in five cohorts of European ancestry (totaling ≈2,000 additional cases), two additional loci achieved genome-wide significance in the combined analysis[37] (Table 7).

Table 7. Genome-wide significant loci from the largest ASD GWAS.

Index variant	Genomic location (GRCh37/hg19)	Analysis	p-value	Nearest genes
rs910805	chr20:21248116:A-G	ASD	$2.04\text{x}10^{-9}$	*KIZ*, *XRN2*, *NKX2–2*, *NKX2–4*
rs10099100	chr8:10576775:C-G	ASD	$1.07\text{x}10^{-8}$	*C8orf74*, *SOX7*, *PINX1*
rs201910565	chr1:96561801:A-AT	Follow-up	$2.48\text{x}10^{-8}$	*LOC102723661*, *PTBP2*
rs71190156	chr20:14836243:GTTTTTTT-G	ASD	$2.74\text{x}10^{-8}$	*MACROD2*
rs111931861	chr7:104744219:A-G	Follow-up	$3.53\text{x}10^{-8}$	*KMT2E*, *SRPK2*

The column "Nearest genes" lists nearest genes from within 50 kb of the region spanned by all SNPs with $r^2 \geq 0.6$ to the index variant. The column "Analysis" indicates whether the index variant yielded significance in the main ASD scan or in the follow-up analysis.

These results parallel those of early NDDs GWAS studies, which, upon reflection, were underpowered. However, later studies with tens of thousands of participants did yield numerous associations. Of note, Grove and colleagues established a strong genetic correlation with other NDDs such as schizophrenia (SZP), ADHD and major depression (MD). The observed differences in allelic architecture among these traits do not seem to have a clear genetic basis. Some apparent distinctions may simply stem from variations in the stage of investigation across traits, and it is reasonable to anticipate an increase in the number of detected variants with larger sample sizes, as further elaborated in Chapter 1[37].

3.2.4.1 Gene discovery in GWAS

In coding regions, mutations associated with ASD are limited to specific types of variants including PTVs, probably damaging missense variants, and evolutionarily conserved nucleotides (refer to section 3.2.2.3 for details).

However, the overwhelming majority of genetic variants associated with disease risk in GWAS are situated in non-coding regions of the genome[109]. As there is currently no equivalent definition for critical non-coding loci, and predicting the regulatory gene target using proximity to the non-coding variant has shown less than 40% accuracy[110], several approaches can be explored to leverage GWAS signals. These approaches aim to annotate variants and assign

USC DE SANTIAGO DE COMPOSTELA

candidate genes, as detailed in the subsequent sections. Of note, eQTL (expression quantitative trait loci) colocalization analysis, yet another method to mine GWAS data, will be discussed in Chapter 1.

3.2.4.1.1 Gene-based association analysis (GBA)

Gene-based association analysis (GBA) enables the examination of the joint association of all markers within a locus. Hence, it enables the integration of all genetic information across the gene, culminating in a single statistic (*i.e.*, a p-value for each gene). The usual definition of a gene in a GBA involves a default window of approximately ± 50 kb around the 5′ and 3′ untranslated regions (UTRs). However, certain algorithms may additionally incorporate LD information from the genomic region to delineate gene boundaries.

Grove *et al.*[37] conducted a GBA by using MAGMA, the most commonly used GBA tool[111]. In their analysis, 15 genes exceeded the significance threshold. As anticipated, the majority of these genes were situated within the genome-wide significant loci identified in the GWAS (Table 7). Notably, 7 genes were located in four additional loci, including *KCNN2*, *MMP12*, *NTM*, and a cluster of genes on chromosome 17 (*KANSL1*, *WNT3*, *MAPT*, and *CRHR1*).

Recently, another study further leveraged summary statistics from Grove *et al.*[37] by using a novel GBA tool (PASCAL (pathway scoring algorithm))[112]. PASCAL allowed for the identification of the following loci associated with ASD: *XRN2*, *NKX2-4*, *KIZ*, *KCNN2*, *NKX2-2*, *CRHR1-IT1*, *C8orf74*, and *LOC644172*. Although some of these genes were previously reported by MAGMA (*XRN2*, *KIZ*, and *KCNN2*), the use of a different GBA tool, that differs in the calculation of the gene statistic and the construction of the LD matrix, has proven valuable in bringing attention to additional genes.

3.2.4.1.2 Transcriptome wide analysis (TWAS)

TWAS (transcriptome wide association studies) which integrate tissue expression and genetic data (namely, GWAS summary statistics) have emerged as valuable tools for investigating the genetic architecture of complex traits[113].

UTMOST (unified test for molecular signatures)[114] has recently been introduced as a novel framework for both single and cross-tissue expression imputation and gene-level association analysis. UTMOST has the capability to collectively consider the impact of SNPs from summary statistics (GWAS) across LD regions (utilizing 1000 Genomes data) and integrate tissue expression data from GTEx[115]. This integration results in the creation of single and cross-tissue covariance matrices, aiding in the delineation of gene-trait associations.

Utilizing UTMOST, Rodríguez Fontenla and colleagues[116] have established the association of *CIPC*, *PINX1*, *NKX2-2*, and *PTPRE* at the brain level. Additionally, they have unveiled the significance of GI tissue in ASD etiology by identifying the association of other genes such as *NKX2-2*, *MANBA*, *ERI1*, and *MITF*, which is consistent with the involvement of GI symptoms in ASD cases. This result reveals the importance of considering expression data (RNA sequencing (RNA-Seq), a method for transcriptome profiling that involves utilizing deep-sequencing technologies to identify and precisely measure RNA molecules at a specific point in time) from different tissues (not only brain) to study the genetic and etiological basis of ASD.

3.2.5 The "missing heritability" in ASD

As explained earlier in this section, the heritability estimates derived from classical twin/family studies lack direct corroboration from the current genetic analysis data. These estimates consistently exhibit lower values, thereby giving rise to what is commonly referred to as the "heritability gap" or "missing heritability" (*i.e.*, the incomplete explanation of the genetic factors that contribute to the development of ASD)[32,117]. When heritability models encompass all common SNPs, the estimates of h^2_{SNP} range from 12% to 50%. This is in pronounced contrast to the conclusions drawn from twin and family studies, where heritability for ASD constitutes 80% of the disease risk.

Several explanations have been proposed to account for this phenomenon of missing heritability. These include:

1. Study designs and ascertainment. As an example, an important aspect to consider is that the vast majority of GWAS and other genetic investigations have primarily focused on populations of European ancestry. However, it is noteworthy that the highest levels of genetic variation are found in populations with recent African ancestry. Studies conducted in non-European populations have proven to be particularly insightful, revealing novel and intriguing genetic variants[32].

2. The existence of rarer variants, potentially with more substantial effects, that are inadequately captured by currently available genotyping arrays. A reduction in reproductive fitness (as in the case of ASD) would generally be expected to diminish the prevalence of high-risk variants. Variants of this nature lack the necessary frequency for detection by GWAS genotyping arrays, and their effect sizes are not substantial enough to be identified through traditional linkage analyses in family studies. When the MAF drops below 0.5%, the likelihood of detecting associations diminishes, unless the effect sizes are exceptionally large, as observed in Mendelian disorders (Figure 4). This could explain the relative lack of variants detected so far in GWAS analysis, which predominantly focus on variants present in 5% or more of the population.

3. The existence of pleiotropic effects of multiple genes, where a single variant can influence multiple disease risks simultaneously. Focusing solely on the reproductive impacts of a singular disease overlooks the pleiotropic effects. For instance, rare CNVs and SNVs associated with ASD exhibit pleiotropic effects, elevating the susceptibility not only for ASD but also for other complex disorders such as SZP, ADHD, ID, and epilepsy[40].

4. The possibility that larger numbers of variants, each exerting a smaller effect, remain undiscovered.

5. Poor capture of SVs by current genotyping arrays.

6. Low power to detect gene-gene/ gene-environment interactions.

7. Incomplete inclusion of shared environment factors in the analysis.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

8. The possibility that the contribution of *de novo* variants to disease liability is lower than their overall frequency in cases due to incomplete penetrance[118].
9. Incomplete exploration of potential causal variants within each gene in current investigations. The relative risks identified for marker SNPs may underestimate the actual risks associated with the true causal variants. Significantly, *KMT2E* emerged as implicated in both recent GWAS[37] and exome sequencing analyses[40] suggesting that genes housing common variants with modest effects on complex traits may also harbor rare variants with more substantial effects.

To address these limitations, several studies aimed towards reducing phenotypic heterogeneity in an attempt to "increase genetic homogeneity". This approach acknowledges diagnostic classification as an additional source of heterogeneity and mirrors the division of endophenotypes in linkage studies. As highlighted in the work by Grove *et al.*, the SNP-heritability of Asperger's syndrome was observed to be twice that of childhood autism and unspecified pervasive developmental disorders. Notably, the SNP-heritability of ASD without a co-occurring ID diagnosis, was found to be three times higher than that of autism with ID[37].

Additionally, stratifying individuals could identify genes exhibiting a high mutation prevalence within specific subgroups of individuals with ASD, such as *CHD8*, which is recurrently mutated in individuals with macrocephaly[119].

Nowadays, much of the speculation surrounding missing heritability is focused on the non-coding genome, primarily due to the prevalence of alterations identified in GWAS within non-coding regions of DNA. The non-coding genome constitutes a fraction nearly 99 times larger than protein-coding genes within our genetic material[35]. Despite its substantial size, research on the pathophysiological effects of the non-coding genome is still in its early stages. This makes it challenging to definitively identify its potential pathogenic role in diseases and the specific genes implicated, a topic that will be further explored in the upcoming section.

3.3 THE NON-CODING GENOME IN ASD

While ASD have been proven to have a strong genetic component, the vast majority of the last decade's research largely focuses on coding regions of the genome, as reviewed in the previous section. Although promising, known genetic contributions including (but not limited to) CNVs, *de novo* SNVs, and common or rare inherited variants, do not explain a substantial portion, so that ~80% of the variance remains challenging to understand[27] (Figure 3) and only a small fraction of ASD patients (<30%) have been associated with specific genetic variants[99].

Therefore, the significant genetic component to ASD suggested by family and twin studies has not been consistently supported by genetic analyses. This has resulted in a discrepancy, referred to as "missing heritability", in our understanding of the genetic basis of ASD, as discussed in the preceding section.

Non-coding DNA, constituting approximately 99% of the human genome, has been thoroughly recognized to play a crucial role in NDDs, finely tuning biological pathways relevant for appropriate brain development[120,121]. In our context, it has been identified as a potential factor contributing to the missing heritability in ASD.

Novel sequencing technologies have been a milestone in opening new directions for the study of the gene regulatory networks embedded within the non-coding regions. Also, progress is being achieved by virtue of projects aiming to characterize all the functional elements present in the human genome, such as the encyclopedia of DNA Elements (ENCODE; see Chapter 3). Indeed, the most recent WGS analysis has revealed numerous ASD-associated variants that provide new insights into the genomic architecture of ASD. However, the enrichment of rare variants affecting non-coding elements in individuals with ASD is limited, underscoring the complexities involved in identifying consistent and replicable signals within a genetically heterogeneous disorder like ASD as well as the urgent need to gain further knowledge about the contribution of non-coding elements to disease risk [122].

The characteristics of genetic variants lying within non-coding regions (Figure 5) are mainly the following: (i) they would affect gene regulation if located within 5′ and 3′UTRs, cis-regulatory elements (promoters, enhancers, insulators) and/or ncRNAs (non-coding RNAs), (ii) they can only be detected using WGS in combination with specific algorithms, (iii) they have been overlooked in previous studies (WES) that do not cover the non-coding genome, (iv) they constitute part of the "missing heritability" in ASD that remains to be discovered. Additionally, regulatory landscapes are structured within self-interacting genomic regions known as topologically associating domains (TADs). This implies that DNA sequences located within a TAD exhibit more frequent physical interactions with each other compared to sequences outside the boundaries of the TAD[123].

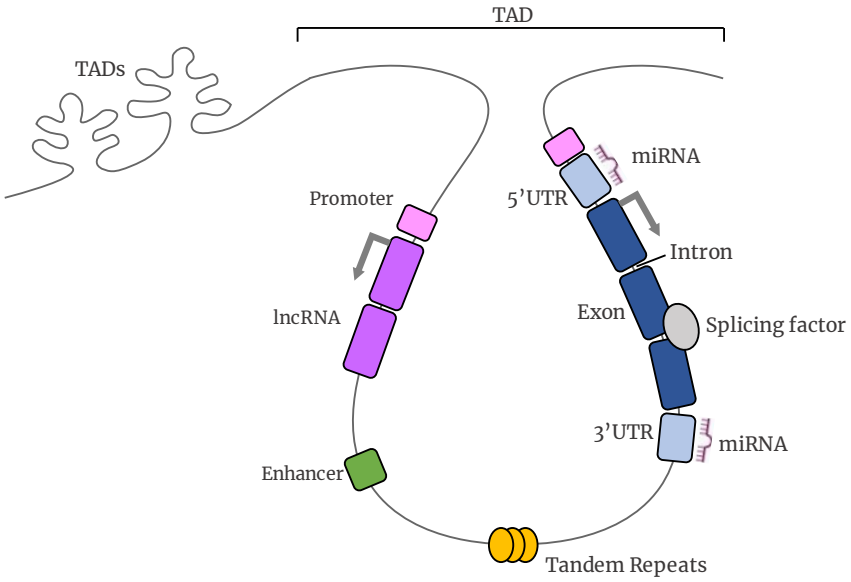


Figure 5. Non-coding regulatory elements affecting gene expression. Three-dimensional conformation of the genome plays a fundamental role in gene regulation: DNA is organized into TADs, self-interacting genomic regions. Within TADs, enhancers may form chromatin loops with proximal or distal promoters through the action of transcription factors. Transcription factors also mediate TR's gene expression regulation. Within a specific gene, variants altering 5' and 3' UTRs, promoters and enhancers also play a role in gene regulation, as well as long non-coding RNAs (lncRNAs) and microRNAs (miRNAs), via different mechanisms. Lastly, splicing factors determine different gene-isoform levels through alternative splicing, a process further discussed in this chapter.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Post-transcriptional mechanisms, such as ncRNAs, that widely influence gene expression levels without altering the underlying DNA sequence, constitute one means of altering entire gene networks[124]. Moreover, the untranslated gene regions, including the 5′ and 3’ UTRs and introns, are the major regulators of gene expression[125]. There are over 1 million distinct TRs that can affect the structure and function of DNA, RNA and proteins, affecting a wide range of molecular and cellular processes[126]. However, the extent by which variants in non-coding elements contribute to ASD risk is poorly understood. This is partly because, to date, only a few WGS studies have been conducted (see Table 8) and techniques for annotating the function of variants in non-coding regions need to further enhancement. The following sections will provide an overview of the current progress and challenges in unraveling the non-coding architecture of ASD.

3.3.1 Biased studies

3.3.1.1 Genetic variants in 5' and 3' untranslated regions (UTRs)

The 5' and 3' UTRs are non-coding regions of DNA that are located at the beginning (5') and end (3') of a protein-coding gene. These regions are transcribed into mRNA but are not translated into protein. The 5' and 3' UTRs harbor several regulatory elements. These include upstream open reading frames, which can function as translational repressors by binding to ribosomes and initiating translation before the main coding sequence. Additionally, there are internal ribosome entry sites that can facilitate cap-independent translation initiation. The secondary structure, sequence composition, and length of the UTRs may also act as regulatory elements and can affect the stability and localization of the mRNA. These elements are recognized by RBPs (RNA-binding proteins) or ncRNAS (mainly microRNAs (miRNAs)) in a sequence- or structure-dependent manner, playing a major role in post-transcriptional gene regulation. This includes regulation of mRNA stability, localization, and secondary structure, as well as translational initiation[127,128].

Because of their relevance on these regulatory mechanisms, variants in UTRs can modify molecular pathways and cellular processes, potentially affecting disease onset. In fact, approximately 3.7% of disease-associated variants in the NHGRI-EBI GWAS catalog localize to the UTRs of protein-coding genes[127]. 5' and/or 3' UTR dysregulation has been linked to ASD in several recent studies[129–132] (Figure 6).

For example, SNPs overlapping the 3' UTR of the oxytocin receptor gene (*OXTR*) may affect the stability of the mRNA due to changes in transcription factor binding. Multiple lines of evidence suggest that *OXTR* variants are associated with ASD owing to the role of oxytocin in modulating social disruptions characteristic of ASD (*i.e.*, regulation of affiliative behavior and social bonding) through the regulation of the underlying neural circuitry[129–131].

Additionally, Turner and colleagues[133] found an excess of non-coding *de novo* SNVs and indels in ASD patients, positively correlated with paternal age, when focusing their analysis on UTRs, promoters and enhancers.

Moreover, alternative cleavage and polyadenylation (APA), a process which affects 3'UTR length influencing the presence or absence of regulatory elements, can be used to control gene expression during neuronal differentiation. Through single cell 3'end sequencing in neural stem

cells Gopferich *et al.*[132] detected altered APA in ASD patients, with longer 3'UTRs resulting in a decreased number of neural stem cells.

Altogether, these data provide evidence that variants in UTRs can impact mRNA and/ or protein levels through various mechanisms, thus becoming an important risk factor. However, regulation of protein abundance is not the only function of UTR regulation. Future work should focus on how UTR regulation may lead to changes in mRNA's subcellular localization, probably affecting protein function and not just abundance.

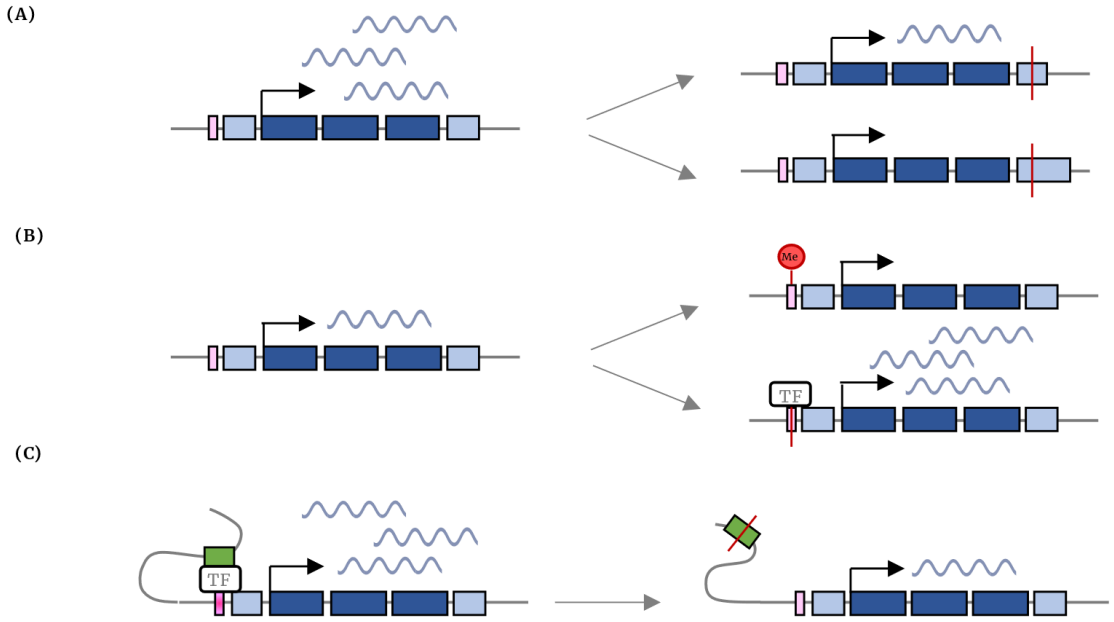


Figure 6. Examples of non-coding ASD-associated variants affecting proper functioning of UTRs, promoters, and enhancers. Representation of a gene divided into exons (dark blue rectangles). UTRs are represented as light blue rectangles, the promoters are represented in pink, and the enhancers are represented in green. mRNA levels are represented as undulating lines. Risk variants are represented as red lines. (A) UTRs. (Top-right) Variants in the 3'UTR may give rise to a decrease in mRNA stability, with lower mRNA levels and thereby protein abundance[130,131,133]. (Bottom-right) Alterations in 3' UTR length, prompted by a variant leading to APA, contribute to modifications in gene expression. These changes involve the introduction of new regulatory element binding sites or the absence of existing ones[132]. (B) Promoters. Risk variants involving epigenetic changes, such as increased levels of methylation[134] (top-right), or gain of transcription factor's binding sites[135] (bottom-right), provoke alterations in gene expression. (C) Cis-regulatory elements. Variants harbored in gene enhancers may reduce/ablate their interaction with target gene promoters, reducing mRNA expression[96,136,137].

3.3.1.2 Promoters

Core promoters are short sequences that can extend ±50 bp (base pairs) upstream and downstream of the transcription start site (TSS), serving as a binding platform for the RNA Pol II and general transcription factors to initiate transcription[138]. Regarding ASD, several studies revealed association with promoter variation[41,134,135], although further investigation is needed in order to explain the specific mechanisms by which gene expression is being altered (Figure 6).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Shorter repeats (microsatellites; usually <10 base pairs repeat units) and DNMs located in promoter regions have been described as associated with changes in the expression of certain genes that are believed to play a role in the development of the autistic phenotype. As an example, Tansey *et al.*[135], among many others, analyzed the implications of variations in the promoter of the arginine vasopressin receptor 1A (*AVPR1A*) for ASD. *AVPR1A* is a key receptor involved in the regulation of social behavior. The researchers found that shorter repeat alleles of the microsatellite decrease *AVPR1A* transcription, suggesting an increased susceptibility to the autistic phenotype.

Moreover, epigenetic changes of state of promoters have also been associated with ASD. Nagarajan *et al.*[134] found significantly increased methylation of the *MECP2* promoter in the frontal cortex of male individuals with ASD compared to controls, leading to reduced MeCP2 protein expression. Furthermore, a rare *MECP2* promoter variant found in an ASD patient was also correlated with reduced expression of MeCP2. Because expression defects of MeCP2 were previously found in the autistic brain, these data further suggest that both genetic and epigenetic mechanisms affecting promoters may be important in ASD pathology[134].

WGS data have also uncovered an increased load of missense variants located in fetal brain promoters and embryonic stem cell enhancers among probands[133]. Furthermore, small non-coding regulatory deletions in *NR3C2* have been linked to ASD through WGS analysis focused on promoter regions[75]. Modeling loss of *NR3C2* function in zebrafish, revealed that the mutant animals exhibit both social deficits and sleep disturbances. Additionally, *DLG2*, which has been associated with cognition and learning in humans but not previously implicated in ASD, was also found to harbor a recurrent deletion, significantly associated with ASD ($p = 0.003$), in its promoter[75].

3.3.1.3 Cis-regulatory elements (CREs)

Cis regulatory elements (CREs) are non-coding elements, such as enhancers, silencers and insulators, that modulate expression of target distal genes and ensure a tight spatio-temporal regulation by the interplay with core promoters[138].

Although the core promoter is sufficient to direct transcriptional initiation by the RNA polymerase II, alone it only produces basal levels of mRNA. Enhancers augment the rate of transcription of a cognate gene, while silencers result in the opposite outcome. Insulators act to either block enhancers from targeting genes or to provide a barrier between a protein and associated silencer[138].

An increasing number of studies have shown that a high fraction of causative variants in neurodevelopmental disorders are linked to variants in CREs[139]. Moreover, rare inherited and DNMs on CREs have been proven to be causative of ASD[96,133,136,140,141] (Figure 6).

Focusing on DNVs within enhancers with characterized *in vivo* activity, Padhi *et al.* have identified a specific enhancer targeting *EBF3*, a gene found to be associated with ASD on a genome-wide scale. This enhancer has been found to accumulate an excess of DNMs that leads to shared phenotypic characteristics among carrier individuals. These common characteristics include being male, lacking ID, and displaying symptoms such as hypotonia or motor delay[136].

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

By examining and scrutinizing the transcriptional and epigenomic characteristics of genes active in the cortex, Zhao and colleagues reported a set of long genes containing broad enhancer-like domains. These genes were found to play a crucial role in synapse function and development, demonstrating a strong association with NDDs and ASD. Their expression is mostly downregulated in autistic individuals, thus challenging RNA polymerase activity and resulting in their transcriptional impairment, contributing to cellular and functional abnormalities found in ASD patients[137].

Brander *et al.*[96] found that rare inherited structural variants disrupting CREs in genes that are highly dosage sensitive, particularly those with a paternal-origin, are significantly associated with the risk of developing ASD. Furthermore, Kim and colleagues[142] have linked non-coding DNMs to their target genes through a DNase-seq dataset (that defines promoters and enhancers based on regions of open chromatin state, see Chapter 3) and a Hi-C dataset of human neuronal progenitor cell (that defines chromatin interactions, see Chapter 3). Following this approach, they reported that non-coding DNMs show relative enrichment for high disease impact score in probands compared to siblings, and further revealed emerging neurodevelopmental genes, not previously implicated in ASD, including *CTNNA2*, *GRB10*, *IKZF1*, *PDE3B*, and *BACE1*. Of note, these genes contribute to ASD-relevant pathogenic mechanisms and low IQ.

3.3.1.4 Non-coding RNAs (ncRNAs)

ncRNAs, molecules that do not harbor canonical open reading frames and are therefore not normally translated into proteins, are emerging as important regulators in various NDDs due to their indispensable roles in brain development. Important types of regulatory non-coding RNAs mainly include miRNA, PIWI-interacting RNA (piRNA), small nucleolar RNA (snoRNA), small interfering RNA (siRNA), long non-coding RNA (lncRNA), enhancer RNA (eRNA) and circular RNA (circRNA)[11,143].

This section will summarize current knowledge of miRNAs and lncRNAs, both of which have been extensively studied in ASD (Figure 7).

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

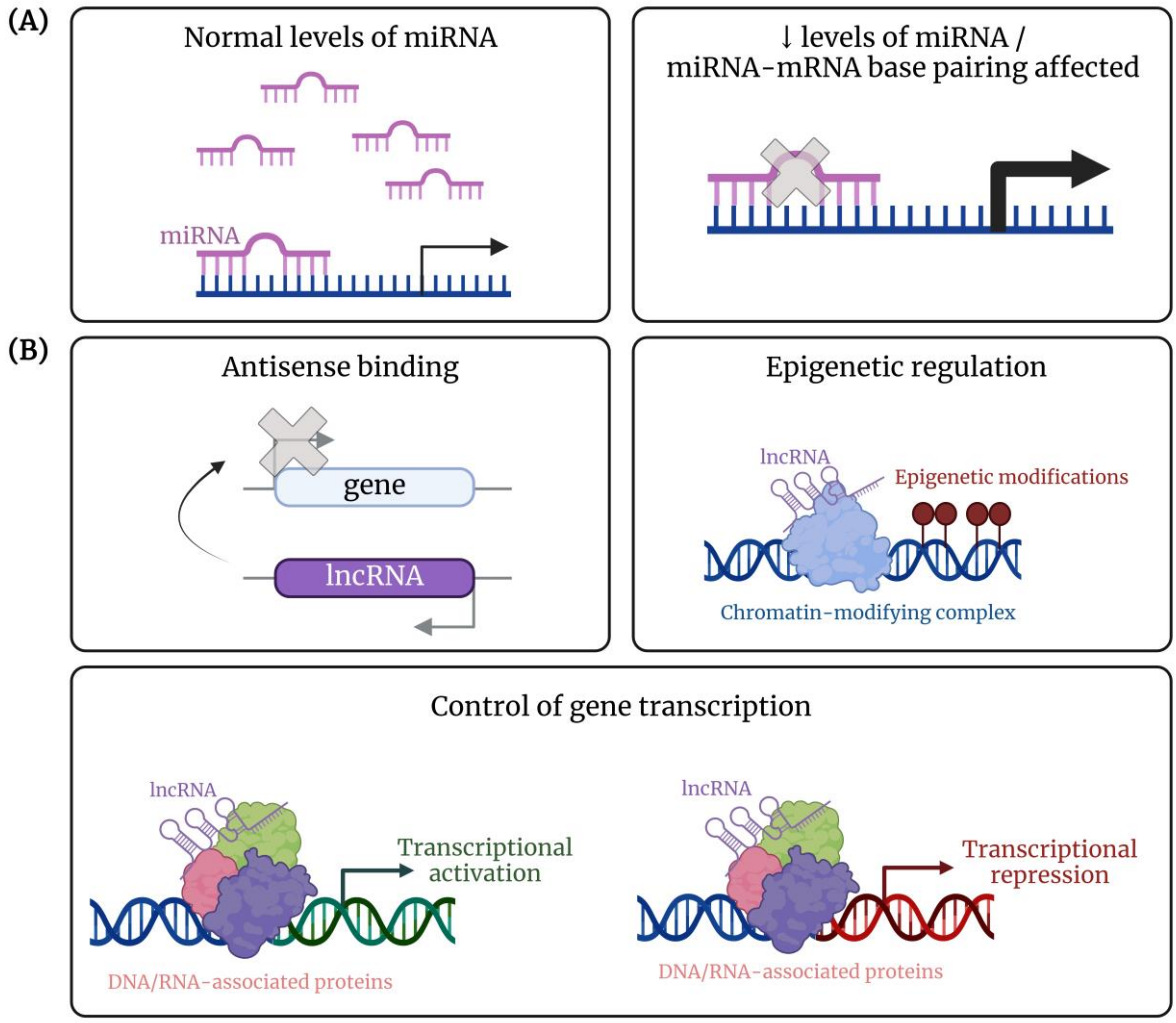


Figure 7. Examples of the impact of miRNAs and lncRNAs in ASD. (A) Pairing of miRNAs (purple) to their target mRNA (blue). (Left) Normal levels of miRNA prompt decreased levels of target gene through (i) inhibition of target mRNA transcription or (ii) silencing of target mRNA translation[144,145]. (Right) Numerous factors (*i.e.*, the reduction in miRNA levels or variants affecting miRNA-mRNA base pairing) may impair miRNA-mRNA pairing, thus increasing the levels of target gene expression. (B) (Top left) LncRNAs (purple rectangle) may function in several ways. One possible role, already reported in ASD, include antisense binding to target gene (blue rectangle), inhibiting its expression[146]. (Top right) LncRNAs may also recruit chromatin-modifying to the target gene locus, leading to activation or repression of local genes[120,143] (Bottom) They can also act as scaffolds for DNA/RNA-associated proteins to regulate transcriptional expression. For example, they can stimulate transcription by bringing key proteins to target gene promoters or inhibit gene transcription by binding to gene repressors[147]. Created with BioRender.com.

3.3.1.4.1 Long non-coding RNAs (lncRNAs)

LncRNAs are a family of non-coding RNA molecules, typically exceeding 200 nucleotides in length, lacking open reading frames, and exhibiting low protein-coding potential[147].

Comprising a substantial portion of cellular transcription, lncRNAs play diverse roles in gene expression regulation. Particularly abundant in the central nervous system, especially in the brain, lncRNAs exhibit specific spatial and temporal expression patterns. Their regulatory

mechanisms include interactions with chromatin complexes for epigenetic control, modulation of proteins or protein complexes, binding to DNA/RNA-associated proteins to influence transcription, maintenance of DNA stability through R-loop and triple helix formation, and participation in higher-order chromatin structure[147].

Implicated in various developmental and physiological processes, such as neural development, brain function, organelle formation, competition for miRNA binding, and regulation of protein localization, lncRNAs contribute significantly to cellular functions and processes[148,149].

Among 166 human diseases that have been found to be associated with lncRNAs (recorded in the lncrnadisease database; http://cmbi.bjmu.edu.cn/lncrnadisease), neurological diseases account for 8.4% of them[150].

Numerous studies have found altered lncRNA levels in ASD brains[42-44]. For example, Ziats and Rennert[151], analyzed over 33,000 annotated lncRNAs from postmortem brain tissue of ASD and control individuals' prefrontal cortex and cerebellum, and detected 222 differentially expressed lncRNAs in ASD that were enriched for genomic regions containing genes related to neurodevelopment and psychiatric disease. Subsequent studies successfully replicated and expanded upon these findings[152,153].

3.3.1.4.2 microRNAs (miRNAs)

miRNAs are endogenous non-coding single-stranded RNAs, approximately 22 nucleotides in length, that are capable of regulating gene expression at post-transcriptional level by binding to the 3'UTR of target mRNAs. They repress protein production by translational repression or messenger RNA destabilization and decay[154]. Each miRNA has the ability to bind many mRNAs, simultaneously influencing the expression of hundreds of target genes. Collectively, miRNAs are predicted to target > 60% of the transcriptome, potentially modulating the corresponding cellular networks[155].

miRNAs are abundant in the brain, with approximately 70% of experimentally detectable miRNAs being expressed in this tissue, playing a vital role in many important brain development facets, such as neuronal survival, function and plasticity. Recent studies have involved miRNA in neurogenesis, synaptogenesis, axon outgrowth and guidance, and neuronal migration and integration[156–158]. In addition, accumulating evidence supports the implication of disturbances in the miRNA system in the pathology of neurological diseases, such as FXS, SZP, spinal muscular atrophy and ASD[159–161].

Numerous studies compared the expression of human miRNAs in individuals with ASD and healthy controls. Altogether, these studies identified more than 200 miRNAs with potential implications in ASD[144,145,155,162–173]. However, no specific miRNA was uniformly dysregulated across all sample sets. In contrast, Vaishnavi *et al.*[174], focusing on ASD-associated CNV loci, as these genomic structural variants are a major contributor to the pathophysiology of ASD, found 71 CNV-associated miRNAs (5 of them were already reported in ASD[161]).

Also, genetic variants in miRNA genes impact their regulatory function[175,176] (Figure 7). As an example, Williams *et al.*[177], identified several rare inherited SNVs within the mature

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

sequence of miRNAs predicted to affect the regulation of ASD-risk genes, such as *NRXN2* and *CNTNAP2*.

Finally, studies in animal models (knockout mice for some ASD candidate genes) have enabled a further characterization of 11 miRNAs in correlation with ASD (for a review, see Schepici *et al.*[178]).

3.3.1.5 Repetitive DNA

It is estimated that over half of the human genome, which consists of approximately 3 billion base pairs of DNA, is made up of repetitive elements that make up the repeatome[126].

Repetitive DNA can be classified as tandem and interspersed repeats and copy number variants. TRs and CNVs may occur in the coding, non-coding and regulatory regions of the genes[179]. Interspersed repeats correspond to retrotransposons, discrete pieces of DNA that are reversed transcribed and inserted at new genomic locations, also occurring in both exons and non-coding regions[180].

3.3.1.5.1 Tandem repeats (TRs)

TRs are consecutive and head-to-tail copies of a sequence, with a compositional complexity ranging from a few bases to more than a hundred, although the most commonly observed repeats in the genome are mononucleotide, dinucleotide and trinucleotide repeats[126].

Particularly, short tandem repeats (STRs), which consist of repeating units of 1-6 base pair motifs, represent a key component of the repeatome. Hundreds of STRs have been involved in gene expression regulation, replication, transcription, translation, recombination and chromatin organization[66,181]. Additionally, TRs have also been associated with molecular and cellular dysfunction, with implications in more than 50 monogenic diseases[126].

Polygenic diseases, such as ASD, are recently being connected with TRs (Figure 8): two studies have identified a large number of complementary, non-overlapping TR variants associated with fetal brain regulatory regions and genes related to the development of the nervous system[66,182]. Moreover, rare TRs are more prevalent in ASD cases than in control siblings and make up to 2.6% of the total risk of ASD[66].

Overall, these results accentuate the importance of considering repeat variants in the study of genetic variants underlying ASD. However, this kind of research faces several limitations: (i) the identification of TR variants remains a challenge and requires stringent filtering in order to achieve high validation rates; (ii) there is a lack of specific datasets, so that validation heavily relies on simulated data; (iii) TRs with complex structures are ambiguous to define and their boundaries highly depend on the choice of parameters used to create the reference, originating huge variance among different methodologies and studies.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

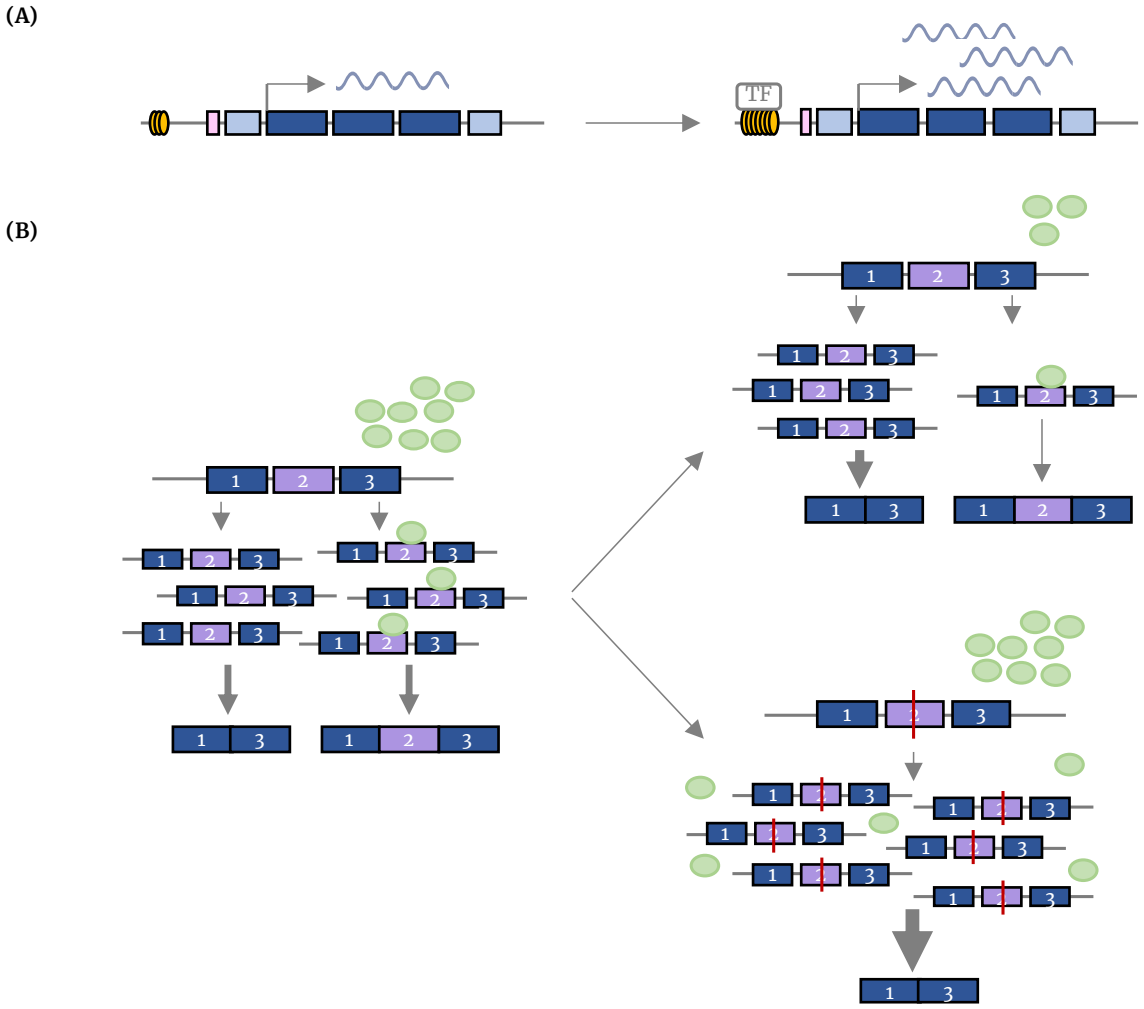


Figure 8. Examples of non-coding ASD-associated variants interfering TR and variants affecting proper functioning of the splicing process. Representation of a gene divided into exons (dark blue and purple rectangles). (A) Tandem repeats. Elongation of TRs (yellow ovals) prompts increased TF binding, and thus, increased mRNA levels/gene expression (represented as undulating lines)[66,182]. (B) Risk variants are represented as red lines; different exons are enumerated (numbers 1 and 3 represent constitutive exons in blue; number 2 represents an alternative exon with regulatory cis-binding sequence, in purple). (Left) Normal levels of trans-acting factor (green circles) result in typical patterns of spliceoform expression. This means that both spliceoforms (*i.e.*, including the alternative exon and/or both constitutive exons) are produced in regular quantities (represented as gray arrows with the same width). (Top-right) Reduced levels of trans-acting factor result in alternative spliced/differentially expressed genes downstream of trans-acting factor (*i.e.*, changes in spliceoform quantities for multiple target RNAs)[183–186]. Here, the spliceoform that includes both constitutive exons, but not the alternative one, has a higher expression due to the unavailability of the transcription factor to bind the cis-binding sequence in the alternative exon. (Bottom-right) Variants in regulatory cis-binding sequence (harbored in a specific exon/intron) are followed by quantitative/qualitative changes in spliceoform expression of a target RNA[187]. Here, the spliceoform including the alternative exon cannot be expressed because trans-acting factors are not able to bind to the mutated cis-binding sequence.

3.3.1.5.2 Retrotransposons

While not as widely implicated in the disease risk by previous research, Borges-Monroy *et al.*[188] also revealed a contribution of non-coding *de novo* retrotransposon insertions

to ASD. In their analysis of retrotransposon insertions (elements LINE-1, Alu, and SVA) in individuals with ASD, a significantly higher rate of *de novo* intronic L1 insertions was observed in probands with syndromic ASD genes curated from SFARI ($p = 0.001$) (*i.e.*, genes implicated in syndromes in which a significant subpopulation develops symptoms of ASD, such as FXS). There was also a trend for more *de novo* intronic LINE-1 insertions in genes with higher pLI ($p = 0.02$). Both requirements are considered in section 3.2.2.3 as evidence for pathogenicity of mutations and candidate risk gene discovery.

Moreover, *de novo* insertions in ASD showed an enrichment trend in regulatory regions of the genes upregulated in the prefrontal cortex (although not significant after multiple test correction, no such enrichment was detected in controls). Additionally, *de novo* Alu insertions were significantly observed in regions with enhancer chromatin marks in fetal brain development ($p = 0.018$). Given that *de novo* retrotransposon insertion rates between cases and controls are similar, but probands had more *de novo* L1 insertions in known ASD genes than expected (similar conclusions than those for coding SNVs or CNVs (see section 3.2.2) and most of the retrotransposon insertions take place in non-coding regions, there is now evidence implicating non-coding insertions in phenotypic impact.

Furthermore, heightened expression of LINE-1 has been detected in specific brain regions of individuals with ASD, along with observed changes in Alu methylation[189].

3.3.1.6 Splicing variants

90% of human genes that encode proteins undergo pre-mRNA splicing, a complex process by which introns are removed and exons are aligned and concatenated to form mRNA. While many exons are constitutively spliced together, alternative splicing (AS) is a process during which genes are spliced and otherwise processed in different ways, thus generating numerous transcripts from a single protein-coding gene[190].

AS is a key regulator of gene expression since precise pre-mRNA splicing is essential for appropriate protein translation. It depends on the presence of consensus cis sequences that define exon-intron boundaries and regulatory sequences recognized by the splicing machinery[190].

Its role in human disease has been widely explored, with an exponential growth on the number of aberrant splicing processes causing human disease[191]. These aberrations may occur at two different levels: cis-acting motifs (exonic/intronic splicing enhancers and silencers) and trans-acting factors that regulate the assembly of the spliceosome by recognition of proximal cis-elements, such as RBPs[192–194]. Moreover, polymorphisms affecting AS have recently been accounted for an estimated 40-60% of general disease risk in numerous diseases[184]. Regarding ASD, the study of AS is currently attracting a lot of attention[195,196] (Figure 8).

The majority of AS studies in ASD patients pertain to individual RBPs that govern the inclusion/exclusion of microexons (3-27 nucleotides)[185,186]. Microexons constitute the most conserved component of the neural-regulated AS program. They generally reside on protein surfaces, specifically in domains with important roles in shaping protein-protein interactions and are significantly enriched in genes with neuronal function which have been genetically linked to ASD, being predominantly neuronal-included[197]. RNA-Seq analysis of post-mortem

brain samples in idiopathic ASD individuals shows misplicing in 30-40% of brain-specific microexons[185]. Analyses of larger cohorts of ASD individuals corroborated these observations and demonstrated that most of the differential splicing events involve the exclusion of these microexons[183,198].

Most microexons are controlled by the neural-specific Ser/Arg-related splicing factor, nSR100/SRRM4, and its misregulation is linked to reduced expression levels of nSR100/SRRM4 in autistic brains[199,200]. Employing heterozygous mutant mice expressing approximately 50% wild-type levels of nSR100, Quesnel-Vallières *et al.*[199] demonstrated that a single variant in a splicing regulator, and thus the disruption of its target splicing program, suffice to reproduce hallmark features of ASD, including altered social behavior, synaptic transmission and neuronal excitability.

Furthermore, Zhou *et al.*[201] revealed, through a deep learning analysis, an increased burden of non-coding mutations that disrupt natural programs of RBPs. Additional examples of RBPs implicated in the differential splicing changes observed in ASD brain are *RBFOX1*[202–204] and *PTBP1*[202,205].

3.3.1.7 Gene discovery in the non-coding genome

Despite many recent efforts, *in silico* definition of the functional effect of non-coding variants is still a challenging task (Figure 9).

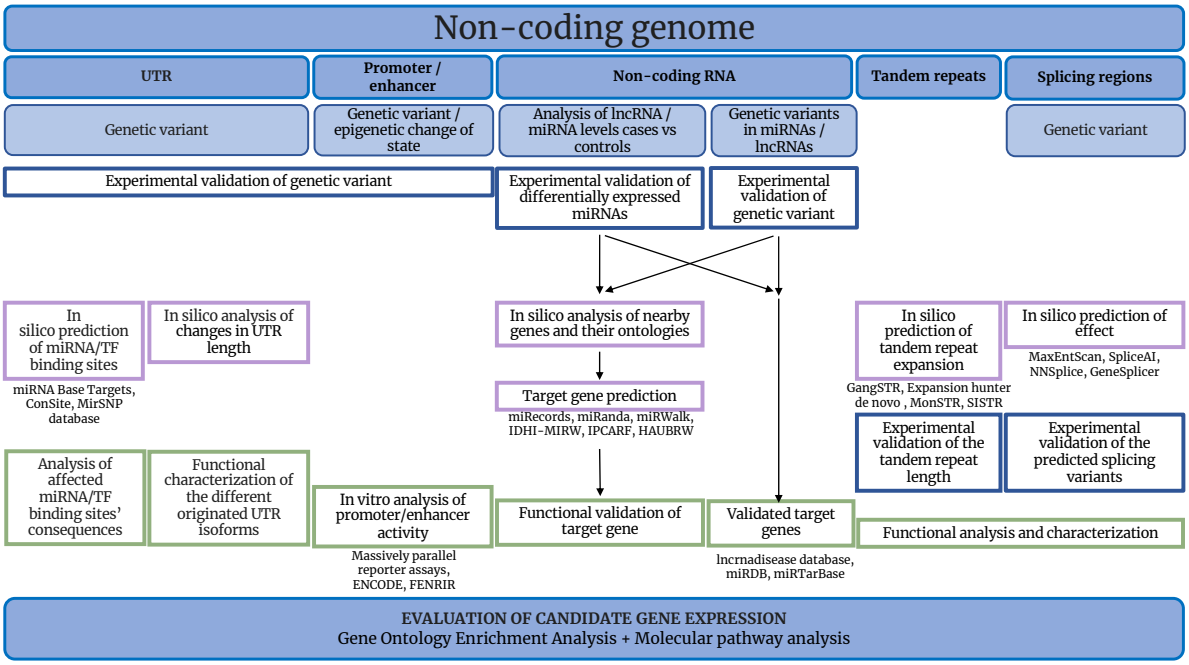


Figure 9. Strategies for the characterization of non-coding variants. This figure depicts examples of *in silico* tools and databases, many of which have already been implemented and assessed in ASD, used to elucidate the functional mechanisms underlying pathogenic variants.

As previously outlined, the most evident role of variants in UTRs relies on their impact on the disruption of existing miRNA binding sites or on the creation of new ones. Thus, most prediction algorithms rely on the imperfect base-pairing between miRNAs and their target sites

(for further detail refer to references [206,207]). Alternatively, some algorithms assess ribosome loading levels[208] or investigate the utilization of alternative initiation sites for protein synthesis, aiming to understand potential alterations in gene expression levels[209].

The vast majority of promoter variation studies in ASD have demonstrated the involvement of genetic variations by investigating specific SNPs and their impact on gene expression through functional analyses[135,210,211]. Alternatively, a global approach involves the use of massively parallel reporter assays (MPRA), allowing simultaneous testing of the regulatory activity of numerous regulatory sequences and their variants[212].

Regarding enhancers, the computational mining of genetic data together with expression data in brain tissue, and the use of different databases providing information on TF binding sites, chromatin and epigenetic marks, is the basis for the development of multiple bioinformatic tools for predicting the effect of variants in these non-coding regions[213]. For example, Zhang *et al.*[214], combined GWAS data with brain region related enhancer-gene networks, providing novel insights into the pathogenesis of ASD by identifying ASD-associated pathways. Other successful approaches include the use of machine learning algorithms such as ExPecto[215], which can predict the tissue-specific transcriptional effects of variants on a DNA sequence without the need for functional experiments.

On another level, thanks to the use of enhancer browsers and databases, such as ENCODE, Roadmap Epigenomics, VISTA, FANTOM5 or dbSUPER (reviewed in Perenthaler *et al.*[140]), along with specific computational approaches to predict functional CREs *in silico*, Chen *et al.*[141], developed a machine learning algorithm called FENRIR that enables functional characterization of tissue specific enhancers and allows for the prioritization of enhancer-disease associations. Among the 2,723 target genes regulated by the FENRIR-predicted ASD-associated enhancers, significant functional enrichment of pathways and processes previously reported to be associated with ASD was detected. These included neuronal development and synaptic functions, chromatin-related functions, and Wnt signaling pathways.

Researchers have started to describe specific pipelines facilitating genome-wide expression profiling, including the identification of novel and rare transcripts like ncRNAs from RNA-Seq data[216,217]. Moreover, various computational approaches have been developed to identify disease-associated lncRNAs, offering potential applications as biomarkers to enhance disease detection, prognosis, and clinical outcomes[218–220]. However, it was only 3 years ago that Wang & Wang developed an efficient machine learning approach to predict and prioritize ASD-associated candidate lncRNAs, providing targets for further investigation[221].

The read length of STRs makes their accurate detection complicated with current technologies, so the development of novel tools to make this possible is needed. Two main algorithms have been used to estimate repeat lengths in ASD: GangSTR, for repeat unit lengths of 1-20 base pairs[222], and ExpansionHunter Denovo, optimized for detecting repeats longer than 150 base pairs[223]. Because of recent estimations that DNM may contribute up to 30% of simplex ASD cases, Mitra *et al.*[182], developed an algorithm called MonSTR that allows for the identification of *de novo* STR variants, and another bioinformatics tool called SISTR that determines the probability of each STR variant to have a deleterious consequence.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Lastly, several algorithms have been employed for the study of AS in ASD, allowing for the prioritization of a large number of variants depending on their potential effect on splicing. These include GeneSplicer, MaxEntScan, BDGP/NNSplice or Human Splicing Finder[224].

Unlike the aforementioned algorithms, that largely only focus on local motifs, computational tools that account for long-range specificity determinants have much greater performance. Xiong *et al.*[225], developed a machine learning algorithm that analyses DNA sequence features (cis elements) and uses them to predict the number of transcripts with the given exon spliced in. The algorithm's implementation in an ASD cohort resulted in ~200 genes with at least one variant predicted to affect splicing. Additionally, they prioritized 19 genes as more compelling ASD disease candidates because of their known neurological, neurobehavioral, or neurodevelopmental phenotypes in humans, with just two of them being previously suggested to have a role in ASD pathology. Further on, obtaining tissue-specific models of alternative splicing from particular relevant tissues implicated in the disease may help to gain greater insight into the etiopathology of ASD. Following this intention, Cheng *et al.*[226], developed Multi-tissue Splicing (MTSplice), that predicts the effect of genetic variants on splicing in 56 human tissues. They assessed the potential of MTSplice on scoring a large number of DNMs in thousands of ASD simplex families. It was found that DNMs from probands disrupted splicing sites in brain tissues more severely than variants of the control group for all tissues. Jaganathan *et al.*[36], modeled a deep neural network called SpliceAI that predicts the splice function of each position in the pre-mRNA sequence, so that each variant can be scored for its splice-altering impact evaluating 10,000 nucleotides of the flanking context sequence but also taking into account nucleosome positioning (because of its influence in chromatin state) and exon-intron lengths (see Chapter 2 for further detail). Due to the wide window analyzed, its accuracy exceeded that from other algorithms for splice site detection and allowed them to estimate that that 9%–11% of the pathogenic variants in patients with rare genetic disorders are caused by cryptic splicing (*i.e.*, the use of unannotated acceptor/donor sites), a process previously underappreciated[36].

3.3.2 Unbiased: WGS (whole genome sequencing) analysis

Previous publications outlined in the preceding section have explored the impacts of various non-coding variant classes. However, these studies primarily focused on individual or a limited number of regulatory features of the non-coding genome, leading to an incomplete representation of the genomic architecture of ASD.

This is understandable considering that WGS association studies require the same statistical precision that has been applied to GWAS-, or WES-based gene discovery. Thus, WGS analyses are subject to addressing the substantial multiple-testing burden, particularly problematic in non-coding portions of the genome.

In the following sections, we will be discussing (i) limitations in WGS analyses, (ii) strategies to solve existing limitations, and (iii) current outcomes of such approaches.

3.3.2.1 WGS limitations

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

The issue when studying the whole genome arises when seeking for association in the elevated number of functional annotations and gene-sets that could potentially be tested[96].

For instance, An and colleagues[41] assessed WGS data trying to characterize the functional role of non-coding mutations. In order to do so, they performed a CWAS (category-wide association study), a data analytic tool employed for stringent association tests aimed at identifying non-coding loci associated with the disease of interest. CWAS utilizes category-based burden tests, utilizing *de novo* variants from WGS data along with diverse annotation datasets. In their analysis, they integrated multiple datasets encompassing gene function, genes associated with NDDs, conservation across species, and epigenetic markers, thereby combinatorially defining 55,143 categories. As stated in their work: "The scope of the problem—testing for an excess of *de novo* mutations in cases relative to controls for each category—is challenging because there are more categories than families", a recurrent issue cited by many other authors in similar analyses[95,227].

Therefore, a priori prediction of which regulatory elements of the non-coding genome are important for disease risk, limits the number of tests evaluated and consequent statistical corrections.

Moreover, there are no reliable estimates for the quantity of non-coding loci capable of influencing risk, the magnitude of such risk, or the genomic features of these loci, unlike the current situation with coding loci. Numerous studies have been unable to detect non-coding categories significantly linked to ASD cases and have struggled to create a reliable predictor of case status[66,95]. In contrast, associations of missense/LoF mutations in ASD genes, an excess of significant coding categories and polygenic risk scores have been thoroughly established using comparable approaches (see sections 3.2.2-3.2.4). Examining these findings within the framework of a power analysis provides crucial insights into genomic architecture: it seems improbable that there exists a non-coding variation class equivalent to coding LoF mutations in both mutation frequency and effect size[95].

These observations underscore the inherent challenges in identifying robust and consistently reproducible non-coding signals within a genetically heterogeneous disorder like ASD. Additionally, they emphasize the necessity for an even greater volume of ASD WGS data and the constant need for larger datasets (Table 4).

3.3.2.2 Characterization of the non-coding genome in WGS

Enhanced characterization of the non-coding functional genome could offer marginal improvement in the aforementioned scenario, facilitated by techniques like Hi-C, for instance.

Hi-C is a chromosome conformation capture technique used in molecular biology to investigate the three-dimensional structure of the genome first described in 2009 by Lieberman-Aiden *et al.*[228]. It allows researchers to study how different regions of DNA physically interact with each other in the nucleus of a cell, *i.e.,* it detects genome-wide chromatin interactions in the cell nucleus by combining chromosome conformation capture and NGS[229].

As an example, Grove *et al.*[37] discovered that common variation in ASD is situated in regions highly enriched with regulatory elements predicted to be active during human corticogenesis. Considering that many gene regulatory events occur at a distance through

chromosome looping, the researchers utilized Hi-C data to identify potential target genes for these variants. After obtaining a set of credible SNPs, considered putative causal SNPs, from the top-ranking loci identified in the ASD GWAS, the resulting 28 loci were cross-referenced with Hi-C reference data for analysis. The credible SNPs exhibited a significant enrichment with enhancer marks in the fetal brain, reinforcing the likely regulatory role of these SNPs during brain development.

These findings, among others, underscore the importance of non-coding regions and their impact on critical processes for neurodevelopment, such as gene expression regulation or splicing. These aspects are elaborated in detail in the following section.

Alternatively, predicting tissue-specific effects, inherent to regulatory regions[230], could further enhance current knowledge. Some methods leverage epigenetic information extracted from diverse cell types, employing techniques like ChIP-Seq (chromatin immunoprecipitation with sequencing), ATAC-Seq (assay for transposase-accessible chromatin with sequencing) and DNase-Seq (DNase I hypersensitive sites sequencing), both sequencing assays which capture regions of open chromatin[231], and PLAC-Seq (proximity ligation-assisted chromatin immunoprecipitation sequencing; refer to Chapter 3). Computational models then utilize this data to predict the tissue-specific expression effects of non-coding variants.

Moreover, there is an absence of a non-coding counterpart to the triplet code found in protein-coding regions, which has been a crucial element for predicting the impact of mutations on gene function within the coding genome. Additionally, unlike coding exons, the borders of non-coding regions lack clear definition leading to arbitrary definition of boundaries. Therefore, various annotations, including evolutionary conservation and sequence constraint (as seen in WES gene discovery; section 3.2.2.3), or focusing on regulatory regions near-known ASD-risk genes[232], serve as valuable guides.

Finally, just as the design of exome studies in simplex or multiplex families enhances the capability to detect *de novo* and inherited mutations, respectively, a similar approach may be taken in selecting autistic probands for WGS: selecting those that are negative for known pathogenic mutations, including large CNVs and LoF DNMs, as identified through WES and SNP microarray analyses. Consequently, the exclusion of most high-impact DNMs in coding sequences enables a specific focus on more challenging cases of the disease where no genetic etiology had been established.

3.3.2.3 Key findings in WGS

Following the strategies highlighted in the previous section, among many others, researchers have come to several conclusions regarding the non-coding genome, all of them resembling those for coding regions (see sections 3.2.2.1, 3.2.2.2, and 3.2.3).

In brief, these include:

1. Increased proband burden of high-effect mutations close to LoF-intolerant genes/ASD gene datasets[61,232], suggesting that damaging non-coding variants probably affect gene function in a similar manner to coding mutation.
2. A consistent significant proband-specific mutation effect associated with brain tissues, which offers compelling evidence that high-impact variants originating

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

from the non-coding genome of ASD probands likely disrupt brain-specific gene regulation, which aligns with prior findings related to mutations in protein-coding regions[233].

3. Individuals with mutations in conserved loci show a trend toward a higher rate of mutations in female ASD cases, again, consistent with results seen in coding mutations[41].
4. Involvement of non-coding mutations in synaptic transmission and neuronal development, revealing a convergent genetic landscape of coding and non-coding mutations in ASD[95].
5. D*e novo* risk scores, especially at sites conserved across species, driven by variation in the distal promoter or in regions targeted by transcription factors, has also revealed a non-coding contribution to ASD.

Considering both the evidence supporting a convergent role of coding and non-coding variation in genes, biological pathways and increased burden in probands, coupled with the greater number of unbiased WES in the research field (as evident in Table 5 *versus* Table 8), we anticipate that, as more extensive ASD datasets are gathered, and approaches similar to those employed in coding regions are adapted for non-coding discovery, numerous non-coding genetic risk influences will be identified.

Table 8. Key supporting evidence for the contribution of non-coding variation to the genetic etiology of ASD.

Study	Data description	Relevant key findings	Associated ASD genes
Turner *et al.*, 2016[232]	WGS 39 simplex families + 13 trios (SSC)	Probands show a significant enrichment of *de novo* and private disruptive mutations within putative regulatory regions (*i.e.*, highly conserved transcription TFBS in the CNS DHSs).	The potentially disruptive events are located in proximity to genes associated with ASD risk, including *DSCAM*, *SCN2A*, *NR3C2*, *PRKCA*, and *ARID1B*.
Yuen *et al.*, 2016[61]	WGS 200 *trios*	There is a significant enrichment of predicted damaging DNMs in ASD cases, of which 15.6% are non-coding.	Genes disrupted by the DNMs are enriched in pathways related to chromatin organization, RNA processing and translation and synaptic transmission: *SHANK2*, *EIF4E* and *DAPK1*, among others.
Doan *et al.*, 2016[234]	2,100 sibling-matched ASD probands (SSC)	There is a 6.5-fold enrichment of rare, *de novo* CNVs within HARs (human accelerated regions, implicated in the regulation of genes important in human evolution) among individuals with autism when compared to sibling-matched controls.	*CUX1*, *CDKL5*, *GPC4* and *PTBP2* (among other genes already implicated in neural function and ASD).
Brandler *et al.*, 2018[96]	WGS 829 families (REACH + SSC)	A depletion of SVs is identified in fetal brain promoters (OR = 0.73; $p < 0.05$). In ASD cases, paternally inherited CRE-SVs in fetal brain promoters, TSSs or 3'UTRs of	Recurrent CRE-SVs are observed in ASD cases, involving genes such as *CNTN4*, *LEO1*, *RAF1*, and *MEST*.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

		intolerant genes are transmitted at an elevated rate (transmission rate = 70.9%; p <0.05).	
An *et al.*, 2018[41]	WGS 1,902 quads	A developed "*de novo* risk score", with 163/238 non-coding annotation categories, results in a significant predictor of case status. The risk score remains a significant predictor of case status when only promoter categories are included.	-
Werling *et al.*, 2018[95]	WGS 519 SPX families (SSC)	Categories restricted to non-coding *de novo* indels in the CWAS analysis (51,801 categories) show a greater number of nominally significant results than expected by chance (p = 0.03). However, no significant enrichments are detected in any class of non-coding variation after correction for multiple hypothesis testing.	-
Zhou *et al.*, 2019[201]	WGS 1,790 SPX families (SSC)	There is a significant association between non-coding mutations and IQ, specifically in intronic regions flanking alternatively spliced exons of LoF-intolerant genes. Enrichment analysis highlights processes such as synaptic transmission (*e.g.*, ion channels and receptors related to neurogenesis, synaptic signaling and synapse organization) and chromatin remodeling.	Previously known ASD-associated genes, such as those involved in synapse organization (*SHANK2*, *NLGN2* and *NRXN2*), synaptic signaling (*NTRK2* and *NTRK3*), ion channels (*CACNA1A*, *CACNA1C*, *CACNA1E*, *CACNA1G* and KCNQ2) and neurotransmission (*SYNGAP1*, *GABRB3*, *GRIA1* and *GRIN2A*).
Castelijns *et al.*, 2020[235]	H3K27ac ChIP-Seq (45 ASD patient brains and 49 controls)	There is a statistically significant enrichment between hominin-specific lineage regulatory gains is selectively deregulated in ASD patient brains.	These gene regulatory elements are linked to genes previously proposed to play a role in ASD such as *c-MET*, *CITED4*, *NLK*, *CAMK2A*, and *DNMT3A*.
Rodin *et al.*, 2021[236]	Ultra deep WGS of PFC (59 ASD cases and 15 controls)	Excess of somatic mutation in DHSs (p <0.05), specifically, nearby to genes with brain-specific expression, ASD cases are significantly enriched for mosaic mutations in brain-active enhancers.	*CNTN2*, *TMEM81*, *RBBP5*, *CDK18*, among others.
Trevino *et al.*, 2021[237]	1,902 SSC families	Significant enrichment of non-coding variation predicted to be disruptive in ASD (> 1.2-fold) in glutamatergic neurons.	*NFIA* and *NPY*.
Trost *et al.*, 2022[122]	WGS 5,100 individuals with ASD and 6,212 relatives (MSSNG)	Significant enrichment in ASD was observed for non-coding TREs but restricted to the African population (OR = 5.0, FDR <0.05). Variants predicted to damage promoters (private to a particular family and having a frequency <0.1%) were over-transmitted to individuals with ASD (FDR <0.05).	134 ASD-associated genes with FDR <0.1.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CNS, central nervous system; DHSs, DNase hypersensitivity sites; FDR, false discovery rate; TFBS, transcription factor binding sites.

3.4 GENES AND MOLECULAR PATHWAYS

In addition to advancing our understanding of the genetic landscape of ASD, the findings from these recent studies provide preliminary insights into the molecular mechanisms influencing the development of the disease. According to the genetic results, susceptibility genes appear to converge within three biological pathways: (i) transcription and translation, (ii) synaptic function, and (iii) chromatin remodeling[204,238].

Simultaneously, multiple studies have pointed towards disruptions in immunity and inflammation, as well as in the serotonin-NAS-melatonin pathway[39,239].

3.4.1 Aberrant mRNA translation of synaptic proteins

Neuronal activity regulates both gene transcription and protein translation[199]. Numerous studies indicate that disruption of this activity-dependent gene transcription and translation are associated with ASD[95,240].

Thus, the impairment of activity-dependent transcriptional regulators or their targets has been linked to ASD, including the Ras-ERK signaling pathway. Iossifov *et al.*[58] analyzed exome sequencing data from 343 ASD families, revealing an enrichment of *de novo* mutations in genes coding for proteins associated with the Fragile X Mental Retardation Protein (*FMRP*). The *FMRP–EIF4E–CYFIP1* complex, which acts downstream of the Ras-ERK signaling pathway and regulates neuronal translation, causes the FXS, a well-known single-gene cause of ASD. This complex modulates the translation of over 1000 genes, many of which are associated with ASD risk[241,242]. Moreover, an analysis of CNV burden in ASD indicated an enrichment of CNVs in targets of FMRP[243].

Furthermore, the PI3K/mTOR pathway, governing global mRNA translation, has been linked to an increased risk of ASD[244]. Mutations in repressors of the mTOR pathway, such as *TSC1*, *TSC2*, and *PTEN*, support dysregulation of neuronal translation in ASD[58]. Notably, monogenic forms of ASD, such as tuberous sclerosis (linked to *TSC1* and *TSC2*) and Cowden syndrome (linked to *PTEN*), are associated with these canonical components of the mTOR pathway.

Additionally, various genes associated with ASD are believed to regulate or interact with the canonical Wnt/β-catenin-dependent signaling pathway, a family of proteins crucial for transcription control in embryonic development and adult tissue homeostasis[245,246]. Key proteins in the pathway, localized at synapses, play essential roles in synaptic growth and maturation. ASD-associated MET tyrosine kinases, such as *CDH8*[73], members of the Brg1-associated factors, and mixed-lineage leukemia complexes, influence β-catenin dynamics, preventing it from activating target gene transcription[247,248]. The *CTNNB1* gene, encoding β-catenin, is also identified among ASD risk variations[239].

3.4.2 Synaptic function

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Numerous proteins encoded by genes associated with ASD play crucial roles in various aspects of the synaptic structure and function by directly encoding synaptic scaffold proteins, neurotransmitter receptors, cell adhesion molecules, and proteins related to actin cytoskeletal dynamics[5]. Examples include proteins involved in excitatory glutamatergic and glycinergic (*e.g.*, *GRIN2B*[57,67,201]) neurotransmission, inhibitory GABAergic signaling (*e.g.*, receptor subunits β3, *GABRB3*[201]), neurogenesis (e.g., *CNTN*[96,236]), synaptic identity establishment (*e.g.*, cadherins and protocadherins[73]), neuronal conduction (*CNTNAP2*[46,177]) and permeable ion channels (voltage-dependent calcium channel subunit α1C (*CACNA1C*[201]) and voltage-gated sodium channel (*e.g.* protein type 2 subunit-α (*SCN2A*)[73,238]). Additionally, postsynaptic density scaffolding proteins like SH3 and multiple ankyrin repeat domains protein (*SHANK*[73]), cell adhesion molecules involved in activity-driven synapse formation (namely, neuroligins (*NLGNs*) and neurexins (*NRXNs*)[249]), among others, are implicated.

Changes in the signaling of these proteins can modulate the strength or number of synapses, ultimately impacting the structure and functional connectivity of neuronal networks in the brain. For instance, deletions, duplications, and coding mutations in the three *SHANK* genes (*SHANK1*, *SHANK2* and *SHANK3*[73]) have been consistently reported in individuals with ASD. SHANK proteins assemble into large molecular platforms in interaction with glutamate receptors and actin-associated proteins.

In summary, dysregulation in synaptogenesis and synaptic transmission has significant implications for ASD. Meanwhile, the glutamatergic and GABAergic synaptic dysfunction raise the hypothesis that disruption in the excitatory/inhibitory (E/I) balance leads to ASD[239].

3.4.3 Chromatin remodeling

Increasing evidence indicates that ASD is the result of a complicated interaction between genes and the environment, with epigenetic factors positioned at the genome-environment interface (Figure 10). Epigenetic mechanisms, including DNA methylation, histone modification, chromatin remodeling, and non-coding RNA activity, are involved in the regulation of social behavior in autism[250].

Much of the epigenetic research in the context of ASD has centered around DNA methylation[29]. A prominent candidate gene linked to ASD is Methyl-CpG-binding protein-2 (*MeCP2*), a crucial regulator in neural development. MeCP2 binds to methylated CpG sites in gene promoters and interacts with chromatin silencing complexes, suppressing gene expression [134,251]. In ASD, there is observed increased binding of *MeCP2* to the promoters of other well-established ASD risk genes: *RELN* and *GAD1*[47,134,252]. Notably, mutations in the *MECP2* gene are the most prevalent cause of classical Rett Syndrome cases, one of the most common genetic disorders associated with ASD (Table 3).

Moreover, recent studies have revealed a characteristic histone acetylation signature in the brains of ASD patients, particularly associated with genes involved in synaptic transmission and morphogenesis[253].

As an example, mutations in *CHD8* have been identified in a myriad of large scale analysis of parent-child trios/quads[58–60,67,69,254] establishing it as one of the most frequently mutated and most penetrant genes in ASD. *CHD8* is located at active transcription sites with the histone

modification H3K4me3 or H3K27ac and recruits histone H1 to target genes by remodeling the chromatin structure. Moreover, it interacts with *ARID1B*, a component of SWI/SNF, an ATP-dependent human chromatin remodeling complex that is frequently mutated in ASD[239].

Furthermore, *KMT2E*, a lysine methyltransferase, is yet another example of a gene involved in epigenetic regulation and was implicated in both the latest ASD GWAS[37] and WES[40] analyses.

3.4.4 Concluding remarks

In brief, there are a multitude of genes linked to ASD that play roles in various cellular functions (namely, chromatin remodeling, mRNA translation and transcription, and synaptic function). Nevertheless, the resultant effects of these associated genes tend to interact and converge towards a disruption in neuronal and synaptic homeostasis rather than functioning independently[239].

For instance, *CHD8*, known for its involvement in epigenetic regulation, has been identified as β-catenin binding partner, crucial for sustaining Wnt signaling during neocorticogenesis[255]. Additionally, *CHD8* influences the development of GABAergic interneurons by modulating the expression of DLX genes, which encode transcription factors pivotal in GABAergic interneuron differentiation[256].

Moreover, post-mortem analyses have revealed DNA methylation differences in cortical regions of ASD patients compared to controls. Notably, altered DNA methylation patterns were identified, impacting the regulation of *SHANK3*, involved in synapse signaling[257]. Furthermore, analysis of cortical co-methylation networks identified several co-methylated modules significantly associated with ASD, enriched for genomic regions annotated to genes involved in the immune system, synaptic signaling, and neuronal regulation[29].

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

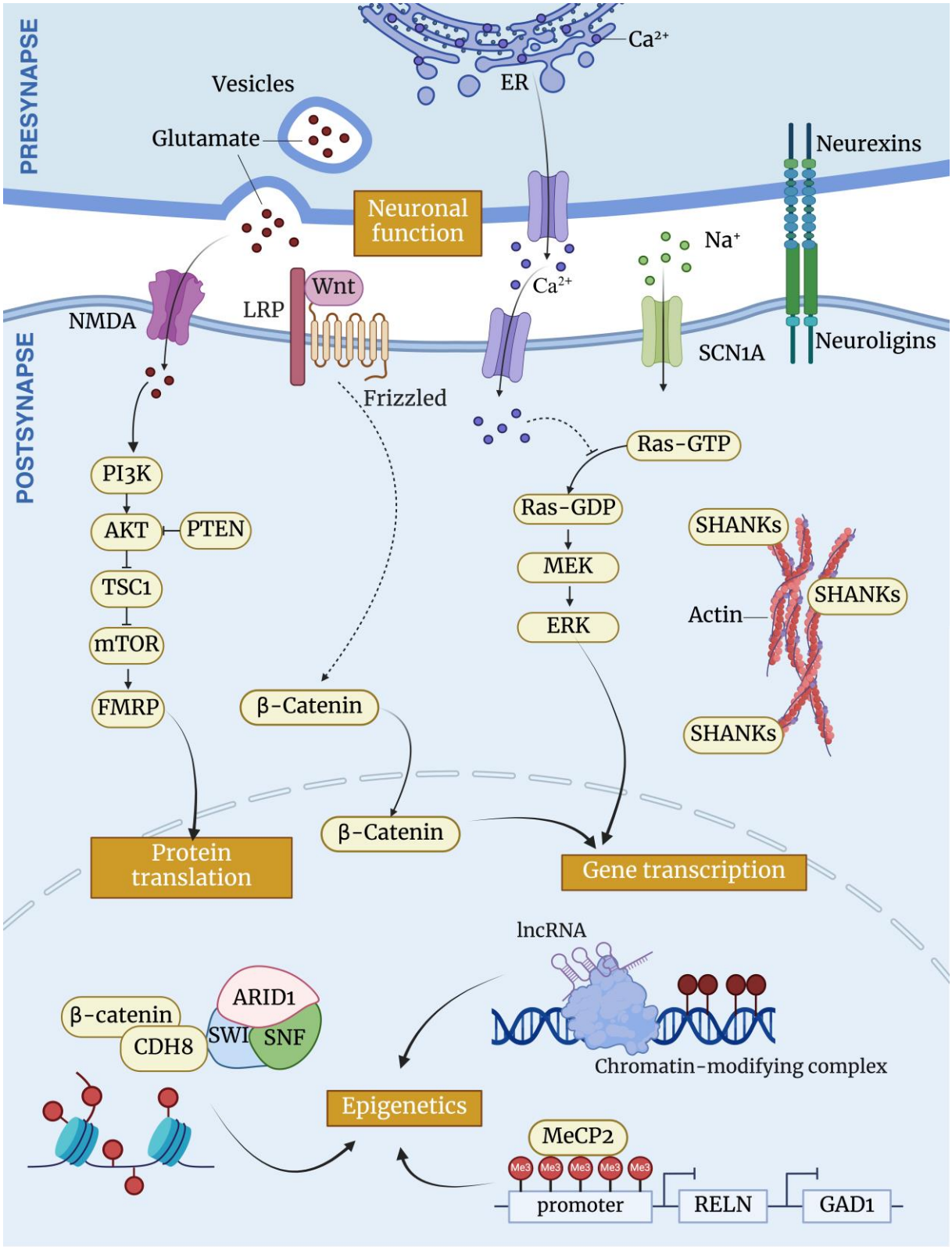


Figure 10. Biological pathways frequently associated with ASD. Current research on ASD susceptibility genes has converged on three biological pathways: (i) transcription and translation, (ii) synaptic function, and (iii) chromatin remodeling. *ER, endoplasmic reticulum.* Created with BioRender.co

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

JUSTIFICATION AND OBJECTIVES

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

4 JUSTIFICATION AND OBJECTIVES

The justification for this research stems from the imperative to address the incomplete understanding of the genetic factors contributing to ASD etiology. Previous studies have predominantly focused on the coding genome, unraveling the implications of hundreds of ASD risk genes, and explaining core biological processes disrupted in the disease.

In contrast, a comprehensive understanding of the non-coding genome remains elusive.

This research is critical because recent evidence has highlighted initial insights into its involvement in disease etiology. As subsequent studies on the non-coding genome continue to progress and ASD risk variants affecting genetic regulatory elements are identified, we anticipate bridging the gap of the missing heritability in ASD.

However, distinguishing deleterious regulatory variants from the extensive background of neutral variation in the genome is challenging. As a result, initial applications of whole-genome exploration have been underpowered to detect the association of non-coding variants with ASD.

Hence, this thesis emerges from the imperative not only to generate new genetic data for enhancing the comprehension of the non-coding genome in ASD but also to leverage the existing data, much of which has been overlooked for the significance of the non-coding genome owing to a prior lack of tools for its comprehensive assessment. Therefore, we will employ recently developed tools for this purpose and outline specific workflows to enhance the discovery of non-coding regulatory roles in ASD disease.

To this end, the primary objectives of this thesis and their respective subobjectives are:

1. Investigate how common variation associated with ASD, predominantly located in the non-coding portion of the genome, influences disease onset.
    - 1.1. Identify specific genes whose expression is altered through the execution of an eQTL colocalization analysis that integrates expression data from GTEX with summary statistics from the largest ASD GWAS conducted to date.
    - 1.2. Conduct follow-up analysis in correlated NDDs to provide additional support for pertinent findings.
2. Evaluate the potential repercussions of AS variation using exome data in an ASD cohort of 360 trios by employing deep learning algorithms.
    - 2.1. Predict variant effects on splicing.
    - 2.2. Conduct *in silico* validations using algorithms that integrate expression data.
    - 2.3. Assess the tissue-specificity of these effects.
3. Characterize the impact of genetic variation (both *de novo* and inherited) on the regulatory landscape in ASD.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3.1. Perform a targeted sequencing study of specific regulatory elements (cCREs) from ENCODE in a cohort of 200 ASD trios, focusing on ASD patients without significant findings in the exome.
3.2. Unravel the regulatory functions of these variants through a deep learning approach.
3.3. Elucidate candidate genes regulated through the integration of PLAC-Seq data.

These objectives have been carefully designed to address the recognized gaps in ASD heritability and contribute substantially to the understanding of its non-coding etiology. The successful achievement of these objectives will not only advance academic knowledge but also open up novel avenues to subsequent research.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CHAPTER 1

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

5 CHAPTER 1

Part of the content of this chapter has been published as Dominguez-Alonso S[1], Carracedo A[2], Rodriguez-Fontenla C[3]. eQTL colocalization analysis highlights novel susceptibility genes in Autism Spectrum Disorders (ASD). Transl Psychiatry. 2023;13(1):1-9. doi:10.1038/s41398-023-02621-0

1. Grupo de Medicina Xenómica, Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidad de Santiago de Compostela, Santiago de Compostela, Spain.

2. Grupo de Medicina Xenómica, Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidad de Santiago de Compostela, Santiago de Compostela, Spain; Grupo de Medicina Xenómica, Fundación Instituto de Investigación Sanitaria de Santiago de Compostela (FIDIS), Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidad de Santiago de Compostela, Santiago de Compostela, Spain.

3. Grupo de Medicina Xenómica, Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidad de Santiago de Compostela, Santiago de Compostela, Spain.

NDDs exhibit a strong polygenic nature, wherein a significant portion of their heritability is ascribed to common genetic variation. In the present day, numerous susceptibility loci have been uncovered through extensive GWAS, particularly in the realm of disorders like SZP, MD, and ADHD, boasting high heritability estimates (ranging from ~40-80% in MD[258,259] up to 70–80% in ADHD[260] and SZP[261]).

Nevertheless, a decade ago, starting with optimistic hypothesis stemming from the elevated heritability estimates, it was posited that if the genetic influences were limited in number, prevalent in the population, and exerted substantial genetic impacts, relatively small sample sizes would prove adequate[262]. This held true for certain traits, as several significant associations within the *APOE* locus were identified with sample sizes as modest as less than 2,000 cases of Alzheimer's disease (AD)[263].

Still, initial negative outcomes in identifying common risk variants in other previously mentioned NDD disorders, with sample sizes of ~1000-2000 cases[264,265], suggested that effects of common ASD risk variants must, individually, be very small and/or that other types of variants may contribute to the disease's genetic liability (*e.g.*, rare variants). This resulted in skepticism regarding the utility of GWAS as a tool for detecting associated genetic components.

However, in spite of the initial difficulty in detecting genome-wide significant effects for most of the disorders, psychiatric genetics investigators stood up under the premise "Don't give up on GWAS"[266], anticipating the effectiveness of GWAS to grow with larger sample sizes and thus, the need to establish global community efforts in psychiatric genetics (Table 4). Notably, a distinct correlation between sample size and the number of significant associations is evident (Figure 11).

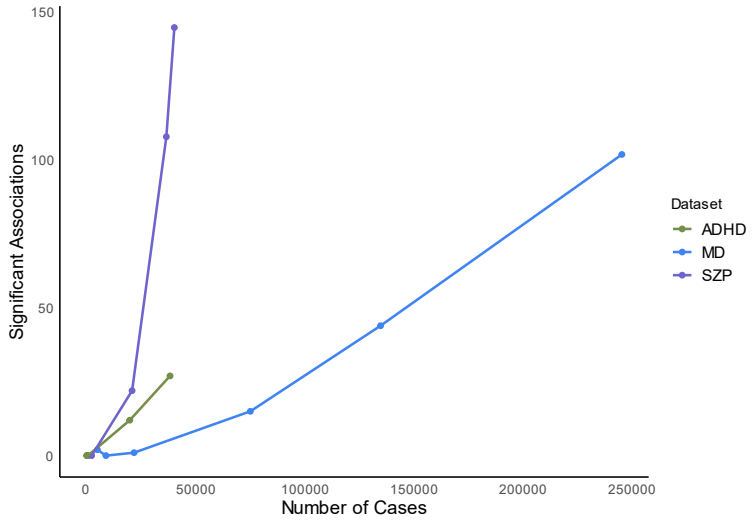


Figure 11. Comparison of GWAS sample sizes and discovery rates. The correlation between the number of cases and genome-wide significant GWAS loci is depicted through lines representing the discovery paths for ADHD, SZP, and MD. Notably, SZP exhibits an exceptional genetic architecture, allowing for substantial discoveries with relatively smaller case numbers. In contrast, MD and ADHD follow more standard and comparable discovery paths, requiring larger sample sizes to achieve a similar number of significant findings as SZP. Data were compiled from references 258–261, 264, 267–275.

Up until now, hundreds of ASD loci have been pinpointed, however, it was not until the most recent ASD GWAS meta-analysis (18,381 cases and 27,969 controls), conducted just four years ago, that the first common risk variants were robustly linked to ASD by more than doubling the discovery sample size compared to previous GWAS (Table 7, Figure 12)[37].

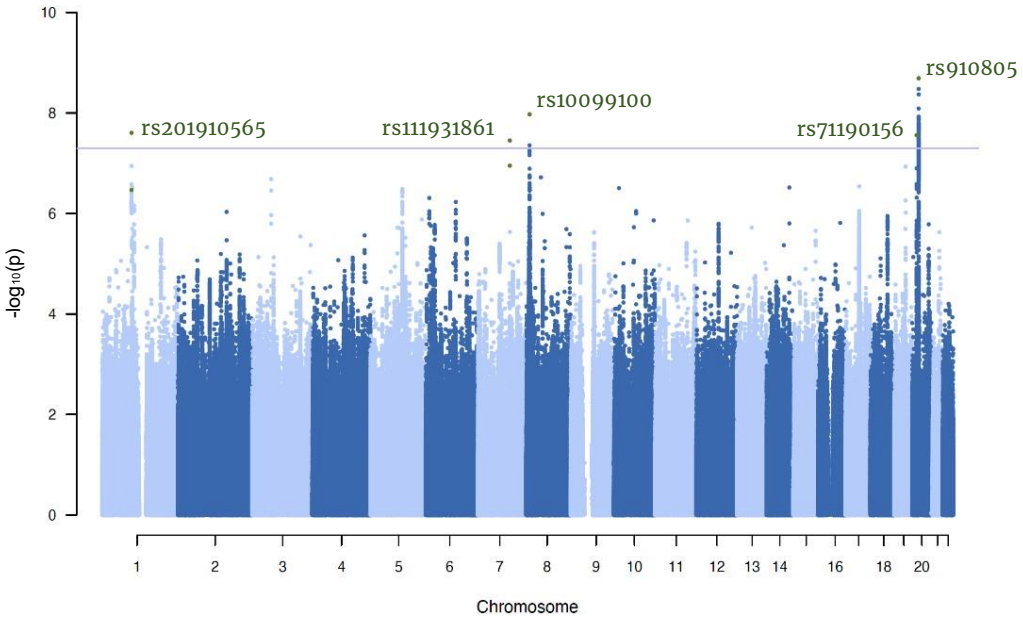


Figure 12. Manhattan plot for the latest ASD GWAS meta-analysis. The x-axis illustrates genomic position (chromosomes 1–22), while the y-axis represents statistical significance as -$\log_{10}$ (p-value). The purple horizontal line represents the genome-wide significance threshold of p-value = 5×10^{-8}. Only summary statistics for the primary ASD scan results (18,381 cases and 27,969 controls) are displayed. Three SNPs reached genome-wide

significance in the primary scan (rs10099100, rs910805 and rs71190156) and two others in the follow-up analysis (rs201910565 and rs111931861), as highlighted in green. This figure was re-created from Grove *et al.*[37] with an adapted script from the R package qqman[276] and public access data downloaded from the PGC website (see Methods), which solely included GWAS summary statistics for the primary scan. Data for the two SPNs reaching genome-wide significance in the follow-up analysis were added manually.

Therefore, ASD exhibit comparable discovery rate trends to MD or ADHD, despite the recruitment of considerably bigger cohorts in the latter two (Figure 11). This underscores the imperative to expand sample sizes.

Nonetheless, while sample sizes play a crucial role in the statistical power of a study, there are many other factors that can contribute to the lack of significant associations in ASD, as discussed in section 3.2.5.

5.1 OBJECTIVE

Besides the logical challenges expected due to high pleiotropy, polygenicity, and phenotypic heterogeneity, most of the signals from GWAS (> 90%) emerge from the non-coding genome, making the unraveling of functional implications not straightforward[277].

One plausible explanation is that these GWAS risk loci may exert their effect by influencing gene expression in different tissues[278]. Providing evidence that a SNP is an expression quantitative trait locus (eQTL), meaning it influences variation in gene expression levels, can enhance our ability to elucidate the nature of the mechanism driving associations.

While only a few studies have explored eQTLs associated with specific genes implicated in ASD[129,279], a recent investigation has identified over 30 eQTL-associated DNA variants with significantly different allele distributions between ASD-affected individuals and controls[280]. Therefore, the primary aim of this chapter is to identify signals where eQTLs colocalize with the ASD GWAS signals. Colocalization (*i.e.*, the correlation between GWAS signals and eQTLs) allows to refine the identification of potentially causal genes with a functional effect within genomic areas (including multiple genes and associated SNPs) associated with the disease by GWAS analyses.

In this context, we have further mined the largest ASD GWAS[37], as well as previous GBA and TWAS[281], altogether with publicly available eQTL data on 49 different tissues (Genotype-Tissue Expression (GTEx) Project v8)[282], by using eQTpLot[283], a newly developed tool that illustrates the colocalization of GWAS and eQTL signals in a locus, and the enrichment of and correlation between the candidate gene eQTLs and trait-significant variants. In summary, GWAS association signals may be located proximate to multiple potential candidate genes. To discern which gene is being regulated, we compared the association patterns of GWAS signals and eQTLs within a locus, aiming to ascertain whether the same genetic signal underlies gene expression and disease susceptibility (Figure 13).

Additionally, we sought to leverage results from other GWAS studies of genetically correlated NDDs to increase sample size and enhance our power in detecting colocalization.

Moreover, while numerous eQTLs are shared across tissues, certain ones exhibit tissue-specific impacts, particularly with consistent patterns of similarity in eQTL regulation among

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

different brain tissues[115]. As a result, we decided to employ tissue-specific eQTL datasets from GTEx with the aim of concentrating on tissues relevant to the disease.

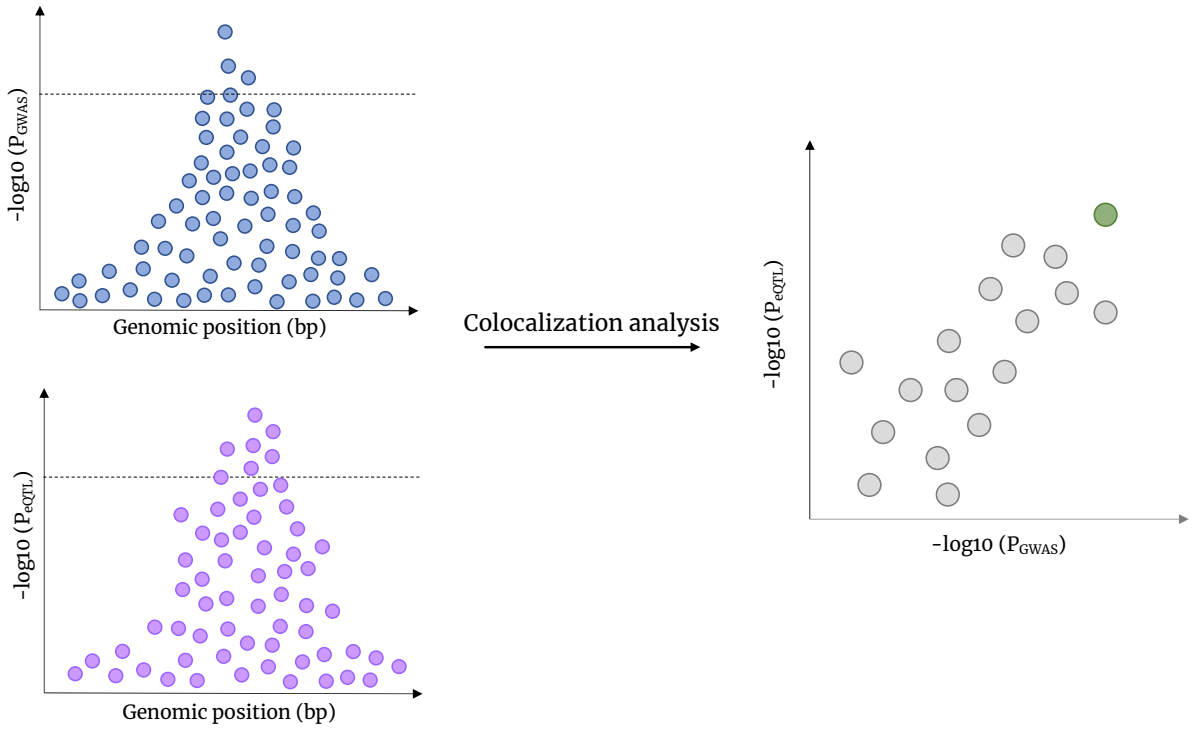


Figure 13. eQTL colocalization analysis. The lead SNP identified in a GWAS may be situated close to multiple potential candidate genes, known as the nearest genes. Determining which gene is under regulation can be achieved through eQTL colocalization analysis, which compares the association patterns of GWAS (top-left) and eQTLs (bottom-left) within a locus to ascertain whether the same genetic signal underlies gene expression and disease susceptibility. Robust colocalization patterns (right) should manifest as a distinct spike toward the upper-right corner of the P-P plot (*i.e.*, a positive and significant correlation). A bona fide signal (*i.e.* a single-causal SNP that influences the trait by regulating gene expression) should be located in the upper-right-most corner of the P-P plot (highlighted in green).

5.2 Materials and methods

5.2.1 Samples and collection

5.2.1.1 Discovery set

Summary statistics from the latest ASD GWAS meta-analysis were obtained from the public repository available in the PGC website (http://www.med.unc.edu/pgc/results-and-downloads). The following dataset, iPSYCH_PGC_ASD_Nov2017.gz, was employed. This dataset includes the meta-analysis of ASD by the Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH) and the PGC released in November 2017 (18,381 cases and 27,969 controls) (Table 9). Additional information about the genotyping and quality control (QC) methods employed are available at the PGC website.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Table 9. Characterization of the cohorts included in the colocalization analysis.

Study name	Trait	Sample size		Design	Ancestry	Reference
		Cases	Controls			
iPSYCH	ASD	13,076	22,664	Case/Control	European	Grove *et al.*, 2019[37]
PGC	ASD	7,387	11,359	Trios[a]	European	Grove *et al.*, 2019[37]
-	ADHD	20,183	35,191	Case/Control	European	Demontis *et al.*, 2019[260]
-	MD	246,363	561,190	Case/Control	European	Howard *et al.*, 2019[284]
INDEPENDENT PGC	SZP	40,675	64,643	Case/Control	European	Pardiñas *et al.*, 2018[261]
CLOZUK+PGC	SZP	36,989	113,075	Case/Control	European	SWGPGC, 2014[269]

[a] For trio designs, the control individuals are pseudocontrols generated from non-transmitted alleles. *SWGPGC, Schizophrenia Working Group of the Psychiatric Genomics Consortium.*

eQTL data were obtained from GTEx v8, which examines 15,201 RNA-Seq samples from 49 tissues of 838 postmortem donors[282]. GTEx v8 data were obtained on February 3, 2022, from the GTEx Portal: GTEx_Analysis_v8_eQTL.tar. Files containing all significant gene-variant pairs were selected for the analysis. These files include information for every variant with p-values below the gene-level threshold (calculated based on the minimum p-value distribution from permutations for each gene), considered as significant.

Initial data for allele count across the European population were obtained from the 1000 Genomes project, an integrated variant call set with phased genotype calls on 1092 samples for SNPs, short indels and large deletions (phase 1 release, v3_20101123; http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/). Files for each chromosome were converted to BED (Browser Extensible Data) format and then r^2 (squared correlations between SNPs' allele counts) was calculated (PLINK version 1.9[285]). These data were used as the pairwise variant LD information.

5.2.1.2 Neurodevelopmental Disorders (NDDs) sets

The largest cross-disorder GWAS meta-analysis of neuropsychiatric disorders to date (comprising more than 725,000 cases and controls of anorexia nervosa, ADHD, ASD, bipolar disorders (BD), MD, obsessive-compulsive disorder (OCD), SZP, and Tourette syndrome), allowed for the identification of 35 novel loci. Moreover, 109 genetic loci were discovered to have an impact on two or more disorders, with nearly 75% of the 146 genome-wide significant SNPs influencing multiple conditions[286].

Given the close genetic and phenotypic relationships between ASD, SZP, MD and ADHD[37], follow-up analyses were performed using GWAS summary statistics from the main PGC studies that are available for download, with altogether 1 study for MD, 2 for SZP and 1 for ADHD (Table 9). Studies for each of the diseases were selected based on sample sizes, with a focus on avoiding studies with overlapping cohorts (*i.e.*, when two studies include partial overlapping of samples, we selected the study with the biggest cohort, thereby maximizing detection power).

5.2.1.3 Data harmonization

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Summary statistics included the aggregate p-values and association data for every variant analyzed in the GWAS: (i) the rsID number, a unique label ("rs" followed by a number) used by researchers and databases to identify a specific SNP, (ii) genomic coordinates, indicated as chromosome and base position, (iii) effect allele, (iv) alternate allele, (v) OR, a statistical measure used to assess the association between a particular genetic variant and a specific trait or disease (*i.e.*, its effect size; how much more likely or less likely individuals with a certain genetic variant are to develop the disease compared to those without the variant), (vi) p-value, and (vii) standard error, among other parameters.

OR was transformed into beta (β) values (an alternative representation of the effect size of the allele, preferred for subsequent analyses) by applying the natural logarithm. For summary statistics files lacking the rsID column, SNPs' genomic coordinates were cross-referenced with the data from 1000 GENOMES phase 3, a catalogue of common human genetic variation (https://ftp.ensembl.org/pub/current_variation/vcf/homo_sapiens/1000GENOMES-phase_3.vcf.gz) and each SNP was assigned a unique ID.

For eQTL data, a unique file was generated for all tissues, with an additional column containing the tissue name to facilitate the selection of specific tissues in the subsequent analyses.

All data underwent processing in R version 4.0.0.

5.2.2 eQTL (expression quantitative trait locus) analyses

5.2.2.1 Selecting ASD associated genes from GWAS, GBA and TWAS

For the upcoming analyses, we have curated a set of genes from prior ASD studies, as detailed in Table 10. These studies aimed to leverage recent insights into ASD-associated loci, primarily utilizing GBA and TWAS methodologies. In order to facilitate the "candidate" gene designation for thorough testing (input genes for eQTpLot), we find it beneficial to explore these additional analyses that leverage the same initial dataset.

For each ASD-associated gene selected, the analyses will incorporate summary statistics obtained from the ASD GWAS meta-analysis and eQTLs from GTEx.

This comprehensive approach enables an exploration of the intersection between eQTLs and loci associated with ASD. Consequently, it facilitates the identification of candidate genes for subsequent in-depth testing of eQTL colocalization.

Table 10. ASD associated genes collected from bibliography and selected for eQTL colocalization study.

Genes	Genomic Region (Index variant)	Chr	Study	Reference
KIZ, XRN2, NKX2–2, NKX2–4	rs910805	20	GWAS	Grove *et al.*, 2019[7]
C8orf74, SOX7, PINX1	rs10099100	8		
LOC102723661, PTBP2	rs201910565	1		
MACROD2	rs71190156	20		
KMT2E, SRPK2	rs111931861	7		
XRN2, KIZ	rs910805	20	GBA (MAGMA)	Grove *et al.*, 2019[7]
KCNN2	rs13188074	5		
KANSL1, MAPT	rs142920272	17		
MACROD2	rs71190156	20		

WNT3	s146122400	17		
MFHAS1	rs17701675	8		
XKR6	rs6992403	8		
MSRA	rs7831557	8		
CRHR1	chr17:43965129	17		
SOX7	rs10099100	8		
NTM	rs529507	11		
MMP12	chr11:102751102	11		
BLK	rs2736342	8		
XRN2, NKX2-4, KIZ, NKX2-2	rs910805	20	GBA (PASCAL)	Alonso-Gonzalez *et al.*, 2019[112]
KCNN2	rs13188074	5		
CRHR1-IT1, LOC644172	rs142920272	17		
C8orf74	rs10099100	8		
PINX1	rs10099100	8	TWAS	Rodríguez-Fontenla
NKX2-2, NKX2-4	rs910805	20	(UTMOST)	*et al.*, 2021[281]
PTPRE	rs227375	10		
MANBA	-	4		
ERI1	-	8		
MITF	-	3		
CIPC	-	14		

Underlined gene names were associated at the GI level.

The selection process for the first 12 genes involved proximity to the five SNPs that achieved genome-wide significance in the ASD GWAS conducted by Grove and colleagues[37]. Specifically, genes were chosen if they were within 50 kb of the region covered by all SNPs with $r^2 \geq 0.6$ to the index variant.

In their study, Grove and colleagues also conducted a GBA using MAGMA[111]. MAGMA takes GWAS summary statistics as input and associates genetic variants with genes rather than individual markers. It considers the aggregate effect of multiple markers within a gene, aiming to identify genes or genomic regions that collectively contribute to the observed trait or disease in the study. This approach identified 15 associated genes surpassing the significance threshold, all of which are included in the subsequent analysis. The majority of these genes were located near the genome-wide significant SNPs identified in the GWAS, but 7 genes are located in four additional genomic regions and a cluster of genes is harbored on chromosome 17 (*KANSL1*, *WNT3*, *MAPT* and *CRHR1*).

Another GBA further mining the same GWAS data but leveraging a novel GBA tool, namely, PASCAL[112], highlighted additional genes which are here considered (*CRHR1-IT1* and *LOC644172*), although most of them were previously reported by MAGMA.

Lastly, Rodriguez-Fontenla and colleagues integrated tissue-expression and genetic data from Grove's ASD GWAS through a TWAS. This analysis revealed the association of four genes at the brain level (*CIPC*, *PINX1*, *NKX2-2* and *PTPRE*) and emphasized the association of *NKX2-2*, *MANBA*, *ERI1* and *MITF* at the GI level. However, due to lack of eQTL data to raise statistical power in the genes associated at the GI level, this study will not consider GI tissues.

USC UNIVERSIDADE DE SANTIAGO

5.2.3 eQTpLot: visualization of colocalization between eQTL and GWAS data

eQTpLot[283] is a R package designed to visualize the colocalization between eQTL and GWAS data (https://github.com/RitchieLab/eQTpLot). This tool takes as input GWAS and eQTL summary statistics, along with optional pairwise LD information. It then generates a series of plots that illustrate the colocalization (between GWAS and eQTL signals), correlation (between GWAS and eQTL p-values), and enrichment between eQTL and GWAS signals for a given gene-trait pair, that is, the enrichment of eQTLs among trait-significant variants.

Moreover, it allows to visualize the LD landscape of the locus in question, and the relationship between the direction of effect of eQTL signals and the direction of effect of colocalizing GWAS peaks.

5.2.3.1 Input data and significance thresholds

We ran eQTpLot using as input ASD GWAS summary statistics data from the PGC and 5 available studies aforementioned (p-values for the SNPs from GWAS analysis) and GTEx (expression data including p-values for the SNPs from the eQTL analysis) (see section above).

eQTpLot requires setting different arguments, as the significance threshold for eQTL data and the significance threshold for GWAS data.

We defined a variant as an eQTL for the candidate gene if its gene expression p-value was ≤ 0.05 (significance threshold for eQTL, default option in eQTpLot); variants with p-value > 0.05 were excluded from the analysis. In order to identify the list of all significant variant-gene pairs in the GTEx eQTL dataset, a genome-wide empirical p-value threshold, pt, was defined as the empirical p-value of the gene closest to the 0.05 false discovery rate (FDR) threshold. pt was then used to calculate a nominal p-value threshold for each gene[287,288]. Variants with a nominal p-value below the gene-level threshold were considered significant and included in the final list of variant-gene pairs.

The significance threshold for GWAS data was set to p-value $<5x10^{-8}$ for colocalization analysis (it corresponds to the horizontal line in plots and GWAS significance thresholds for the eQTL enrichment plot, default option in eQTpLot). The p-value $<5x10^{-8}$ corresponds to the most commonly accepted threshold in GWAS analyses, which is based on performing a Bonferroni correction for all the independent common SNPs across the human genome.

5.2.4 eQTL-GWAS colocalization plots

5.2.4.1 Multi and pan-tissue analysis

In some instances, it might be of interest to visualize how a genetic variant affects the expression of candidate genes across multiple types of tissues or even across all available tissues. Such analyses can be accomplished by setting the argument tissue to a list of specific tissues within the eQTL dataframe for a multi-tissue analysis. Alternatively, in a pan-tissue analysis, eQTL data from all the tissues in the eQTL dataframe will be combined, on a per-variant basis, to create a single pan-tissue eQTL. This is achieved by setting "tissue" argument to "all". The multi-tissue analysis follows a similar approach but consolidates eQTL data for the variant only across the specified tissues.

In order to visualize the variants' effect on candidate gene expression across all tissues, we performed pan-tissue analysis by setting the argument tissue to "all". In order to visualize its

USC COMPOSTELA

effect across multiple tissues contained in a specified list of tissues, we performed the multi-tissue analysis focusing on brain tissues from GTEx ("Brain_Amygdala", "Brain_Anterior_cingulate_cortex_BA24", "Brain_Caudate_basal_ganglia", "Brain_Cerebellar_Hemisphere", "Brain_Cerebellum", "Brain_Cortex", "Brain_Frontal_Cortex_BA9", "Brain_Hippocampus", "Brain_Hypothalamus", "Brain_Nucleus_accumbens_basal_ganglia", "Brain_Putamen_basal_ganglia", "Brain_Spinal_cord_cervical_c-1" and "Brain_Substantia_nigra").

The analyzed phenotype corresponded to the disease in the GWAS dataset, where we tested colocalization in each case. The assembly of the human genome was GRCh37 (genome reference consortium human build 37; hg19) in all instances.

eQTL-GWAS colocalization plots were created defining a locus of interest (LOI) for each of the selected genes to include the target gene's coordinates along with ± 1000 kb of flanking genome (specified with the argument range), so that GWAS data (summary statistics) are filtered out to include only those variants which fall within the LOI. The mapping window was defined as 1 megabase (Mb) on either side of the gene of interest, based on the consideration that only cis-eQTLs (associated with expression of a gene within a 1 Mb distance[289]) were included in our analysis. The significance threshold was set to 5×10^{-8} for colocalization analysis (red line in colocalization plots).

5.2.4.2 Congruous and incongruous variants

An additional parameter, "congruence", was set to "TRUE" in order to divide variants into two groups: congruous (variants with the same direction of effect on gene expression and the GWAS trait; *e.g.*, a variant that is significantly associated with a target's gene increased expression also results in an increase in the GWAS trait) and incongruous (variants with opposite effects on gene expression and the GWAS trait, *e.g.*, a variant that is significantly associated with a target's gene increased expression but results in a decrease in the GWAS trait).

5.2.4.3 eQTL enrichment plots

eQTL enrichment plots were generated for the selected LOI to test if there is a significant enrichment (p-value by Fisher's exact test) for eQTLs among GWAS-significant variants, so that each variant in the chromosomal space defined inside the LOI is represented along the horizontal axis, with the inverse log of the p-value of association with the specified GWAS trait on the vertical axis.

5.2.4.4 p-p correlation plots

So as to analyze the correlation between GWAS and eQTL's p-value, we plotted a best fit linear regression over the points, so that each of the variants within the LOI is plotted with -log10 (p-value eQTL) along the horizontal axis and -log10 (p-value GWAS) along the vertical axis. The Pearson correlation coefficient (r) and p-value of correlation (p) are also computed and displayed in the plot, differentiating between variants with congruous and incongruous

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

direction of effects. Significant correlations are defined as having a $r \geq 0.7$ and p-value of correlation ≤ 0.05.

5.2.5 Gene expression, gene network and related human disease-associated genes analysis

5.2.5.1 Gene expression heatmap

GENE2FUNC, a tool of FUMA (Functional Mapping and Annotation of Genome Wide Association Studies) GWAS (https://fuma.ctglab.nl/)[290] was employed to carry out a gene expression heatmap with default settings.

Those genes in which eQTL and GWAS signals colocalize (Table 11), were used as the input.

Expression values are RPKM (read per kb per million) for BrainSpan. Heatmaps display the normalized expression value (zero mean normalization of log2 transformed expression), and darker red means higher relative expression of that gene in each label (developmental stage), compared to a darker blue color in the same label.

5.2.5.2 Gene network analysis

HumanBase (https://hb.flatironinstitute.org/) was used to build a gene network for genes yielding significant colocalization in brain tissue (selected by the multi-tissue eQTL colocalization analysis). HumanBase serves as a comprehensive resource for biological research and offers data-driven predictions related to gene expression, function, regulation, and interactions within the human domain, with a particular focus on specific cell types, tissues, and diseases.

In order to capture tissue-specific gene function we used the "tissue specific gene networks: GIANT" analysis tool. The HumanBase GIANT analysis tool, constructs comprehensive genome-scale functional maps for various human tissues by integrating extensive datasets from over 14,000 distinct publications, covering thousands of experiments. The platform automatically evaluates the relevance of each dataset to 144 tissue- and cell lineage–specific functional contexts. The resulting functional gene maps offer detailed insights into protein function and interactions in specific human tissues and cell lineages.

MAPT, *NKX2-2*, and *PTPRE* were selected as the input genes along with brain tissue in the 5 existing data types (co-expression, TF binding, interaction, GSEA (Gene Set Enrichment Analysis) microRNA targets, GSEA perturbations). The resultant network contains the subset of functionally related genes (Appendix 10.2, Supplementary Table 1, n = 50), all of which were used to test for functional enrichment using genes annotated to Gene Ontology (GO) biological processes (BP) (see Chapter 2 for further reference).

Within each cluster, representative processes and pathways that showed enrichment were identified and presented along with their corresponding q-values. The q-value of each term is calculated using one-sided Fisher's exact tests and Benjamini–Hochberg corrections to correct for multiple tests. We also carried out an enrichment analysis in DIsGeNET (https://metascape.org), containing a large collections of genes and variants associated with human diseases (429,036 gene-disease associations, linking 17,381 genes to 15,093 diseases

USC DE COMPOSTELA

and 72,870 variant-disease associations, between 46,589 SNPs and 6,356 diseases and phenotypes)[291].

5.3 RESULTS

Initial steps from this analysis included eQTL colocalization analysis with ASD GWAS data from the total of 27 LOI here interrogated (Table 11).

We found 8 genes in which eQTL and GWAS signals colocalize ($r \geq 0.7$, p-value ≤ 0.05): *CRHR1*, *KANSL1*, *MANBA*, *MAPT*, *MMP12*, *NKX2-2*, *PTPRE* and *WNT3*.

Although not reaching the positive correlation coefficient threshold of $r \geq 0.7$, *SRPK2* yields a marginal positive correlation with strong statistical support ($r = 0.69$, $p < 1x10^{-6}$).

Moreover, we addressed whether the SNPs-eQTLs within these LOI yield a colocalization signal when analyzed in the multi-tissue analysis and found that *MAPT* (r = 0.76 incongruent variants), *NKX2-2* (r = 0.71 incongruent variants), and *PTPRE* (r = 0.70 congruent variants and r = 0.97 incongruent variants) exhibit a strong correlation when restricting the eQTL analysis to brain tissue.

Table 11. eQTL colocalization analysis results for the 27 LOI in ASD.

Gene	Pan-tissue Analysis				Multi-tissue Analysis			
	Congruent		Incongruent		Congruent		Incongruent	
	r	p-value	r	p-value	r	p-value	r	p-value
BLK	0.46	$<1 \times 10^{-6}$	0.48	$<1 \times 10^{-6}$	0.22	7.38×10^{-1}	0.64	2.49×10^{-1}
C8orf74	0.39	5.09×10^{-2}	0.43	3.29×10^{-3}	**	**	**	**
CIPC	0.48	2.35×10^{-6}	0.41	1.12×10^{-4}	0.69	1.43×10^{-6}	0.41	5.07×10^{-3}
CRHR1	0.96	$<1 \times 10^{-6}$	0.96	$<1 \times 10^{-6}$	-0.52	$<1 \times 10^{-6}$	-0.50	$<1 \times 10^{-6}$
ERI1	0.58	$<1 \times 10^{-6}$	0.62	$<1 \times 10^{-6}$	-0.21	1.40×10^{-2}	-0.25	3.93×10^{-3}
KANSL1	0.82	$<1 \times 10^{-6}$	0.85	$<1 \times 10^{-6}$	-0.20	$<1 \times 10^{-6}$	-0.07	8.05×10^{-3}
KCNN2	-0.08	2.33×10^{-1}	-0.15	1.96×10^{-2}	*	*	*	*
KIZ	0.47	$<1 \times 10^{-6}$	0.53	$<1 \times 10^{-6}$	0.22	1.49×10^{-1}	-0.38	9.94×10^{-3}
KMT2E	0.27	8.58×10^{-3}	0.12	2.85×10^{-1}	*	*	*	*
MACROD2	-0.02	8.46×10^{-1}	-0.01	9.39×10^{-1}	*	*	*	*
MANBA	0.62	$<1 \times 10^{-6}$	0.70	$<1 \times 10^{-6}$	0.02	7.61×10^{-1}	0.00	9.87×10^{-1}
MAPT	0.96	$<1 \times 10^{-6}$	0.96	$<1 \times 10^{-6}$	0.69	$<1 \times 10^{-6}$	0.76	$<1 \times 10^{-6}$
MFHAS1	0.2	$<1 \times 10^{-6}$	0.46	$<1 \times 10^{-6}$	*	*	*	*
MITF	0.46	$<1 \times 10^{-6}$	-0.38	$<1 \times 10^{-6}$	***	***	***	***
MMP12	0.72	2.14×10^{-4}	0.56	1.01×10^{-3}	**	**	**	**
MSRA	0.27	$<1 \times 10^{-6}$	0.23	$<1 \times 10^{-6}$	0.01	9.18×10^{-1}	0.2	1.47×10^{-2}
NKX2-2	-0.36	2.25×10^{-1}	0.71	2.26×10^{-2}	-0.36	2.25×10^{-1}	0.71	2.26×10^{-2}
NKX2-4	**	**	**	**	**	**	**	**
NTM	0.30	$<1 \times 10^{-6}$	0.37	$<1 \times 10^{-6}$	0.12	3.96×10^{-1}	0.06	7.82×10^{-1}
PINX1	0.42	9.56×10^{-5}	0.57	$<1 \times 10^{-6}$	0.26	8.07×10^{-2}	0.48	6.49×10^{-4}
PTBP2	0.35	3.21×10^{-17}	-0.40	1.35×10^{-20}	0.17	1.26×10^{-1}	0.14	2.24×10^{-1}
PTPRE	0.02	8.48×10^{-1}	0.09	3.19×10^{-1}	0.70	1.59×10^{-3}	0.97	2.63×10^{-4}
SOX7	-0.35	$<1 \times 10^{-6}$	-0.37	$<1 \times 10^{-6}$	0.05	6.45×10^{-1}	-0.03	7.88×10^{-1}
SRPK2	0.69	$<1 \times 10^{-6}$	0.66	$<1 \times 10^{-6}$	*	*	*	*
WNT3	0.93	$<1 \times 10^{-6}$	0.94	$<1 \times 10^{-6}$	-0.15	6.82×10^{-2}	-0.24	1.39×10^{-2}
XKR6	-0.18	3.12×10^{-2}	-0.13	1.34×10^{-1}	***	***	***	***

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

XRN2	0.38	$<1 \times 10^{-6}$	0.44	$<1 \times 10^{-6}$	0.35	3.19×10^{-1}	-0.23	5.57×10^{-1}

Significant results ($r \geq 0.7$, p-value ≤ 0.05) are underlined. Boldfaced gene names indicate a significant correlation signal when the analysis is restricted to brain tissue. *Not enough data to generate P-P plot for congruent/incongruent eQTLs; **No data are available for the gene meeting the eQTL p-value significance threshold; ***Not enough finite observations.

For every gene exhibiting a significant association between ASD GWAS and eQTL's p-value ($r \geq 0.7$, p-value ≤ 0.05), we subsequently addressed colocalization in an additional set of GWAS studies for SZP, ADHD and MD, prompted by previously reported highly significant genetic correlations[37,292].

When addressed in the follow-up analysis, all genes, besides *NKX2-2*, were confirmed to exhibit a significant association between GWAS and eQTL p-values in at least one other correlated NDD. Remarkably, *CRHR1*, *KANSL1*, *MAPT*, *PTPRE* and *WNT3* yielded a significant correlation when restricted to brain tissue (Table 12).

Table 12. Colocalization results for potentially causal genes selected from the ASD eQTL colocalization analysis in correlated NDDs.

Gene	Pan-tissue Analysis				Multi-tissue Analysis				Trait
	Congruent		Incongruent		Congruent		Incongruent		
	r	p-value	r	p-value	r	p-value	r	p-value	
CRHR1	0.82	$<1 \times 10^{-6}$	0.97	4.53×10^{-6}	0.68	$<1 \times 10^{-6}$	0.97	1.50×10^{-6}	SZP[261]
	0.81	$<1 \times 10^{-6}$	0.82	$<1 \times 10^{-6}$	0.87	$<1 \times 10^{-6}$	0.73	$<1 \times 10^{-6}$	SZP[269]
	0.87	$<1 \times 10^{-6}$	0.89	$<1 \times 10^{-6}$	0.83	$<1 \times 10^{-6}$	0.89	$<1 \times 10^{-6}$	MD[284]
KANSL1	0.85	$<1 \times 10^{-6}$	0.14	2.92×10^{-2}	0.61	$<1 \times 10^{-6}$	0.18	1.28×10^{-2}	SZP[261]
	0.80	$<1 \times 10^{-6}$	0.84	$<1 \times 10^{-6}$	0.35	$<1 \times 10^{-6}$	0.73	$<1 \times 10^{-6}$	SZP[269]
MANBA	0.81	$<1 \times 10^{-6}$	0.84	$<1 \times 10^{-6}$	0.29	1.78×10^{-5}	0.30	2.02×10^{-5}	MD[284]
MAPT	0.87	$<1 \times 10^{-6}$	0.36	5.29×10^{-3}	0.17	1.76×10^{-1}	0.45	$<1 \times 10^{-6}$	SZP[261]
	0.86	$<1 \times 10^{-6}$	0.87	$<1 \times 10^{-6}$	0.57	$<1 \times 10^{-6}$	0.49	$<1 \times 10^{-6}$	SZP[269]
	0.90	$<1 \times 10^{-6}$	0.89	$<1 \times 10^{-6}$	0.61	$<1 \times 10^{-6}$	0.70	$<1 \times 10^{-6}$	MD[284]
MMP12	0.75	8.94×10^{-5}	0.63	1.16×10^{-4}	*	*	*	*	ADHD[260]
	0.88	$<1 \times 10^{-6}$	0.67	3.23×10^{-5}	*	*	*	*	MD[284]
PTPRE	0.20	4.57×10^{-2}	0.15	1.24×10^{-1}	0.34	2.87×10^{-1}	0.78	8.46×10^{-3}	ADHD[260]
	0.04	7.09×10^{-1}	0.09	3.64×10^{-1}	0.37	1.53×10^{-1}	0.94	6.15×10^{-4}	MD[284]
WNT3	-0.62	$<1 \times 10^{-6}$	-0.59	$<1 \times 10^{-6}$	-0.70	$<1 \times 10^{-6}$	-0.68	$<1 \times 10^{-6}$	ADHD[260]
	0.86	$<1 \times 10^{-6}$	0.41	3.62×10^{-6}	0.68	$<1 \times 10^{-6}$	0.21	4.37×10^{-2}	SZP[261]
	0.90	$<1 \times 10^{-6}$	0.93	$<1 \times 10^{-6}$	0.88	$<1 \times 10^{-6}$	0.91	$<1 \times 10^{-6}$	SZP[269]
	0.79	$<1 \times 10^{-6}$	0.81	$<1 \times 10^{-6}$	0.71	$<1 \times 10^{-6}$	0.82	$<1 \times 10^{-6}$	MD[284]

SWGPGC, Schizophrenia Working Group of the Psychiatric Genomics Consortium. * No data are available for the gene meeting the eQTL p-value significance threshold. Significant results are underlined. Boldfaced gene names indicate a significant correlation signal when the analysis is restricted to brain tissue. Only genes with reported colocalization in ASD and replicated colocalization in at least one other dataset are depicted. For each gene, only studies with at least one significant colocalization signal in the follow-up analysis are included.

5.3.1 Genes with ASD colocalization in the multi-tissue analysis

Out of the 8 genes with colocalizing signals in ASD, 3 of them yielded significant results in the multi-tissue analysis, namely: *PTPRE* (Protein Tyrosine Phosphatase Receptor Type E), *MAPT* (Microtubule Associated Protein Tau) and *NKX2-2* (NK2 Homeobox 2) (Table 11).

Of note, *NKX2-2*, while demonstrating its highest correlation signal in brain tissue (incongruent r = 0.71, p-value = 2.26 x 10^{-2}), is the only gene that was not replicated in any other NDD included in the analysis. If we were to perform multiple hypothesis testing corrections for 27 different hypotheses tested here (*i.e.*, the existence of a significant colocalizing signal in 27 LOI here tested), *NKX2-2* would not surpass the corrected p-value threshold (data not shown). Moreover, its association signal comes from a limited number of variants (Appendix 10.1, Supplementary Figure 1), so further analysis should be performed in order to detect a more reliable and robust colocalization signal.

MAPT displays a significant correlation in ASD in the multi-tissue analysis for incongruent variants (r = 0.76, p < 1 x 10^{-6}) (Figure 14) and follows the same trend in the multi-tissue analysis for MD (r = 0.70, p < 1 x 10^{-6}) (Table 12).

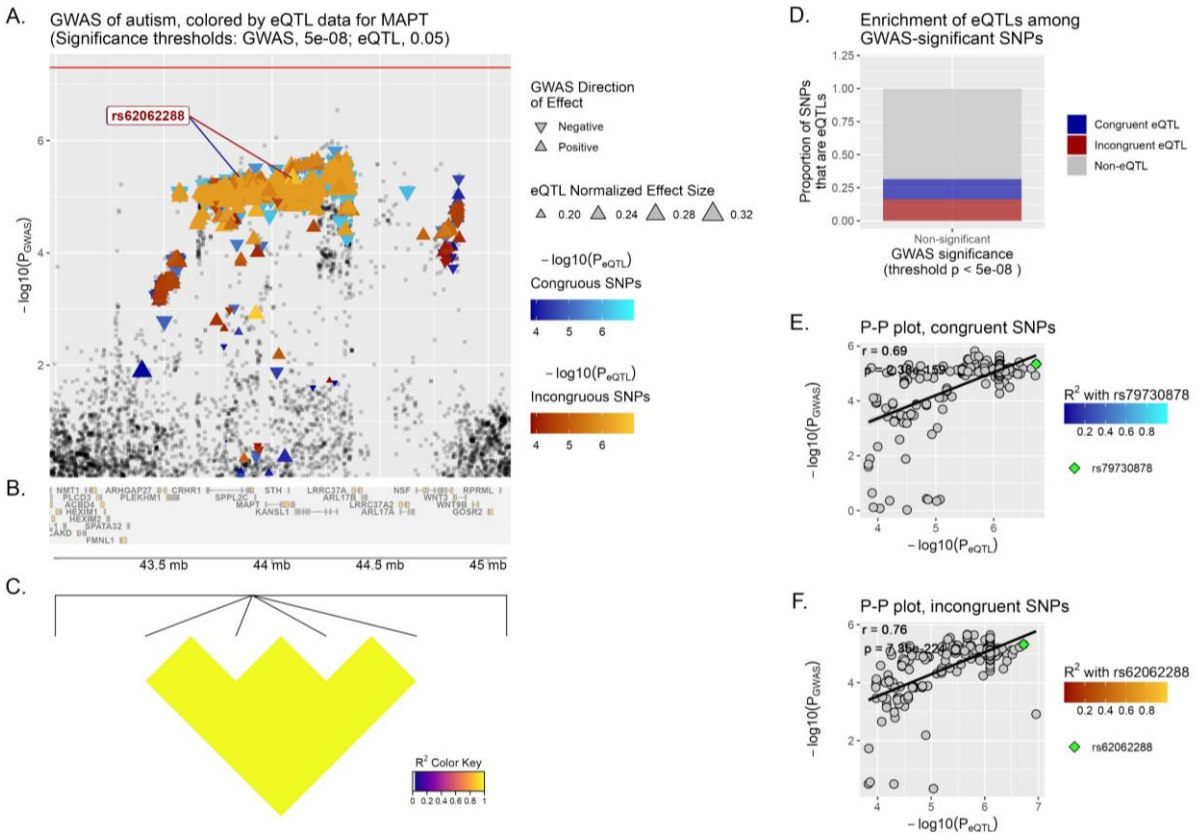


Figure 14. eQTpLot for ASD and *MAPT* multi-tissue analysis. Panel A shows the LOI, containing *MAPT*, with chromosomal space indicated along the horizontal axis. Within this plot, variants that lack eQTL data for the specified gene or do not meet the significance threshold are plotted as grey squares. On the opposite, variants that are determined to act as eQTLs for *MAPT* are plotted as colored triangles, with a color gradient corresponding to the inverse magnitude of its effect on gene expression or p-value. Congruous effects are plotted using a blue color scale, in contrast to variants having an incongruous effect, in red. The size of each triangle is proportional to the eQTL normalized effect size (NES), while the directionality of each triangle corresponds to the direction of effect of the variant on the GWAS trait. Horizontal red line marks the significance threshold in GWAS (p-value = 5x10^{-8}). Finally, we show at the bottom of the plot a depiction of all genes' genomic positions that fall inside the LOI (Panel B) and a heatmap of LD information of all *MAPT* eQTL variants (Panel C). Panel D depicts the enrichment

of *MAPT* eQTLs among GWAS-significant and/or non-significant variants, while panels E-F depict the correlation between p_{GWAS} and p_{eQTL} for *MAPT* and ASD, with the computed Pearson correlation coefficient (r) and p-value (p) displayed on the plot. The lead variants, rs79730878 and rs62062288, are identified in each graph (by default the upper-right-most variant on the P-P plot), with all other variants plotted using a color scale corresponding to their squared coefficient of linkage correlation with this lead variant (if available). For reference, the same lead variants are also labelled in panel A. In this case, as well as in other plots, both labels overlap. For the sake of preserving picture quality, we decided not to edit it. Lead variants can be checked in panels E-F.

PTPRE has yielded the strongest correlation in brain tissue for ASD incongruent variants (incongruent r = 0.97, p-value = 2.63 x 10^{-4}) (Figure 15).

Nevertheless, the association signal exhibits a degree of weakness, particularly concerning the limited number of variants employed in the analytical calculation and the notable low p_{GWAS} values. As such, we further tested this colocalization in the follow-up analysis and observed that, when examining this gene in other NDDs, significant correlations for the multi-tissue analysis were confirmed in both ADHD and MD (incongruent r = 0.78, p-value = 8.46 x 10^{-3}; incongruent r = 0.94, p-value = 6.15 x 10^{-4}, respectively) (Table 12).

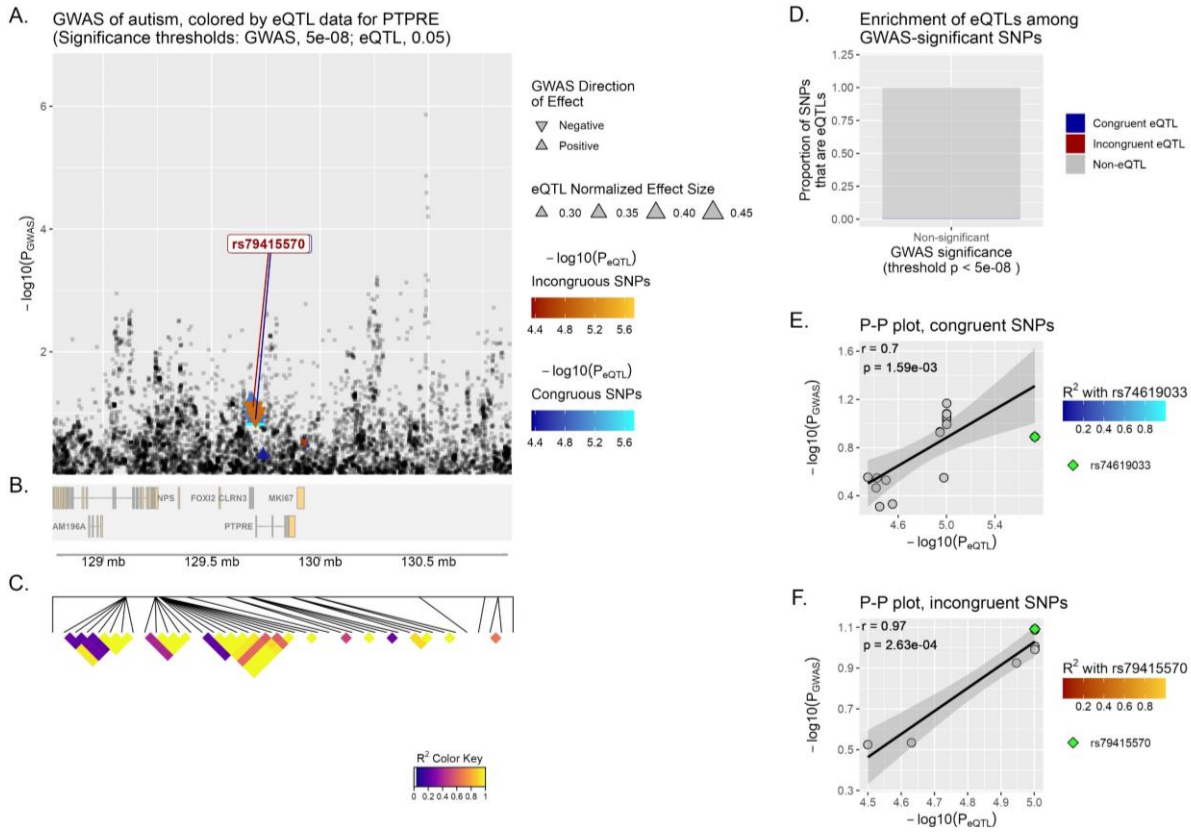


Figure 15. eQTpLot for ASD and *PTPRE* multi-tissue analysis. Please refer to the caption on Figure 14 to view the legend for this figure.

5.3.2 Genes with ASD colocalization in the pan-tissue analysis

Out of the 8 genes with colocalizing signals in ASD, 5 of them yielded significant results only in the pan-tissue analysis, namely: *CRHR1*, *KANSL1*, *MANBA*, *MMP12*, and *WNT3* (Table 11).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

However, when analyzed in the follow-up studies (Table 12), all genes besides *MANBA* and *MMP12* yielded significant colocalization signals in brain.

5.3.2.1 *CRHR1* (Corticotropin Releasing Hormone Receptor 1)

The available data are currently insufficient to draw definitive conclusions regarding the role of *CRHR1* in ASD related to brain tissue. However, in the pan-tissue analysis for ASD, this gene has exhibited significant colocalization of congruent and incongruent variants ($r = 0.96$, p-value $<1 \times 10^{-6}$) (Appendix 10.1, Supplementary Figure 2; Table 11). Similar trends were observed in both SZP and the MD datasets with congruent and incongruent variants yielding significant colocalization results in the pan-tissue analysis (Table 12).

Of note, *CRHR1* also achieved significance, in both SZP datasets and in MD, when restricting to brain tissue (Table 12).

5.3.2.2 *KANSL1* (KAT8 Regulatory NSL Complex Subunit 1)

KANSL1 has shown positive correlation in the pan-tissue analysis for ASD and congruent/incongruent variants (congruent $r = 0.82$, p-value $<1 \times 10^{-6}$; incongruent $r = 0.85$, p-value $<1 \times 10^{-6}$) (Appendix 10.1, Supplementary Figure 3; Table 11).

Moreover, thanks to the follow-up analysis, we could confirm an incongruent correlation for *KANSL1* in the general set of tissues (pan-tissue analysis) but also specific to brain (incongruent $r = 0.73$, p-value $<1 \times 10^{-6}$) in one of the SZP datasets[269] (Table 12).

5.3.2.1 *MANBA* (Mannosidase Beta A)

In the ASD pan-tissue analysis, incongruent variants for *MANBA* proof significantly correlated ($r = 0.70$, p-value $<1 \times 10^{-6}$) (Appendix 10.1, Supplementary Figure 4; Table 11), being these findings replicated in MD (congruent $r = 0.81$, p-value $<1 \times 10^{-6}$; incongruent $r = 0.84$, p-value $< 1 \times 10^{-6}$)

However, analyses restricted to brain tissue have not shown any significant correlation in any of the NDDs.

5.3.2.2 *MMP12* (Matrix Metallopeptidase 12)

MMP12 has exhibited a positive correlation with congruent variants in ASD in the pan-tissue analysis ($r = 0.72$, p-value $= 2.14 \times 10^{-4}$) (Appendix 10.1, Supplementary Figure 5; Table 11), and the same direction of effects was observed in ADHD ($r = 0.75$, p-value $= 8.94 \times 10^{-5}$) and MD ($r = 0.88$, p-value $< 1 \times 10^{-6}$) (Table 12).

The multi-tissue analysis for *MMP12* in these two NDDs could not be performed because no variant met the eQTL p-value significance threshold in any brain tissue (Table 12).

5.3.2.3 *WNT3* (Wnt Family Member 3)

For *WNT3*, the correlation signal in ASD comes from congruent/incongruent variants in the entire set of tissues (congruent $r = 0.93$, p-value $< 1 \times 10^{-6}$; incongruent $r = 0.94$, p-value $<1 \times 10^{-6}$) (Appendix 10.1, Supplementary Figure 6; Table 11).

When colocalization was analyzed in other NDDs, the MD and SZP datasets resulted in a significant signal for the pan-tissue analysis, both for congruent and for incongruent variants.

However, the strongest correlation signal was observed in brain tissue for incongruous variants in one SZP dataset (r = 0.91, p-value <1 x 10^{-6})[269].

5.3.3 Gene expression, gene networks and related human disease-associated genes

Finally, we aimed to assess the potential functional relevance in terms of disease risk for those genes highlighted by having a significant colocalization association in ASD.

With this regard, we first performed expression analysis using BrainSpan RNA-Seq data for the set of genes selected by eQTL analysis. Notably, expression analyses have shown overexpression of *MAPT* across every single developmental stage in comparison with the remaining studied genes, and expression of *KANSL1* restricted to relevant periods of ASD risk (early-to-late prenatal) (Figure 16 A, B).

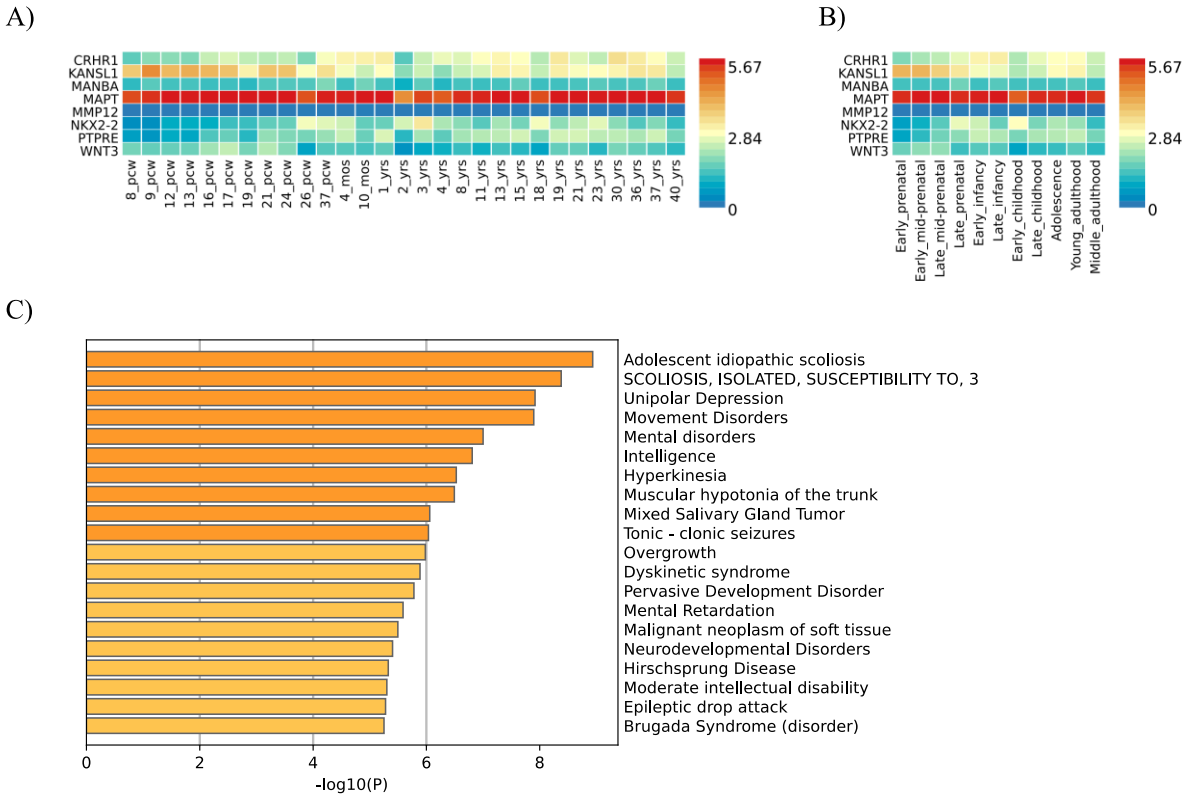


Figure 16. Expression and disease enrichment analysis. Heatmaps for BrainSpan (A) 29 different ages of brain samples and (B) 11 general developmental stages of brain samples. (C) Disease enriched terms are represented in darker orange. These figures were created in the GTEx and DisGeNET portals[115,293].

Moreover, *MAPT, NKX2-2* and *PTPRE* (genes with a significant colocalization signal in the multi-tissue analysis) and their interactors in brain tissue (Appendix 10.2, Supplementary Table 1) show functional enrichment in processes associated with neurogenesis, neuron differentiation and nervous system functioning and development (Table 13).

Furthermore, these genes underscore diverse human diseases, offering significant insights into implications for unipolar disorders, mental disorders, tonic-clonic seizures, and intelligence, that have been associated with ASD, as illustrated in Figure 16 C.

Table 13. Functional enrichment for *MAPT, NKX2-2, PTPRE* and their interactors in brain tissue.

Name	Database	q-value
peptidyl-threonine phosphorylation	Gene Ontology (BP)	0.014
peptidyl-threonine modification	Gene Ontology (BP)	0.014
neuron differentiation	Gene Ontology (BP)	0.014
regulation of establishment of protein localization to mitochondrion	Gene Ontology (BP)	0.014
generation of neurons	Gene Ontology (BP)	0.014
neurogenesis	Gene Ontology (BP)	0.016
regulation of neuron differentiation	Gene Ontology (BP)	0.028
regulation of membrane potential	Gene Ontology (BP)	0.036
regulation of neurogenesis	Gene Ontology (BP)	0.036
negative regulation of neuron differentiation	Gene Ontology (BP)	0.036
regulation of nervous system development	Gene Ontology (BP)	0.036
negative regulation of transferase activity	Gene Ontology (BP)	0.036
negative regulation of neurogenesis	Gene Ontology (BP)	0.039
negative regulation of nervous system development	Gene Ontology (BP)	0.040
establishment of protein localization to mitochondrion	Gene Ontology (BP)	0.044
protein localization to mitochondrion	Gene Ontology (BP)	0.044
positive regulation of transmembrane receptor protein serine/threonine kinase signaling pathway	Gene Ontology (BP)	0.044
regulation of cell development	Gene Ontology (BP)	0.047

5.4 DISCUSSION

One of the main difficulties when interpreting GWAS association signals is to identify the disease susceptibility genes underlying these associations, as the vast majority (> 90%) fall within non-coding regions of the genome. Here, we show how approaches such as eQTL colocalization may help leverage GWAS data by connecting disease associated loci with underlying mechanisms (*i.e.*, the contribution to variation in gene expression level), being of service to determine which gene is most probably affected in specific tissues and to identify potential biological pathways.

5.4.1 Colocalization in brain tissues

In this analysis, we performed an eQTL colocalization analysis using the GTEx expression data and the largest ASD GWAS to date in order to identify potentially causal genes. Using this approach, we identified a total of eight genes with a significant colocalizing eQTL signal (*CRHR1*, *KANSL1*, *MANBA*, *MAPT*, *MMP12*, *NKX2-2*, *PTPRE* and *WNT3*), with three of them (*MAPT*, *NKX2-2*, *PTPRE*) yielding a significant signal when restricted to brain tissue (Table 11). Moreover, *SRPK2* achieved a high coefficient of correlation with nominal statistical support, barely reaching the defined threshold of $r \geq 0.7$.

Grove *et al.*[37] found an intronic SNP in the *MAPT* gene (rs141455452), although not reaching the genome-wide significance association threshold (p-value = 9.00×10^{-7}). *MAPT* shows ubiquitous expression across all brain ages and developmental stages (Figure 16 A, B) and it is differentially expressed in the nervous system, depending on the stage of neuronal maturation and neuron type[294]. Mutations in *MAPT* have been linked to various

neurodegenerative conditions, including AD, frontotemporal dementia, cortico-basal degeneration, and progressive supranuclear palsy[295], and its functional interactors in brain tissue have shown enrichments for diseases relevant to ASD (Figure 16 C). Considering the robust genetic correlation between AD and SZP[296], our identification of a positive correlation in the SZP dataset (Table 12) for the colocalization of *MAPT* is not unexpected. Moreover, in another study, Huang *et al.* [297] further performed a TWAS of 7,805 ASD proband-parent trios, which was subsequently replicated using 35,740 independent samples, and identified *MAPT* as a gene with a significant transcriptome-wide association with ASD.

As of now, *MAPT* is classified as SFARI gene score 3, and was first associated with ASD in Grove *et al*. through a GBA[37]. In this context, we contribute further evidence of its association with ASD, supported by significant colocalization results, specifically in brain tissue (Figure 14).

In regard to *PTPRE,* Rodriguez-Fontenla *et al.*[281] demonstrated for the first time its association at the brain level through TWAS. It is worth noting that this specific gene demonstrates the highest correlation within brain tissue (r = 0.97) among all genes analyzed in this analysis (Figure 15).

PTPRE encodes the protein tyrosine phosphatase, receptor type E, a member of the Protein tyrosine phosphatases (PTPs), which represent a superfamily of enzymes that play essential roles in normal development and physiology[298]. It has low human brain regional specificity and has not, to date, been involved in any NDD. However, PTPs have been increasingly implicated in the control of neuronal synapse formation and function[299]. Moreover, another member of the PTP family, PTPRD (Protein Tyrosine Phosphatase Receptor Delta) has been genetically associated with NDDs and psychiatric diseases, including ASD, ADHD and SZP[300], in line with results from the colocalization analysis for *PTPRE* performed here. Overall, these findings support the idea that *PTPRE* may be acting through similar molecular mechanisms leading to disorder onset, although further studies that characterize the role of *PTPRE* in the brain are required to elucidate its role in ASD. Moreover, these findings warrant careful interpretation. Despite the heightened significance observed in the colocalization analysis, one should anticipate higher association p-values in the GWAS analysis. Furthermore, to establish a conclusive association, additional eQTL data are imperative to demonstrate the connection between variants associated with GWAS and *PTPRE* expression.

As a third relevant finding in the ASD multi-tissue analysis, we detected a positive correlation for *NKX2-2* in brain (Appendix 10.1, Supplementary Figure 1). *NKX2-2* was selected as being one of the nearest genes to a genome-wide significant locus detected by Grove *et al.* (index variant rs910805, p-value = 2.04×10^{-9})[37] (Table 7). This region spans three other genes (*KIZ*, *XRN2*, *NKX2–4*), but this analysis points to *NKX2-2* through eQTL colocalization, although we could not confirm these results in any other NDD, and these results should thus be interpreted with caution.

Moreover, *NKX2-2* was also associated by Rodriguez-Fontenla *et al.*[281] at the brain level, but also at the GI level. Our detected association was limited to the pan-tissue analysis due to

USC DE SANTIAGO DE COMPOSTELA

limited eQTL data for the gene in brain tissues. Another TWAS performed by Huang *et al.*[297] also showed significant association in brain nucleus accumbens basal ganglia. Of note, eQTL variants for *NKX2-2* in our dataset were only present in cerebellum, although the previous TWAS study found a significant association for *NKX2-2* in brain nucleus accumbens basal ganglia and colon transverse. Ideally, future eQTL datasets specific for brain tissue and/or GI tissues would help us establish more robust conclusions.

Overall, this shows how the utilization of both TWAS and eQTL colocalization methods provides a valuable strategy to prioritize potentially causal genes underlying GWAS associations. Also, GO enrichment analysis of the three prioritized genes in the multi-tissue analysis has revealed enrichments in various brain-relevant ontologies associated with ASD etiology, including neuron differentiation. Investigating enrichments in associated human diseases, we observed that these three genes, along with their co-expressed interactors in brain tissue, are enriched in several conditions such as mental retardation, neurodevelopmental disorders, and psychiatric disorders, among others. This integrated approach enhances our ability to identify genes that are likely to play a role in the observed associations and provides further insights into the biological mechanisms involved in complex traits and diseases. By considering the expression patterns of genes and their genetic regulation, we can better understand the functional relevance of the identified genes and their potential contributions to the GWAS findings. In this context, an association for *PTPRE* and *NKX2-2* cannot be definitively asserted due to the aforementioned limitations. However, promising results are evident for *MAPT*.

Although not reaching the correlation threshold for defining colocalization in this study ($r \geq 0.7$), *SRPK2* draws our attention since it has not been previously linked to ASD by any analyses, nor in GBA or TWAS.

SRPK2 is located within one of the 5 genome-wide associated loci from Grove *et al.* (index variant rs111931861, p-value = 3.53×10^{-8}), next to *KMT2E* (Table 7), which has been already associated with a spectrum of NDDs, including ASD[301]. *SRPK2* triggers cell cycle progression in neurons and induces apoptosis and neurodegeneration, which are one of the main hallmarks in ASD[302,303], suggesting that *SRPK2*'s functional characterization in ASD may be of high relevance. On top of that, *SRPK2* has been widely linked to AD by phosphorylating tau, a neuronal microtubule-associated protein, and inducing its polymerization, thus leading to suppression of tau-dependent microtubule polymerization, and inhibiting axonal elongation in neurons[302,304,305]. Depletion of *SRPK2* in dentate gyrus inhibits tau phosphorylation in APP/PS1 mouse (early-onset AD mouse model) and alleviates the impaired cognitive behaviors. This may be of special interest since recent studies have demonstrated that tau also enables ASD-like behaviors and that even partial reduction of this protein prevents such behaviors and related neural abnormalities in independent mouse models[306]. Of note, the detected signal comes from congruous variants (*i.e.*, variants that increase gene expression and also result in an increase in the GWAS trait) (Table 11), which aligns with the aforementioned evidence and suggests a possible role for *SRPK2* in ASD, which should be further tested.

USC DE SANTIAGO DE COMPOSTELA

In this context, results for both *SRPK2* and *NKX2-2* are quite relevant since: (i) rs910805 maps to *XRN2* (not previously implicated in ASD) but our analysis shows that associated variants at this locus are most probably altering *NKX2-2* expression, although a greater statistical power is needed; (ii) rs111931861 maps to *KMT2E* (already associated with ASD) but our analysis gives support to *SRPK2* (novel candidate association). Overall, this highlights the relevance of performing complementary analysis to link GWAS variants with genes, and not just focusing on the "closest gene" strategy[307].

Up to this point, we have identified two previously reported ASD genes (*NKX2-2* and *MAPT*, score 3 in SFARI rank) and two possible novel findings (*PTPRE* and *SRPK2*). However, it is important to note three main limitations when exploring ASD common genetic variation: (i) early GWAS of ASD have failed to detect strong signals, and, as a consequence, colocalization analysis are intrinsically limited in terms of statistical power; (ii) numerous studies have found an important role of AS in ASD onset, so it may be possible that the detected variants affect regulation at a different level than transcription regulation; (iii) due to intrinsic tissue heterogeneity, it may be needed to deepen our knowledge on eQTLs in specific cell types relevant to the disease, but also in different developmental stages, where eQTLs are not fully characterized.

5.4.2 Colocalization in the pan-tissue analysis

Of note, the other five genes showing significant correlations restricted to the pan tissue analysis (*CRHR1*, *KANSL1*, *MANBA*, *MMP12* and *WNT3*) underscore the same limitations as those discussed in the previous sections. The fact that correlations in the brain did not reach significance is likely attributed to the lack of specific eQTL datasets and the limited statistical power when considering the GWAS p-values (as larger cohorts are genotyped, significance values will increase).

5.4.2.1 *CRHR1*, *KANSL1* and *WNT3* in adjacent loci

However, a relevant finding regarding the adjacent LOIs harboring *CRHR1*, *KANSL1*, *WNT3*, should be considered. These three genes (along with *MAPT*, see previous section) are present in close proximity in the genome (Appendix 10.1, Supplementary Figures 2, 3 and 6) and have been selected for this analysis based on GBA results. Their biological roles all give suggestive evidence of their possible implication in ASD risk, especially the fact that *KANSL1* is overexpressed in ASD-risk relevant periods (Figure 16 A, B), and still, we are here able to point to *CRHR1* as the most probable causal gene in this locus.

The dysregulation of the stress response within the hypothalamic–pituitary–adrenal axis, specifically involving the corticotropin-releasing hormone and its primary receptor (*CRHR1*), is recognized as a significant factor contributing to the pathogenesis of MD and SZP[308,309], in agreement with the results from the colocalization analysis. Furthermore, a bidirectional model has been suggested to explain the role of *CRHR1* in anxiety regulation[310]. Refojo *et al.*[310] characterized an imbalance between *CRHR1*-regulated anxiogenic glutamatergic (thus, the absence of *CRHR1* in forebrain glutamatergic circuits is linked to decreased anxiety) and anxiolytic dopaminergic systems (the removal of *CRHR1* in midbrain dopaminergic neurons

USC UNIVERSIDADE DE SANTIAGO

leads to heightened anxiety-like behavior). These results are in line with significant correlations for both congruent and incongruent signals: variants that either elevate or diminish *CRHR1* expression are linked to the onset of the disease through distinct mechanisms (Tables 11, 12).

Moreover, dysregulation of WNT signaling has been documented in various psychiatric disorders, including ASD and SZP[311]. Remarkably, *WNT3* expression is heightened in the prefrontal cortex of ASD patients, implying that an overactivation of WNT signaling is involved in the pathogenesis of ASD. Conversely, studies in mice have demonstrated that loss of *WNT3*a prevents hippocampal development, and its role in synapse development and plasticity is well established. Considering all the evidence, both congruent and incongruent effects of *WNT3* may be linked to the pathogenesis of ASD and other NDDs, as shown here (Tables 11, 12).

Furthermore, the *KANSL1* gene has been widely characterized in the Koolen-de Vries syndrome (KdVS), characterized by congenital malformations, developmental delay and/or intellectual disability, and behavioral features. KdVS diagnosis is established in a proband who has either a heterozygous 500-650 kb deletion at chromosome 17q21.31 that includes *KANSL1* or a heterozygous intragenic pathogenic variant in *KANSL1*[312]. Accordingly, our findings indirectly suggest a potential association between a decrease in *KANSL*'s expression and ASD (the significant correlation in brain is limited to incongruent SNPs; Table 12).

Nonetheless, when correlation signals are scrutinized in detail, we can conclude the following: although *KANSL1* and *WNT3* yield significant correlation signals overall, these are sustained by a limited number of variants, with relative low p-values (clustered towards the bottom-left corner of the P-P plot) and without any variant representing a bona fide colocalized variant. On the contrary, *CRHR1*'s signal is sustained by a considerable number of variants, with great P-P distribution, and with two variants (congruent and incongruent) situated in the top-right corner of the P-P graph. Furthermore, the proximity of *MAPT* (discussed in the previous section) and *CRHR1* within adjacent loci suggests that this genomic region is likely to attain genome-wide significance in future GWAS with larger sample sizes, which would reinforce the association of these two genes with disease risk.

5.4.2.2 Additional implications

As additional results, we found significant effects of *MANBA* in ASD (Table 11, Appendix 10.1, Supplementary Figure 4), but these results could not be confirmed in any other pathology. Higher levels of *MANBA* expression in cerebellum have been correlated with a reduced risk of ADHD[313]. However, the analysis for ADHD did not reveal any significant correlation, and we were unable to confirm its implication in the brain. Although further research is necessary to establish a robust explanation for this discovery, in this cohort, we can conclude that higher expression levels are not regulated by common variation acting as eQTL, or at least, not in brain. Alternatively, potential scenarios encompass (i) pleiotropic effects of *MANBA* in other tissues, and (ii) the prospect that eQTLs regulate *MANBA's* expression, although we are currently unable to confirm this due to insufficient signal strength, preventing validation of these associations in the current ASD cohort. Within the aforementioned study (ref 313), the association signal for the gene originated from a Han Chinese population,

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

introducing the possibility that variants for this gene may not be discernible with adequate statistical power due to population biases in their frequency.

Furthermore, in regard to *MMP12*, matrix metalloproteinases (MMPs) stand out as key modulators of the extracellular matrix (ECM). The ECM in the brain has gained recognition as a crucial repository of signaling molecules that can impact processes such as synaptic plasticity, synaptogenesis, and neurite outgrowth within the central nervous system. Of note, *MMP12* has been shown to be expressed in the brain[314], although its implication in NDDs has yet to be detected. However, many lines of evidence support the role of *MMP9*, another protein member of the same family, in ASD, as well as in attentional dysfunction[314] (in line with the colocalizing signal obtained in ADHD (Table 12)). Particularly, both mRNA and protein levels were demonstrated to be higher in patients with different NDDs, when compared to healthy controls. This last piece of evidence is in agreement with the congruous effects of *MMP12* here highlighted. However, no conclusion can be obtained for this gene, since: (i) the correlation signal is sustained by only a few variants, (ii) these do not show particular elevated p-values, even when considering the lower significances in current ASD cohorts, and (iii) no eQTL in any brain tissue achieved the threshold for significance (Appendix 10.1, Supplementary Figure 5).

5.4.3 Implications of the eQTL colocalization analysis

Altogether, given all the aforementioned limitations, our results show how eQTL analysis constitutes another way to identify the most probable gene (*i.e.*, with the strongest support for causality) than just focusing on the closest gene to GWAS significant SNPs, thus providing further evidence for functional validation.

Furthermore, the division of SNPs with congruous and incongruous effects provided a clearer basis for interpretation, aligning, in most instances, with prior knowledge of the gene's impact on disease pathology in the vast majority of the aforementioned scenarios.

Finally, our results further highlighted the need of creating eQTL datasets in specific tissues, cell types and neurodevelopmental stages related to the disease in order to identify statistically significant eQTLs. Moreover, ongoing GWAS projects with larger sample sizes, as the one presented here, will more than likely contribute to identify risk variants of modest effect not previously detected. Additionally, the fact that most of the ASD results are supported in datasets for genetically correlated diseases, suggests that understanding the regulatory mechanism at these loci may reveal the basis for pleiotropic effects across psychiatric disorders and that leveraging GWAS data from each disease might help to build a better comprehension of underlying common mechanisms for NDDs.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CHAPTER 2

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

6 CHAPTER 2

ASD have a strong heritability estimated around 80%[24], although their complex genetic etiology has limited progress towards understanding its molecular basis.

Genomic approaches, including GWAS[37,104], WES[59,69] and WGS of families[61,315], as well as transcriptome analyses by RNA-seq[316–318] and microarray techniques[71,94], have yielded association with hundreds of genes over the last decades. However, a large portion of its genetic architecture remains uncharacterized, and ASD diagnosis continues to be a major challenge in both clinical and research settings.

Emerging advances have implicated AS as yet another process influencing the development of the disease. AS is the mechanism by which introns are excised from the pre-mRNA primary transcript, and exons are selected and concatenated in different arrangements, generating multiple transcript isoforms and, consequently, protein products, from a single gene.

More than 95% of multiexon human genes that encode proteins undergo AS, and most mRNA splice patterns exhibit tissue- and cell-type-specificity[319]. Consequently, AS is mainly accountable for generating the vast proteomic diversity observed in complex organisms and has been increasingly linked to functional intricacies within the central nervous system. It plays crucial roles in nervous system development, neuronal differentiation and maturation, as well as complex neuronal processes, such as the control of synaptic plasticity associated with cognition[194,320].

Certain neuronal genes, such as neurexins, have the capacity to generate hundreds of mRNA isoforms, leading to different sets of proteins with potentially diverse functions essential for efficient neuronal operation[184]. Notably, aberrations in AS have been associated with several NDDs including SZP and BD[198,321], Rett Syndrome[322,323], FXS[324] and, importantly, in ASD (see section “3.3.1.6 Splicing variants”).

6.1 OBJECTIVE

Given the mounting evidence of splicing disruptions in ASD patients, we sought to investigate the possible effects of DNVs in AS within a Spanish cohort of 360 ASD trios (unaffected parents and affected proband) that has not been previously ascertained for this kind of variation.

Thus, we employed SpliceAI, a machine learning tool that robustly predicts splice sites and splice-disrupting variants. It exhibitis notable performance when compared to previously developed *in silico* detection tools[36], serving as the gold standard for assessing AS nowadays.

Moreover, we utilized SpliceVault[325] and ABSsplice[326], two newly developed tools that further assess AS, in order to bolster robustness of the *in silico* splicing prediction and delve into the specific effects within human tissues.

This approach provides initial insights into the potential functional roles of genes harboring splicing variants, thereby highlighting molecular pathways and biological processes in which they are implicated.

USC UNIVERSIDADE DE SANTIAGO

6.2 METHODS

6.2.1 Subjects

The analysis described herein builds upon the complete sample set examined in Alonso Gonzalez *et al.*[327], which explored the biological roles of postzygotic and germinal coding mutations but did not specifically assess splicing.

DNA extraction from the Spanish ASD samples, consisting of 360 trios, was performed using the GentraPuregene blood kit (Qiagen Inc., Valencia, CA, USA) from peripheral blood.

Participants from Santiago (n = 136) were recruited from the Complexo Hospitalario Universitario de Santiago de Compostela and Galician ASD organizations. Meanwhile, subjects from Madrid (n = 224) were enrolled through the AMITEA program at the Child and Adolescent Department of Psychiatry, Hospital General Universitario Gregorio Marañón.

Inclusion criteria stipulated that only individuals aged 3 years or older were included in the study.

Enrolled participants received a clinical diagnosis of ASD from trained pediatric neurologists or psychiatrists, following the criteria outlined in both the DSM Fourth Edition Text Revision (DSM-IV-TR) and Fifth Edition (DSM-5). Additionally, when deemed necessary, the Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R) were administered.

All participants, along with their parents or legal representatives, provided written informed consent, and the study was conducted in accordance with the principles outlined in the Declaration of Helsinki.

Ninety trios from the Spanish cohort were already analyzed by Lim *et al.*[328]. The entire Spanish cohort was included in Satterstrom *et al.*[40] as part of the Autism Sequencing Consortium (ASC), a large-scale international genomic consortium integrating ASD cohorts and sequencing data from over one hundred investigators. All data generated as part of the ASC were transferred to dbGaP with Study Accession: phs000298.v4.p3.

6.2.2 *In silico* prediction and prioritization of splicing variants

To perform subsequent analyses, we leveraged previously reported biallelic DNVs from our previous work[327]. Here, DNVs were defined as those present exclusively in probands and not in parents (*i.e.*, genotypes 1/0 or 1/1 in probands and 0/0 in parents). However, the filtering steps carried out in Alonso-Gonzalez *et al.*[327] were not used with SpliceAI. Instead, appropriate quality filtering for splicing variants was applied to the raw DNVs of the analysis, as detailed below.

To carry out *in silico* annotation and prioritization of splice variants, we used SpliceAI (https://github.com/Illumina/SpliceAI, Figure 17), which has been already utilized to assess the clinical impact of non-coding variants that act through altered splicing in patients with ASD[36].

SpliceAI was executed with default parameters (-D = 50; -M = 0). On one hand, the parameter -D refers to the maximum distance between the variant and the gained/lost splice site, defining a threshold based on the sensitivity of SpliceAI in detecting splice-altering variants. Notably, exonic and near intronic variants (≤ 50 bp from an exon-intron boundary) exhibit the highest sensitivity scores when assessing high-scoring predicted cryptic splice

variants (delta Δ score ≥ 0.8). Additionally, deep intronic variants (> 50 bp from the exon-intron boundary) demonstrate a depletion of high-impact cryptic splice mutations, suggesting increased difficulty in predicting their functional consequences. On the other hand, annotated raw files include splicing changes associated with the reinforcement of annotated splice sites and the weakening of unannotated splice sites. Generally, these changes are less pathogenic than the reverse scenarios, *i.e.*, weakening annotated splice sites and strengthening unannotated splice sites. Therefore, we sought to apply score masking (parameter -M) to set the Δ scores of these splicing changes to 0, as recommended for variant interpretation.

For the gene annotation file, we used the GENCODE V24 canonical annotation files included in the package. In cases where variants overlapped multiple genes, separate predictions were generated for each gene. The variant with the highest impact, indicating greater confidence in the prediction, was retained.

SpliceAI annotates variants solely within the boundaries defined by the gene annotation file. Moreover, it refrains from annotating variants in proximity to chromosome ends (within 5 kb on either side), deletions exceedingly twice the input parameter -D, or those inconsistent with the reference fasta file. As a result, a SpliceAI annotation was available for 252,520 *de novo* variants, which accounts for 72.91% of the initial 346,352 variants. Variants with Δ scores ≥ 0.8 (indicating confidently predicted splice-altering effects) were retained (n = 1,836).

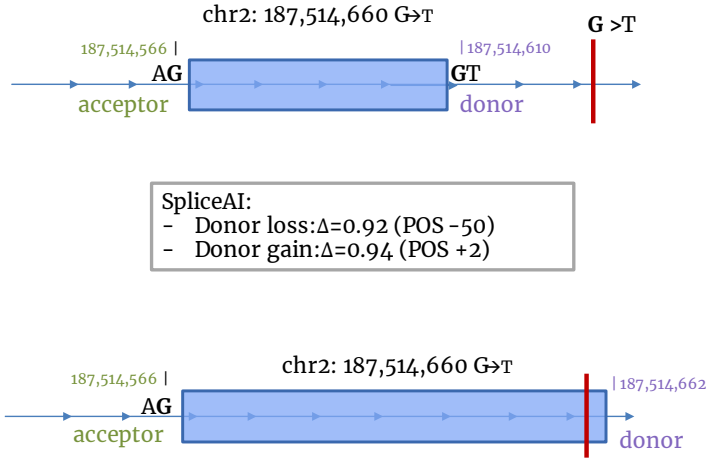


Figure 17. SpliceAI interpretation. The plot reflects an example of an intronic variant (red line) at position 187,514,660, which is situated 50 positions downstream of the canonical donor site (conserved GT nucleotides). Δ scores indicate that the G>T variant results in the canonical donor loss 50 positions upstream. With a high confidence Δ score of 0.94, SpliceAI predicts that this variant results in the creation of a cryptic donor 2 positions downstream of the variant, leading to partial intron retention and consequently, the elongation of the exon.

In addition, we considered only *de novo* variants with genotype quality (GQ) ≥ 20 and alternate read depth ≥ 7 and removed any call if it had an allele frequency > 0.1% across the samples in our dataset or in the non-psychiatric subset (which includes samples not ascertained for neurological or psychiatric phenotypes) of gnomAD (1,793 putative *de novo* variants excluded), leaving 43 high confidence splicing variants.

6.2.3 *In silico* variant validation

6.2.3.1 SpliceVault

The tool SpliceVault[325] was employed to accurately predict the exact manner in which genetic variants affect the splicing process.

SpliceVault is a tool designed to forecast how genetic variants influence splicing. It ranks the four most common unannotated splicing events across 335,663 reference RNA-Seq samples (300K-RNA Top-4) from GTEx v8 and the sequence read archive (SRA), the largest publicly available repository of high throughput sequencing data.

300K-RNA is built in hg38 (GRCh38), so genome coordinates were converted between assemblies, from GRCh37 to GRCh38, using the UCSC web server's LiftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver).

Splice variants were then interrogated against the SpliceVault web portal (https://kidsneuro.shinyapps.io/splicevault/) using default settings recommended for clinical use: top 4 events, skipping of ≤ 2 exons and cryptic positions within +/- 600 nucleotides (the highest sensitivity and positive predictive values for 300K-RNA predictions were obtained by applying these filters).

A variant was considered validated if it appeared in the top 4 events of SpliceVault, based on Dawes *et al.*[329] concluding remarks: “We find that the most frequently recurring natural mis-splicing events at each exon-intron junction, predict with accuracy which cryptic-donor will be activated in rare disease”. In their study, they discovered that when annotated splicing donors are disrupted by variants, the activation of a cryptic donor occurs in 87% of cases among the top four events.

Moreover, SpliceVault’s sensitivity outperforms SpliceAI in exon/double-exon skipping and cryptic splice site activation[325]. However, 300K-RNA Top-4 is not applicable for variants that impact canonical splice sites, prompting us to integrate both approaches.

6.2.3.2 Tissue specificity: ABSplice

Given the high tissue specificity of AS, each variant was examined for its specific effect in any given human tissue with ABSplice[326] (https://github.com/gagneurlab/absplice), a model which maps acceptor and donor splice sites and quantifies their usage in 49 human tissues from GTEx.

ABSplice was ran in a conda environment. We documented the tissue exhibiting the highest score for aberrant splicing outputted by ABSplice for each variant. This score indicates the likelihood that a specific genetic variant causes abnormal splicing in a particular tissue.

ABSplice thresholds are defined as 0.01 (low), 0.05 (intermediate), and 0.2 (high), which have approximately the same recalls as the high, medium, and low cutoffs of SpliceAI. Consequently, we extended the use of ABSplice as an additional tool for variant validation. A variant was considered validated if it appeared in the ABSplice dataset with a score exceeding 0.2 in at least one tissue.

Additionally, to enhance our understanding of the variant effects across various tissues, we developed a custom script. This script annotated our variants by cross-referencing them with precomputed ABSplice-DNA scores for all tissues, not just the highest scoring one. The

precomputed ABSplice-DNA scores for 49 human tissues and all possible SNVs genome-wide for hg38 were made available at Zenodo (https://zenodo.org/records/7871809).

6.2.4 Complementary datasets

Comparing functional profiles can reveal functional consensus and differences among different experiments and helps in identifying differential functional modules in different datasets.

For further validation, candidate gene selection, and interpretation of hereafter interrogated genes harboring splice variants, as well as the analysis of GO enrichments and cluster comparison of enriched terms (detailed in the next section), we utilized previously reported ASD-related sets of genes (Table 14).

After variant validation, genes harboring validated DNVs were contrasted against different complementary gene datasets, including ASD-related genes implicated in AS, as well as the genes in the SFARI database. Overlapping genes were considered as candidate genes and were subsequently included in further GO analyses.

Moreover, we used the same gene datasets, which included ASD-related genes implicated in AS, as well as a study integrating *de novo* and inherited variants in 42,607 ASD cases[73] for further cluster comparison of enriched terms between different categories of genetic variants (*i.e.*, *de novo* and inherited coding variants not predicted to alter splicing, and splicing variants).

Table 14. Complementary datasets included in gene validation and GO analysis.

Description*	Study type	Samples
The Human Gene module of the SFARI database (up-to-date reference for all known human genes associated with ASD) was accessed (SFARI 07-17-2023 release) and queried against 1-2 scoring genes (which indicates high confidence and strong evidence of association with ASD, respectively)[317].	Mutation screening, family-based association, case-control, WES, WGS and CNV array	-
A highly conserved program of neural microexons primarily regulated by the neuronal-specific splicing factor nSR100/SRRM4[185].	RNA-Seq custom pipeline	22 autistic individuals, 20 controls
Unique patterns of alternative splicing and gene co-expression in ASD-affected dizygotic twins compared to their parents[195].	AS and co-expression analyses	Two pairs of DZ twins and their parents
Distinct alternative splicing patterns in the blood of patients with ASD compared to typically developing individuals[196].	Whole genome exon arrays	30 ASD patient, 20 controls
Dysregulated splicing pattern of *RBFOX1*-dependent alternative exons in the ASD brain[204].	Post-mortem brain tissue	19 ASD samples, 17 controls
Detection of a 1.30-fold enrichment of *de novo* splicing mutations in ASD ($p = 0.0203$) compared to healthy controls when employing SpliceAI[36].	High-depth mRNA sequencing	36 ASD samples
Analysis of *de novo* and inherited variants identifies 60 genes with exome-wide significance implicated in ASD, including five new risk genes (*NAV3, ITSN1, MARK2, SCAF1* and *HNRNPUL2*)[73].	WES/WGS	42,607 ASD cases

*The first 6 studies were used for gene validation; and the last 6 were used for cluster comparison of enriched terms.

6.2.5 Gene ontology analysis

6.2.5.1 Gene network analysis

HumanBase (refer to section "5.2.5.2 Gene network analysis" for further detail) was used to build a gene network for genes already associated, and thus, validated, in the aforementioned complementary datasets (Tissue Specific Gene Networks: GIANT).

CACNA1I, *CBLB*, *CLTB*, *DLGAP1*, *DVL3*, *KIAA0513*, *OFD1*, *PKD1*, *SLC13A3* and *SCN2A* were selected as the input genes along with brain tissue in the 5 existing data types (co-expression, TF binding, interaction, GSEA microRNA targets, GSEA perturbations).

The resultant network contains the subset of functionally related genes specific to brain tissue, capturing tissue-specific gene function, all of which were used to test for functional enrichment using genes annotated to GO terms. The Gene Ontology describes our understanding of the biological domain by considering three key aspects: (i) biological process (BP; larger processes, or "biological programs" achieved through multiple molecular activities), (ii) cellular component (CC; denotes the location relative to cellular compartments and structures) and, (iii) molecular function (MF; molecular-level activities performed by gene products) terms.

6.2.5.2 Enrichment analysis

ClusterProfiler (https://github.com/YuLab-SMU/clusterProfiler)[330], an R package tailored for contrasting biological themes among gene groups, was utilized to perform both GO over-representation tests and to deduce enriched functional profiles on separate gene clusters.

6.2.5.2.1 Gene ontology (GO) enrichment analysis

The package org.Hs.eg.db, provided by Bioconductor, was used as the genome wide annotation for humans.

We employed the Biological Id TRanslator tool, already implemented in the clusterProfiler package (with parameters: fromType = "SYMBOL", toType = "ENTREZID", OrgDb = "org.Hs.eg.db"), to obtain Entrez unique gene identifiers for the genes of interest. For genes failing conversion to an Entrez ID, we further employed the mygene module in Python (https://pypi.org/project/mygene/), which obtains the gene annotation data from several public data resources (NCBI Entrez, Ensembl, Uniprot, NetAffx, PharmGKB, UCSC and CPDB) and keep them up to date.

GO enrichment analysis was performed with specific significance thresholds (p-valueCutoff = 0.01, q-valueCutoff = 0.05) adjusted by the Benjamini-Hochberg procedure. Highly similar GO terms (*e.g.*, > 0.25) were removed by applying the "simplify" function to retain the most representative terms (*i.e.,* the most significant) with parameters: cutoff = 0.25, by = "p.adjust" and select_fun = min.

For results visualization, we adopted the dotplot graph, which depicts the enrichment scores (*i.e.,* adjusted p-values) and gene count as dot color and size, respectively. The gene ratio is plotted on the horizontal axis, and each GO category is represented on the y axis (the number of terms (most significant) were set to 10 with the showCategory parameter).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

6.2.5.2.2 Cluster comparer

In order to perform a biological theme comparison between the aforementioned sets of ASD-related genes, we used the "compareCluster" function, which calculates enriched functional profiles of each gene clusters and aggregates the results into a single object.

For visualization purposes, the "showCategory" parameter, indicating the display of the topmost significant categories, was set to 5.

This tool was utilized to compare clusters (*i.e.*, sets) of genes, namely (i) a custom gene list of hereafter validated genes and functional interactors in brain tissue *versus* previously ASD-associated genes implicated in AS and (ii) the custom gene list *versus* genes harboring *de novo*/inherited variants with no predicted roles in AS.

6.2.6 Gene expression analysis

The GTEx Multi Gene Query tool of the GTEx Project version 8[115] (https://www.gtexportal.org/home/) was employed to carry out the gene expression heatmaps.

Those genes harboring tissue-specific splice variants predicted by ABSplice were used as an input.

Expression values are represented as TPM (Transcripts Per Million), calculated as the number of reads for a gene and normalized by gene length. Additionally, different transcripts for each gene are collapsed during the normalization process.

Heatmaps display the average expression per tissue. Darker blue means higher relative expression of that gene in each label (tissue type), compared to a yellow/light-green color in the same label. Genes and tissues are ordered by cluster.

6.3 RESULTS

6.3.1 *In silico* variant prediction and validation

Using previously reported DNVs in a Spanish cohort of 360 ASD trios[327], we implemented several variant-level quality control filters and validation procedures to assess the potential implications of AS.

Out of the initial 252,250 variants scored with a SpliceAI annotation, we identified 43 high-confidence splicing variants, meaning they yielded a SpliceAI $\Delta \geq 0.8$ in at least one of the interrogated splice sites, thus resulting in four predictions: acceptor gain (AG), acceptor loss (AL), donor gain (DG) and donor loss (DL) (Figure 17).

These variants were detected in only 11.39% of the families (n = 41 probands), with two probands harboring 2 splicing variants in different genes (Appendix 10.2, Supplementary Table 2).

It is worth noting that achieving ideal *in vitro* validation would necessitate access to the tissue of relevance (presumably developing brain), which was not feasible. Consequently, we have undertaken validation through the utilization of diverse methods and supplementary RNA-Seq datasets from brain and other tissues (Figure 18).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

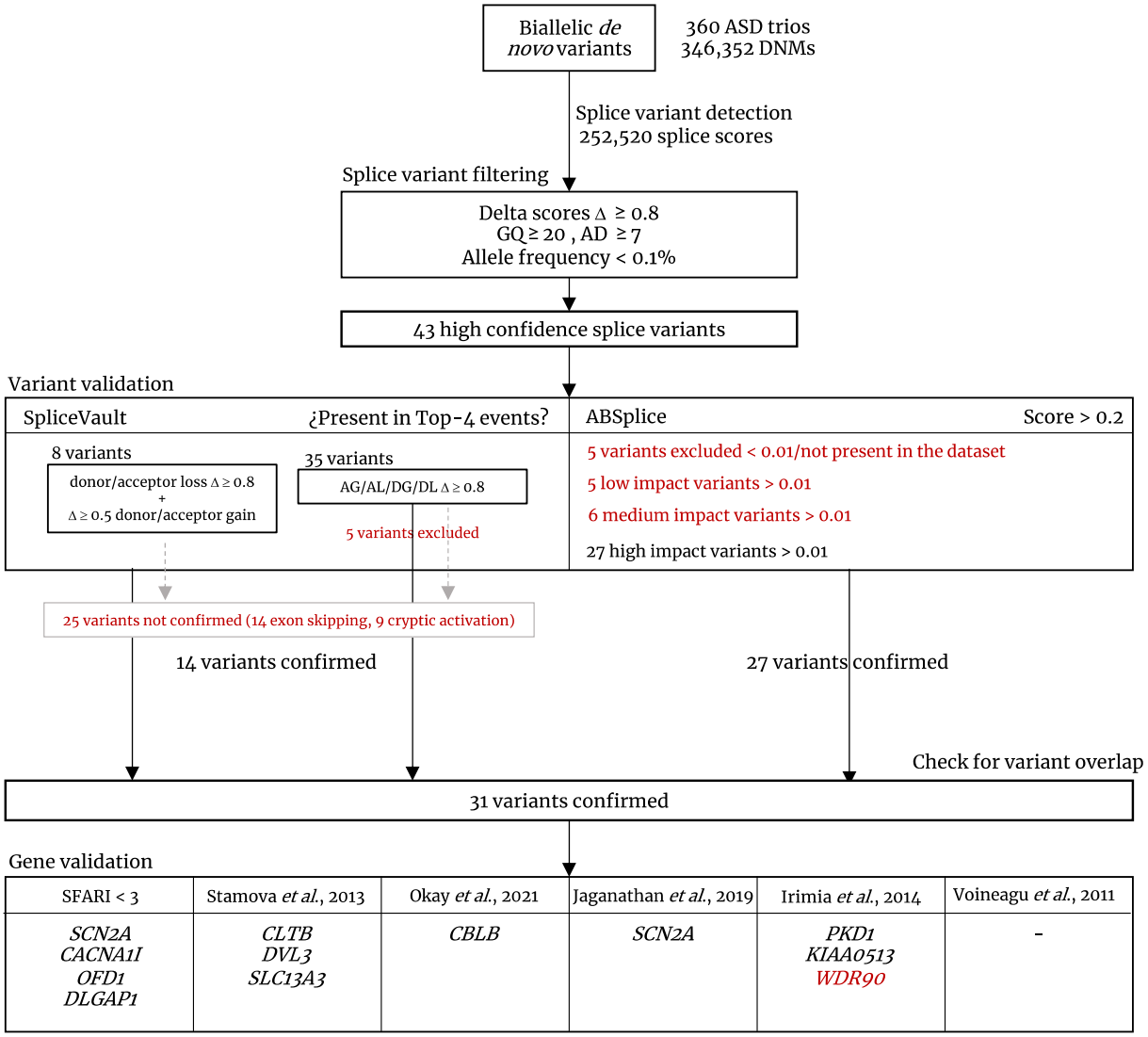


Figure 18. Workflow for *in silico* variant prediction, variant validation, and gene validation. Red font color is used to indicate variants that have not been confirmed or genes harboring unconfirmed variants.

Several procedures were followed in order to ensure robustness in the *in silico* prediction of the splicing effects of these variants.

Although the recommended threshold for splice-altering variants is $\Delta \geq 0.5$, we adopted a much more conservative threshold of $\Delta \geq 0.8$, which yields higher precision[36]. In fact, this cutoff showed the highest validation rate and outperformed other popular classifiers that have been referenced in the literature for rare genetic disease diagnosis (GeneSplicer, MaxEntScan and NNSplice)[36].

In addition, we reassessed all variants using SpliceVault, which quantifies natural variation in splicing and potentially predicts variant-related splicing changes (*i.e.,* exon-skipping events and cryptic splice sites).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Our dataset included 8 variants exhibiting cryptic donor/acceptor sites scores of ≥ 0.8 for site loss and ≥ 0.5 for site gain (Table 15). Among these, 87.5% of the cryptic activation variants were validated (n = 7 present in the Top-4 events ranked by SpliceVault).

The remaining variant (chr1-16895732-C-T) resulted in exon 23 skipping in 51.9% of the unannotated splice sites. In order to detect single exon skipping events, we would need to observe: (i) one single variant with AL and DL ≥ 0.8, or (ii) one individual harboring two different variants flanking the same exon with AL and DL ≥ 0.8, respectively. Methodological limitations prevented us from validating this phenomenon. The use of exome sequencing data introduces the potential limitation that deep intronic variants, not detectable through this method, might be associated with exon exclusion.

Furthermore, 35 variants yielded ≥ 0.8 in only one out of four scored positions by SpliceAI (AG/AL/DG/DL) (Table 16).

Excluding 5 variants with (i) no annotated splicing, (ii) an associated gene not present in the SpliceVault server (*FAM27B*), (iii) no cryptic annotation (*i.e.,* no unannotated splicing events); 30 variants were queried against the SpliceVault server.

One variant with AG ≥ 0.8 and five with DL ≥ 0.8 were confirmed by SpliceVault to be correctly predicted and result in cryptic activation.

For the rest of the variants (n = 24), 54.17% of the Top-1 event resulted in exon skipping (12 single exon skipping and 1 double-exon skipping), while 9 variants resulted in cryptic activation, not detected in our method.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Table 15. Variants with an AL/DL Δ ≥ 0.8 and AG/DG Δ ≥ 0.5.

Variant (GRCh37)	Gene	AG	AL	DG	DL	SpliceVault check?	Total Samples
chr9-114176268-T-C	*KIAA0368*	0.66 (20)	0.98 (-2)	0 (21)	0 (-2)	Y	404 (0.2%)
chr12-3649770-A-C	*PRMT8*	0.76 (10)	0.98 (2)	0 (8)	0 (1)	Y	4 (0.0%)
chr16-85105388-G-T	*KIAA0513*	0.92 (5)	1.00 (1)	0 (2)	0 (-32)	Y	195 (0.1%)
chr1-155981618-G-A	*SSR2*	0 (33)	0 (-47)	0.98 (2)	0.92 (-17)	Y	24 (0.0%)
chr7-1538341-A-C	*INTS1*	0 (-26)	0 (39)	0.90 (34)	1.00 (2)	Y	19 (0.0%)
chr9-139407471-A-C	*NOTCH1*	0 (28)	0 (34)	0.90 (34)	1.00 (2)	Y	375 (0.2%)
chrX-47003870-A-C	*NDUFB11*	0 (45)	0 (32)	0.71 (12)	1.00 (2)	Y	28959 (10.9%)
chr1-16895732-C-T	*NBPF1*	0.74 (-2)	0.94 (-1)	0 (-2)	0 (-1)	N (exon 23 skipping)	77931 (51.9%)

Variants are in GRCh37. Columns AG, AL, DG and DL show SpliceAI Δ scores. Delta position (*i.e.*, the location where splicing changes in relation to the variant's position) is shown between parenthesis (negative numbers refer to positions upstream of the variant while positive numbers refer to downstream positions).

Table 16. Variants with AG/AL/DG/DL Δ ≥ 0.8.

Variant (GRCh37)	Gene	AG	AL	DG	DL	SpliceVault check?	Total Samples	Top 1 non-annotated event
chr16-5141894-G-C	*EEF2KMT*	0.94 (-1)	0 (45)	0 (-1)	0 (-32)	Y	15696 (9.4%)	NA
chr1-20650027-T-C	*VWA5B1*	0 (-35)	0 (-41)	0 (-8)	0.94 (-2)	Y	1579 (23.9%)	NA
chr1-67242087-G-A	*TCTEX1D1*	0 (3)	0 (-50)	0 (3)	0.98 (-1)	Y	3 (0.0%)	NA
chr3-122629685-A-C	*SEMA5B*	0 (-32)	0 (49)	0.48 (24)	1 (2)	Y	414 (0.6%)	NA
chr5-843723-C-A	*ZDHHC11*	0 (37)	0 (-1)	0 (-8)	0.81 (0)	Y	16583 (21.8%)	NA
chr16-2163160-A-C	*PKD1*	0 (7)	0 (25)	0.03 (-2)	0.86 (2)	Y	117305 (65.6%)	NA
chr3-105421304-C-A	*CBLB*	0.17 (-21)	0.92 (-1)	0 (-21)	0 (-15)	NA	8263 (4.9%)	exon skipping (12)
chr5-176958524-T-G	*FAM193B*	0.17 (-22)	0.98 (-2)	0 (-22)	0 (-2)	NA	25211 (16.9%)	exon skipping (5)
chr5-179133258-G-A	*CANX*	0 (2)	0.94 (1)	0 (-50)	0 (29)	NA	19391 (7.8%)	exon skipping (3)
chr11-376072-A-C	*B4GALNT4*	0 (37)	0.96 (2)	0 (-36)	0 (1)	NA	63 (0.1%)	exon skipping (12)
chr15-42168847-T-G	*SPTBN5*	0.15 (-8)	0.85 (-2)	0 (47)	0 (-2)	NA	53 (0.1%)	double exon skipping (19-20)
chr16-56904007-G-A	*SLC12A3*	0.11 (2)	1.00 (1)	0 (-22)	0 (0)	NA	106(1.0%)	exon skipping (5)
chr4-110749291-T-G	*RRH*	0 (32)	0 (-2)	0.37 (12)	0.90 (-2)	NA	4 (0.0%)	exon skipping (4-5)
chr9-78711019-G-A	*PCSK5*	0 (17)	0 (-1)	0 (10)	0.96 (-1)	NA	16110 (13.5%)	exon skipping (8)
chr16-22269096-G-T	*EEF2K*	0 (31)	0 (-37)	0 (50)	0.81 (-5)	NA	50781 (27.9%)	exon skipping (9)
chr16-30910856-T-G	*CTF1*	0 (14)	0 (-16)	0 (-30)	0.81 (-2)	NA	77 (0.1%)	exon skipping (2)

chr19-7686019-A-C	*XAB2*	0 (0)	0 (35)	0.05 (13)	0.96 (2)	NA	-	exon skipping (9)
chr20-3641171-A-C	*GFRA4*	0 (-30)	0 (15)	0 (23)	0.98 (2)	NA	27 (1.0%)	exon skipping (3)
chrX-13767653-G-C	*OFD1*	0 (3)	0 (-45)	0.46 (3)	1 (-1)	NA	556 (0.3%)	exon skipping (9)
chr17-40835837-A-G	*CNTNAP1*	0.02 (16)	1 (2)	0 (21)	0 (-49)	NA	73 (0.1%)	cryptic activation -158
chr10-118620666-A-G	*ENO4*	0 (-1)	0 (35)	0.90 (-1)	0.40 (35)	NA	35 (0.1%)	cryptic activation -39
chr16-711712-C-T	*WDR90*	0 (-2)	0 (-38)	0.92 (-2)	0 (-29)	NA	175 (0.1%)	cryptic activation -77
chr16-29473043-G-A	*SULT1A4*	0 (2)	0 (-43)	0.81 (1)	0.02 (16)	NA	80211 (38.8%)	cryptic activation -25
chr2-95539855-T-G	*TEKT4*	0 (-38)	0 (-40)	0.04 (-40)	0.98 (-2)	NA	40 (0.6%)	cryptic activation +512
chr2-166170276-G-A	*SCN2A*	0 (6)	0 (-30)	0 (10)	0.85 (-5)	NA	3373 (8.4%)	cryptic activation +213
chr8-91033285-G-T	*DECR1*	0 (-2)	0 (-48)	0 (-2)	0.98 (-1)	NA	42 (0.0%)	cryptic activation -51
chr13-114005162-A-C	*GRTP1*	0 (1)	0 (2)	0 (-44)	0.88 (2)	NA	36753 (34.7%)	cryptic activation -52
chr16-20638576-A-T	*ACSM1*	0 (47)	0 (2)	0 (-25)	0.98 (2)	NA	15744 (55.7%)	cryptic activation -67
chr19-10572358-T-G	*PDE4A*	0 (-14)	0 (-2)	0.18 (-14)	0.92 (-2)	NA	32888 (22.3%)	cryptic activation +35
chr11-65784647-T-G	*CATSPER1*	0.44 (-9)	0.96 (-2)	0 (43)	0 (-6)	NA	1912 (3.7%)	cryptic activation +31
chr22-40060742-A-C	*CACNA1I*	0.14 (6)	1 (2)	0 (36)	0 (1)	*	*	*
chr11-118938598-C-G	*VPS11*	0 (-1)	0 (-18)	0.96 (-1)	0.01 (-13)	*	*	*
chr10-51130591-A-C	*PARG*	0 (-38)	0 (30)	0 (-43)	0.98 (2)	*	*	*
chr9-67793896-C-A	*FAM27B*	0 (-47)	0 (-42)	0.08 (-46)	0.96 (1)	**	**	**
chr18-3502489-A-G	*DLGAP1*	0 (-17)	0 (2)	0.04 (-17)	0.98 (2)	***	***	***

*No annotated splicing, **gene not present in the dataset, *** no cryptic annotation. Variants are in GRCh37. Columns AG, AL, DG and DL show SpliceAI Δ scores. Delta position (*i.e.,* the location where splicing changes in relation to the variant's position) is shown between parenthesis (negative numbers refer to positions upstream of the variant while positive numbers refer to downstream positions). For exon skipping events, the skipped exon is shown in parenthesis. For cryptic activation events in SpliceVault, the cryptic position is depicted.

Further on, we sought to revalidate predicted variants against ABSplice[326] with 62.79% of the variants yielding scores ≥ 0.2 (equivalent to the high precision cutoff ≥ 0.8 in SpliceAI) and were thus, confirmed (Table 17).

Table 17. ABSplice prediction.

Variant (GRCh37)	Gene	ABSplice score	ABSplice tissue
chr22-40060742-A-C	*CACNA1I*	**0.43**	Brain Cerebellum
chr12-3649770-A-C	*PRMT8*	**0.4**	Brain Nucleus accumbens basal ganglia
chr1-67242087-G-A	*TCTEX1D1*	**0.38**	Brain Frontal Cortex BA9
chr8-91033285-G-T	*DECR1*	**0.36**	Adipose Subcutaneous
chr9-139407471-A-C	*NOTCH1*	**0.36**	Adipose Subcutaneous
chrX-13767653-G-C	*OFD1*	**0.36**	Adipose Subcutaneous
chr9-78711019-G-A	*PCSK5*	**0.35**	Adipose Visceral Omentum
chr16-2163160-A-C	*PKD1*	**0.34**	Adipose Subcutaneous
chr16-85105388-G-T	*KIAA0513*	**0.34**	Adrenal Gland
chr2-95539855-T-G	*TEKT4*	**0.34**	Testis
chr7-1538341-A-C	*INTS1*	**0.34**	Adipose Subcutaneous
chr16-22269096-G-T	*EEF2K*	**0.33**	Adipose Subcutaneous
chr3-122629685-A-C	*SEMA5B*	**0.31**	Artery Coronary
chr5-179133258-G-A	*CANX*	**0.29**	Brain Amygdala
chr13-114005162-A-C	*GRTP1*	**0.28**	Adrenal Gland
chr16-56904007-G-A	*SLC12A3*	**0.28**	Kidney Cortex
chr2-166170276-G-A	*SCN2A*	**0.28**	Brain Cerebellar Hemisphere
chr16-30910856-T-G	*CTF1*	**0.27**	Adrenal Gland
chr19-7686019-A-C	*XAB2*	**0.26**	Adipose Subcutaneous
chr9-114176268-T-C	*KIAA0368*	**0.26**	Adipose Subcutaneous
chr11-376072-A-C	*B4GALNT4*	**0.24**	Brain Amygdala
chr17-40835837-A-G	*CNTNAP1*	**0.23**	Brain Anterior cingulate cortex BA24
chr3-105421304-C-A	*CBLB*	**0.23**	Adipose Subcutaneous
chr11-65784647-T-G	*CATSPER1*	**0.21**	Testis
chr18-3502489-A-G	*DLGAP1*	**0.21**	Brain Anterior cingulate cortex BA24
chr19-10572358-T-G	*PDE4A*	**0.21**	Testis
chr15-42168847-T-G	*SPTBN5*	**0.2**	Nerve Tibial
chr1-20650027-T-C	*VWA5B1*	0.18	Testis
chr5-176958524-T-G	*FAM193B*	0.17	Adipose Visceral Omentum
chr16-29473043-G-A	*SULT1A4*	0.09	Brain Cerebellum
chr16-5141894-G-C	*EEF2KMT*	0.07	Adipose Visceral Omentum
chr16-20638576-A-T	*ACSM1*	0.055	Testis
chr5-843723-C-A	*ZDHHC11*	0.045	Brain Cerebellar Hemisphere
chr1-155981618-G-A	*SSR2*	0.04	Adipose Subcutaneous
chr20-3641171-A-C	*GFRA4*	0.04	Brain Amygdala

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

chr10-118620666-A-G	*ENO4*	0.032	Testis
chr16-711712-C-T	*WDR90*	0.032	Adipose Subcutaneous
chr1-16895732-C-T	*NBPF1*	0.18	Brain Cerebellar Hemisphere
chr4-110749291-T-G	*RRH*	<0.01	NA
chr11-118938598-C-G	*VPS11*	<0.01	NA
chr9-67793896-C-A	*FAM27B*	*	NA
chr10-51130591-A-C	*PARG*	*	NA
chrX-47003870-A-C	*NDUFB11*	*	NA

Variants are sorted by ABSplice scores. Variants with scores ≥ 0.2 (bold font) were confirmed. *Variant not present.

After the validation process, 75.61% (n = 31) of the initially predicted splicing variants (excluding those variants not present in any of the complementary datasets (n = 2)) were confirmed (Appendix 10.2, Supplementary Table 2).

6.3.2 Tissue specificity of predicted splice variants

Following cross-referencing with SpliceVault and ABSplice, we further evaluated tissue-specific effects of the *in silico* validated variants (n = 31), albeit the score provided by ABSplice. This approach allowed us to globally assess tissue-specific effects of all validated variants, acknowledging that some may not be high confidence in ABSplice. Thus, variants that did not reach the 0.2 impact score threshold in ABSplice but were validated in SpliceVault were also included in this analysis (Appendix 10.2, Supplementary Table 2).

After removing one variant not present in the dataset, we evaluated tissue-specific effects in 26 variants with score ≥ 0.2 (high impact), 2 variants with score ≥ 0.05 (medium impact) and 2 with score ≥ 0.01 (low impact). Notably, adipose tissue yielded the highest scores for 38.7% of the variants (n = 12), followed by brain with 32.1% (n = 9), testis and adrenal gland with 9.6% each (n = 3), and the remaining 3 variants (each comprising 3.6% of the total) were distributed among nerve tibial, kidney cortex, and artery coronary.

Then, genes harboring the *in silico* validated splice altering variants were queried against the GTEx portal to assess whether the predicted tissue-specific effects were attributable to gene expression restricted to that particular tissue.

Overall, genes like *CACNA1l*, *SCN2A*, *DLGAP1* and *PRMT8*, which harbor variants predicted to have their highest impact in brain tissue, do show higher expression values restricted to brain tissues (Figure 19, gene name in green). Only one of these genes (namely, *CANX*), with a validated variant predicted to have the highest impact in amygdala, did not show a high expression limited to the brain.

However, genes that host splicing variants with the highest impact in adipose tissue (Figure 19, gene name in purple) do not exhibit expression limited to any specific tissue. This is in contrast to the expectation, as a specific expression limited to adipose tissue would provide a logical rationale for the increased burden of variants yielding the highest scores in adipose tissue.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

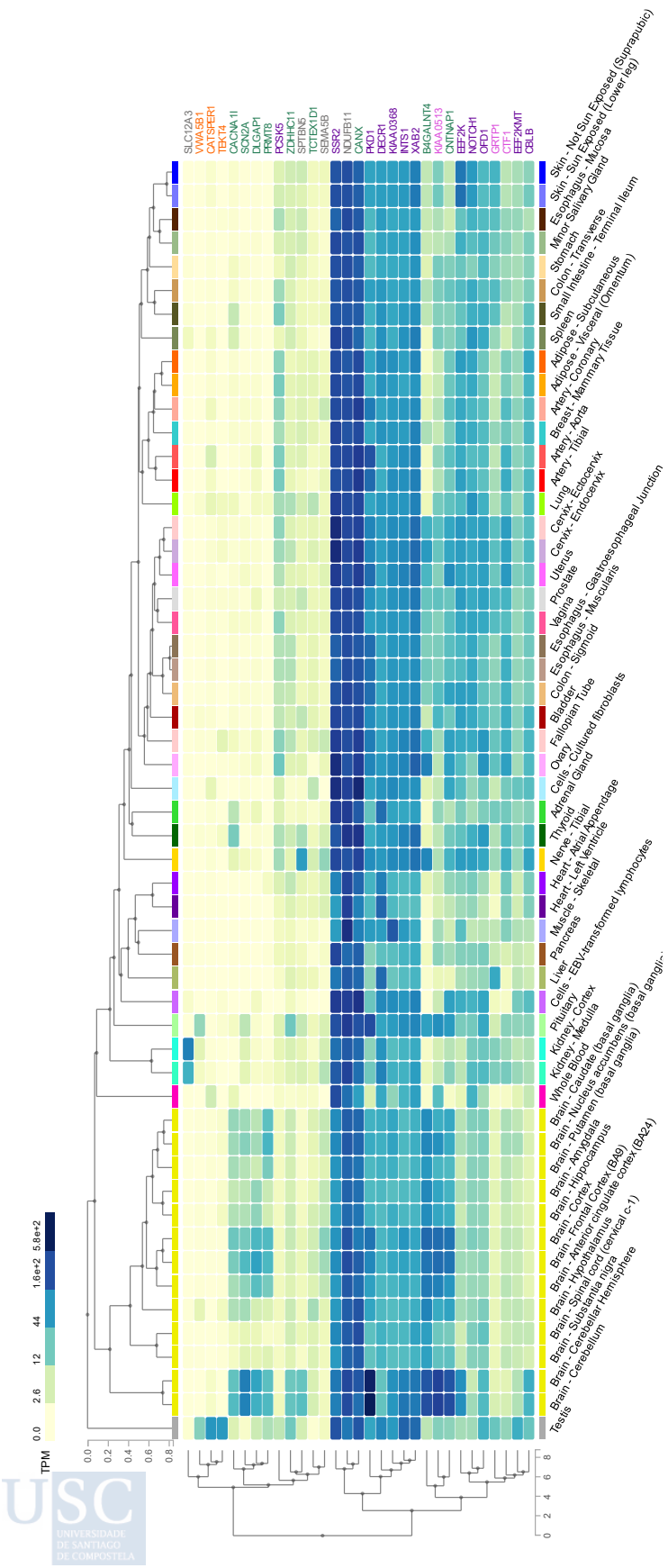


Figure 19. Tissue specific expression of genes harboring 31 *in silico* validated variants. Gene expression heatmap for genes harboring splice variants with predicted tissue-specific impact. Genes and tissues are ordered by cluster. Genes are highlighted based on the tissue where their splicing variant yields the highest score: green for the brain, purple for adipose tissue, orange for testis, and gray for other tissues. This plot was created with the tool GTEx Multi Gene Query (https://www.gtexportal.org/home/multiGeneQueryPage) and edited in Adobe Illustrator.

Nonetheless, variants were checked against the whole set of tissues, and it was observed that most of the variants with predicted highest scores in brain, and all the variants with predicted highest effect in adipose tissue, yielded the same high scores in other tissues (*i.e.*, ABSplice scores were not exclusive for that tissue) (data not shown).

In contrast, 3 variants with the highest score identified in testis, exhibited a clear tissue-specificity (*i.e.*, the ABSplice scores retrieved for the rest of tissues were notably lower) (Appendix 10.1, Supplementary figure 7). However, genes harboring these variants (*CATSPER1*, *TEKT4*, *VWA5B1*) had tissue specific expression (Figure 19, gene name in orange) restricted to testis.

6.3.3 Cluster enrichment

Genes harboring high-confidence splice variants predicted by SpliceAI and validated with SpliceVault and ABSplice (n = 31 variants, one variant per gene) were cross-referenced with:

i. previously reported genes associated with ASD aberrant splicing[36,185,195,196,204], to check for similarities in splicing relevant pathways and perform gene validation.
ii. ASD-associated genes in the SFARI Gene (category 1: high confidence, category 2: strong candidate), for gene validation only.
iii. Genes harboring *de novo* and inherited coding mutations[73], to see if different types of variants (*i.e.*, coding *versus* splicing variants) act through distinct mechanisms.

CACNA1I, *CBLB*, *CLTB*, *DLGAP1*, *DVL3*, *KIAA0513*, *OFD1*, *PKD1*, *SLC13A3* and *SCN2A* were present in at least one of the above-mentioned datasets of ASD risk genes, and were thus confirmed and used to construct a brain-specific network of functionally-related genes, hereafter referred to as the splicing gene list (Appendix 10.2, Supplementary Table 3; n = 60). The resultant gene network was interrogated for enrichment in the GO categories of BP, MF, and CC.

The analysis revealed that these genes were significantly enriched for BP related to proper neuronal functioning, such as modulation of chemical synaptic transmission (gene ratio 12/60, q-value = 1.68×10^{-5}), trans-synaptic signaling (gene ratio 12/60, q-value = 1.68×10^{-5}) and synaptic plasticity (gene ratio 8/60, q-value = 1.07×10^{-4}). Other enriched BP were cognition (gene ratio 8/60, q-value = 1.07×10^{-4}) and memory (gene ratio 9/60, q-value = 1.68×10^{-4}) (Figure 20).

Enriched CC terms were all related to the synapse, with significant findings including postsynaptic specialization (gene ratio 12/60, q-value = 1.12×10^{-7}), synaptic membrane (gene ratio 10/60, q-value = 7.84×10^{-6}), glutamatergic synapse (gene ratio 10/60, q-value = 1.03×10^{-5}), and presynaptic synapse (gene ratio 12/60, q-value = 3.34×10^{-4}), among others (data not shown).

Top enriched MF terms were associated with calmodulin binding channels (gene ratio 7/60, q-value = 3.25×10^{-4}), transmembrane receptor protein kinase activities (gene ratio 5/60, q-value = 3.25×10^{-4}), calcium ion channels (gene ratio 6/60, q-value = 3.25×10^{-4}), and tyrosine activities (gene ratio 4/60, q-value = 1.67×10^{-4}), among others (data not shown).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

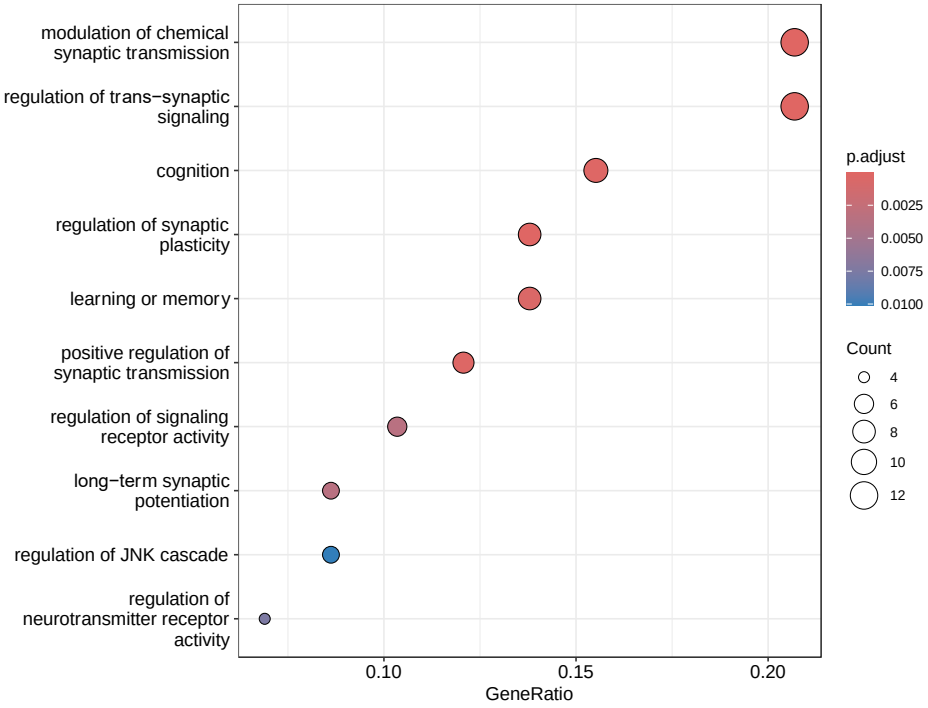


Figure 20. Top 10 enriched biological processes in the splicing gene-list.

Furthermore, we compared functional profiles amongst the different splicing datasets[36,185,195,196,204] and calculated enriched functional profiles for each one. In analyzing datasets for genes implicated in AS in ASD, we found that our gene network clusters together with the other sets of "splicing genes" in terms of BP, CC and MF. Top significant enriched terms included: (i) protein autophosphorylation and modulation of chemical synaptic transmission, for BP; (ii) postsynaptic specialization and cell leading edge, for CC; and (iii) actin/calmodulin binding, for MF (data not shown).

However, when contrasting these findings with genes harboring coding variants (henceforth designated as the non-splicing gene list), the overlap between enriched categories is notably dissipated, especially for MF and BP (Figure 21). CC terms did not exhibit a clear separation between datasets, with all common enriched terms relating to synapse components or postsynaptic density (data not shown).

While some BP (*e.g.,* cognition, learning and memory) were significantly enriched in both datasets, others (*e.g.,* histone modification (gene ratio 17/72, q-value = 3.77×10^{-9}) and chromatin remodeling (gene ratio 15/72, q-value = 2.45×10^{-8}) were specifically enriched in the non-splicing gene list, absent in our 60-gene list (Table 18).

On the other hand, genes from both datasets were incorporated into categories associated with the proper function and organization of the synapse. However, the majority of the significant enrichments in the splicing gene list (48 out of 62 enriched BP terms) were exclusively identified within that particular dataset. Some examples of these enrichments are provided in Table 18.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Table 18. Common/unique significantly enriched BP in splicing-related gene lists and the non-splicing gene list.

Gene list	Description	Gene ratio	q-value
Splicing gene list	regulation of signaling receptor activity	6/60	2.79×10^{-3}
	regulation of neurotransmitter receptor activity	4/60	6.00×10^{-3}
	protein autophosphorylation	6/60	8.17×10^{-3}
	regulation of JNK cascade	5/60	8.17×10^{-3}
	dendrite development	6/60	8.17×10^{-3}
	regulation of protein catabolic process	7/60	1.02×10^{-2}
	positive regulation of MAPK cascade	8/60	1.08×10^{-2}
	JNK cascade	5/60	1.34×10^{-2}
	regulation of phosphatidylinositol 3-kinase signaling	4/60	2.15×10^{-2}
	regulation of synapse assembly	4/60	2.15×10^{-2}
	calcium ion transport	7/60	2.25×10^{-2}
	excitatory postsynaptic potential	4/60	2.46×10^{-2}
Non-splicing gene list	histone modification	17/72	3.77×10^{-9}
	chromatin remodeling	15/72	2.45×10^{-8}
	histone lysine methylation	7/72	4.11×10^{-5}
	histone methylation	7/72	1.20×10^{-4}
	histone H3-K4 methylation	5/72	3.40×10^{-4}
	regulation of histone methylation	3/72	2.07×10^{-2}
	regulation of histone modification	4/72	2.57×10^{-2}
	positive regulation of histone H3-K4 methylation	2/72	3.05×10^{-2}
	histone lysine demethylation	2/72	3.50×10^{-2}
	histone demethylation	2/72	3.67×10^{-2}
Common	learning or memory (splicing gene list)	8/60	4.34×10^{-4}
	learning or memory (non-splicing gene list)	2/72	2.52×10^{-7}
	cognition (splicing gene list)	9/60	1.68×10^{-4}
	cognition (non-splicing gene list)	13/72	1.24×10^{-7}

The most pronounced difference emerged when analyzing enriched terms in the MF category: genes with predicted splice variants were significantly enriched in terms such as calmodulin binding (gene ratio 7/60, q-value = 3.25×10^{-4}), calcium ion channel/transporter activity (gene ratio 5/60, q-value = 3.25×10^{-4}), and transmembrane receptor protein kinase/tyrosine activity (gene ratio 6/60, q-value = 3.25×10^{-4}), while genes in the non-splicing list were specific to histone lysine N-methyltransferase activity (gene ratio 6/72, q-value = 3.85×10^{-6}) and beta-catenin binding (gene ratio 7/72, q-value = 3.85×10^{-6}) (Figure 21).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

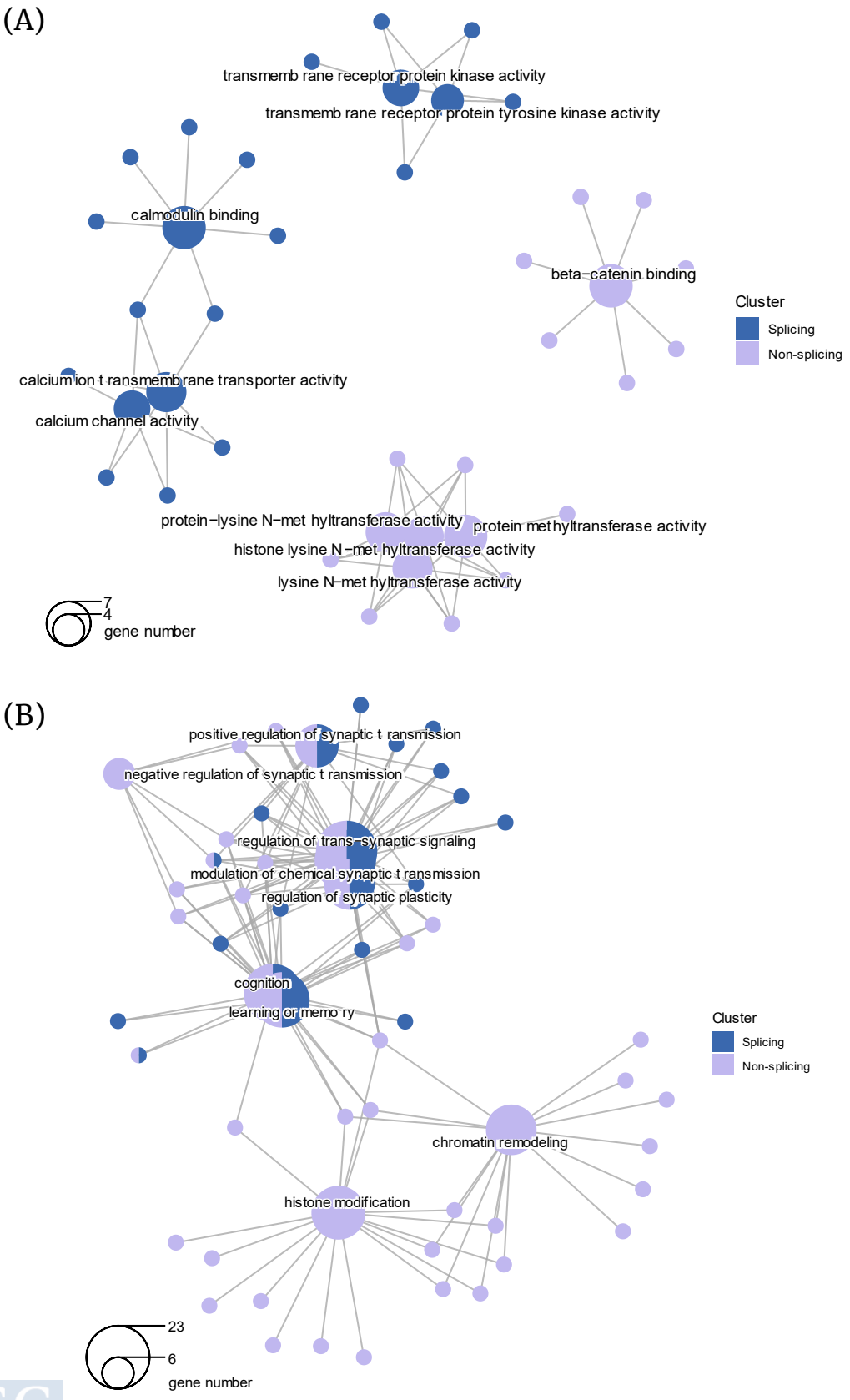


Figure 21. Cluster enrichment analysis for genes harboring splicing *vs.* coding variants (splicing gene list *vs.* non-splicing). Graphs depicting the number of genes of each list included in (A) molecular function, (B)

biological process. The circle size is proportional to the number of genes included in each category, and not significance, as depicted in the legend (both panels have different size ratios).

6.4 DISCUSSION

The multifaceted genetic etiology of ASD, characterized by substantial phenotypic and genetic heterogeneity, has long been a challenge in unraveling its underlying molecular basis. The identification of splicing variants has not been included in the major WGS or WES genetic studies involving large ASD cohorts. However, AS, an intricate mechanism that diversifies protein isoforms from a single gene, has recently garnered attention as a potential contributor to ASD pathogenesis.

The present study delved into the intricate landscape of AS in ASD through *in silico* prediction and validation of splicing variants. However, the conservative threshold (SpliceAI $\Delta \geq 0.8$) chosen for splice-altering variant prediction[36] may, in turn, result in elevated numbers of false negatives. While this approach was necessary, *in vitro* confirmation of the predicted variants (*e.g.,* by Sanger sequencing), validation of the predicted alterations on AS (by reverse transcription polymerase chain reaction (RT-PCR)) and functional analysis of their molecular impacts (RNA-Seq and/or minigene reporter systems[331]), would prove much more adequate and sensitive.

Still, *in vitro* validation was unfeasible due to the lack of sample availability and the difficulty in contacting participants for resampling. Thus, our validation strategy using SpliceVault and ABSplice, at least partially, provided further evidence on the robustness of the predicted splicing effects. On the one hand, SpliceVault exhibited superior sensitivity and positive predictive value than SpliceAI when it comes to exon- and double-exon skipping predictions or cryptic splice site activation, and represents the first evidence-based method for predicting the nature of variant-associated mis-splicing[325]. On the other hand, another study demonstrated that applying SpliceAI on the tissue-specific splice sites defined by SpliceMap (integrated into ABSplice) increased the precision of SpliceAI to 22% at 20% recall, with a significantly higher auPRC consistently across tissues[326]. Notably, our validation approach revealed that the majority of cryptic activation events were successfully corroborated when leveraging evidence-based data (Tables 15, 16). This further emphasizes the role of these predicted cryptic sites in ASD-associated splicing perturbations.

Nevertheless, splicing branchpoints present an additional source of potentially damaging non-coding variants which are amenable to systematic analysis in WGS data[332], but remain undetectable in WES data. This represents another methodological constraint in our approach. Of note, the availability of WGS data could enhance our understanding of the splicing landscape in ASD by enabling the detection of intron retention, a splicing aberration already associated with ASD and other NDDs[333,334] .

Furthermore, 35% of cryptic splice variants with weak and intermediate predicted scores (Δ 0.35 – 0.8) exhibit significant differences in the fraction of normal and aberrant transcripts produced across tissues. Variants with high predicted scores are significantly less likely to produce tissue-specific effects[36]. Therefore, being able to choose a less conservative threshold and perform *in vitro* validation, would be tremendously helpful in gaining insight into tissue-

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

specific effects. However, the tissue-specific effects of splicing variants gained prominence through our assessment using ABSplice.

6.4.1 Tissue-specificity in alternative splicing (AS)

Research in ASD has primarily focused on neurological aspects, looking at factors such as brain structure and function, and neurotransmitter systems. However, it is worth noting that there is ongoing research in the field of neuroimmunology and the gut-brain axis, which explores the connections between the gut and the brain. Recently, GI dysfunction has been described in various neurodevelopmental and psychiatric disorders including ASD[116,335]. Moreover, some studies have suggested a possible link between GI tissue, adipose tissue and brain, with accumulating evidence suggesting that the communication pathways linking them might be promising intervention points for metabolic disorders[336]. However, since adipose tissue can produce certain signaling molecules, such as adipokines, it's possible that there could also be indirect connections between adipose tissue and neurodevelopmental conditions like ASD[337], but this area of research is still emerging and not yet well-understood. These adipokines, especially leptin, have been shown to play roles in the central nervous system, including neurogenesis, neuroprotection, synaptogenesis, and synaptic plasticity[338,339]. In fact, the circulating leptin levels have already been linked to neurodegenerative diseases such as AD[340,341]. Based on emerging evidence, the adipose tissue's ability to produce signaling molecules, suggests potential indirect connections between adipose tissue and neurodevelopmental conditions like ASD.

Importantly, preliminary findings of this study indicated a role of adipose tissue in ASD (Table 17). Yet, upon examining the expression of genes containing variants with adipose tissue-specific effects we noted a relatively uniform expression across all tissues (Figure 19). Notably, these genes exhibited elevated expression levels in the brain, and the associated splicing variants also yielded high scores in brain tissues. As a result, it remains unclear whether a function in adipose tissue has any relevance here or if it is merely speculative. Further studies could test this hypothesis by interrogating whether these genes are driving pleiotropic effects in both sets of tissues, or by performing overall comparisons between splice site usage in neuronal *versus* adipose tissue.

Moreover, three variants in the final set of splicing validated variants, show unique values of tissue-specificity in testis (Appendix 10.1, Supplementary Figure 7). Nonetheless, when comparing transcriptomes, it has been observed that the brain and testis significantly surpass other tissues in terms of the diversity of expressed splice variants[342]. Consequently, our findings may lack sufficient power to attribute specific significance to testis in the context of ASD risk.

6.4.2 Biological underpinnings of AS

On another note, the convergence of genes harboring validated splicing variants with previously reported ASD-associated genes from various datasets substantiates the potential significance of AS in ASD. Our creation of a brain-specific network encompassing functionally related genes demonstrated enrichment in BP intricately tied to neuronal functioning, synaptic transmission, synaptic plasticity, cognition, and memory (Figure 20). Although we identified a

USC DE COMPOSTELA

relatively small number of genes with *in silico* validated splice variants, these findings align with previous studies showcasing aberrant splicing patterns in genes critical to neural development, which may collectively contribute to the complex ASD phenotype.

Additional support for this evidence includes: (i) analyzing larger ASD cohorts under the same criteria, both to augment our splicing gene-list (and thus provide more statistical support for enrichments in GO categories) and to perform a burden test analysis of numbers of mutations (are the numbers of splicing variants per gene consistent with gene length, conservation, and the number of different isoforms?), (ii) testing whether the splicing genes carry an excess of non-splice DNVs in autism probands to further correct this measure, and (iii) using multiplex family/case-control cohorts to check if splice DNVs are enriched in affected individuals when compared with healthy siblings.

In addition, we cross-referenced our gene list with genes that have exome-wide significance when combining evidence from both coding DNVs and rare inherited variants (non-splicing gene list), to encompass a broader spectrum of the disorder. Interestingly, our gene list shows significant enrichment in MF different from those enriched in the non-splicing gene list (Figure 21 A). Moreover, none of these non-splicing enriched terms were found in any other splicing complementary dataset analyzed in this study. Similar results were observed for the category of BP (Figure 21 B), where cognition and learning or memory terms are common amongst both lists (splicing *versus* non-splicing), but chromatin remodeling and histone modification are specific to the non-splicing gene list, and synaptic-related terms are specific to the splicing gene list (Table 18).

The fact that GO enrichment analysis points out to different MF and BP in genes harboring splicing variants and in the non-splicing gene list, might suggest a divergence of affected pathways and mechanisms, thus pointing to different mechanisms in which they participate in the development of the disease.

In fact, a previous study on the full-length isoform transcriptome of the developing human brain[320] has shown that differentially expressed isoforms (DEIs) reveal distinct signals relative to differentially expressed genes (DEGs). The GO enrichment analyses demonstrated stronger enrichment of DEI in neurodevelopment-relevant processes compared with DEGs. In contrast, DEGs were enriched in basic biological function-related processes, such as mitotic cell cycle, metabolic processes, protein targeting, and localization. Also, molecular functions such as kinase activity (one out of three most significant molecular functions strictly associated in our gene list) were solely linked to DEIs.

Furthermore, another investigation discovered that when human cortical neurons were exposed to the anticonvulsant valproic acid (VPA), a recognized environmental risk factor for ASD, differentially expressed genes (DEGs) exhibited enrichment in distinct molecular functions compared to genes with differential transcript isoform usage (DTU). Intriguingly, their results presented a contrast to ours: genes specific to DEGs were enriched for ion channel activity and transmembrane transporter activity (*i.e.*, molecular functions present in our splicing gene list). In contrast, genes specific to DTUs showed enrichment in histone binding, RNA II polymerase binding, and binding to transcription machinery. Nevertheless, they observed a

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

significant overlap between both DEGs and genes exhibiting DTU with known ASD risk genes[343].

These findings align with those presented in this study, suggesting that the full-length isoform transcriptome provides better biological insights into brain development than the gene transcriptome. However, proteins essential for synapse organization, such as neurexins and cadherins, as previously discussed, have the capacity to generate hundreds of mRNA isoforms. The fact that cluster enrichment analysis suggests a role for splice variants in synaptic organization and transmission, in contrast to non-splice genes, may reflect that synaptic genes have more alternative spliced forms than chromatin remodeling genes, rather than process specialization. Moreover, one of the limitations already discussed is that, based on methodological constraints, our gene list is relatively small, so the clustering into specific categories may be arbitrary.

In general, this situation mirrors the still unresolved question of whether *de novo* and inherited variations affect the same biological pathways. Ruzzo *et al.* illustrated that inherited variations cluster in specific biological pathways, introducing novel pathways linked to ion transport, the cell cycle, and the microtubule cytoskeleton[75] distinct from those enriched for *de novo* variants[69]. In a separate study, Wilfert and colleagues stated that DNMs and transmitted LGD variants converge on the same pathway but may be targeting distinct sets of genes[76].

6.4.3 Future perspectives

In essence, this study exemplifies the intricate genetic landscape of the disorder and aims to raise new questions regarding the involvement of AS in ASD, proposing novel avenues for future research. The comprehensive *in silico* validation pipeline employed here showcases the potential of such methods in deciphering splicing perturbations.

However, as with any computational approach, *in vitro* validation is paramount to fully comprehend the functional consequences of these predicted splicing changes. Further studies should expand on the identified splicing events' downstream effects on protein function, signaling pathways, and cellular processes. Integration of our findings with multi-omics data, such as full-length isoform transcriptomics, may provide a more holistic view of the intricate ASD molecular network. Moreover, examining the potential interplay between splicing and other regulatory mechanisms, such as epigenetics, could elucidate additional layers of complexity in ASD etiology.

While numerous questions remain unanswered, and the functional validation of most predicted splice-disrupting variants is still necessary to affirm a molecular diagnosis, *in silico* tools' predictions can serve as supportive evidence in variant classification. Commonly used variant annotation tools are not designed to assess the deleterious impact of splicing variants and their predictions are largely restricted to canonical splice sites. However, if multiple computational sources, such as the framework presented here, indicate that a variant has a deleterious effect, these predictions can be employed. Furthermore, they can aid in prioritizing splice-disrupting variants for subsequent functional testing or experimental validation.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CHAPTER 3

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

7 CHAPTER 3

While there is no question that non-coding variation plays a role in ASD, initial WGS studies were limited in scope to less than 100 parent-child trios[232,344]. It was only 5 years ago that WGS studies significantly expanded, encompassing larger cohorts ranging from 200 to 500 families[75,84,96,133].

Considering the expectation that non-coding mutations may vary widely in functional impact, with only a small fraction likely to exhibit strong effect sizes, detecting associations has proven challenging[41,95]. Under this rationale, power calculations indicate that identifying such signals would necessitate a very large cohort[49,95]. To overcome this, most published studies have restricted to "candidate" non-coding elements, which involves a priori prediction of which regulatory elements of the non-coding genome are important for disease risk[232,345]. In analogy to candidate gene studies, which have consistently struggled to produce robust and replicable results (see section 3.2.1), the selection of candidate regulatory regions is anticipated to yield similar outcomes, due to the multitude of possible combinations involving annotations, cell types, brain regions, and developmental stages[315].

Conversely, other studies have demonstrated that the overall contribution of *de novo* non-coding mutations is comparable to that of LoF coding mutations and missense mutations, although not all ASD proband will have impactful non-coding variation[201]. However, a major challenge has been the definition of functional elements and the interpretation of mutational effects.

7.1 OBJECTIVE

In order to overcome the aforementioned limitations, we have adopted an unbiased methodology, in which candidate non-coding regulatory elements were chosen solely based on their activity in tissues relevant to ASD, specifically the brain and GI tissues. This selection was facilitated by leveraging large-scale initiatives such as the ENCODE project, which aims to systematically catalog the non-coding genome across various cell types and tissues. Moreover, samples were selected on the basis of previous negative results for microarray or WES, a widely discussed approach known for significantly enhancing the diagnostic yield in WGS[346].

Following targeted sequencing of the designated regions, we conducted the necessary filtering and quality control procedures to accurately identify both *de novo* and rare inherited mutations within the selected regulatory elements. These mutations were then prioritized using Sei, a recently developed framework for systematically forecasting sequence regulatory activities. The objective was to analyze the mutational burden of probands not merely based on the count of mutations, which are susceptible to both statistical power challenges and confounding factors, such as the rise in mutation counts with parental age. Instead, the focus

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

was on evaluating the functional impact of mutations, enabling an assessment of their relevance and potential association with ASD.

Moreover, in order to gain a deeper insight into how these variants may influence disease risk, we performed several approaches. These included candidate gene selection (by using PLAC-seq data from others), gene prioritization (based on previous literature) and biological *in silico* characterization of possible effects.

7.2 METHODS

7.2.1 Targeted sequencing and data processing

7.2.1.1 Cohort description

Please refer to section 6.2.1 in the preceding chapter for the description of the initial cohort from which we obtained the results for CNV and WES analysis.

7.2.1.2 Sample selection

Out of the 360 trios with complete phenotype information, microarray and exome sequencing data, we selected cases with negative results for CNVs and WES (mutations classified as benign, likely benign or VOUS (variant of uncertain significance)). By selecting cases on the basis of the absence of *de novo* LoF mutations or large *de novo* CNVs in prior WES and microarray data, we are intentionally enhancing the sample for undiscovered non-coding risk variants.

After this exclusion, individuals with a syndromic form of ASD were also discarded, in order to avoid ascertainment biases.

Following this, samples were randomly selected, leading to a total sample size of 200 trios (100 from Santiago, 100 from Madrid), including data for 39 female probands and 161 male probands.

Among the most prevalent comorbidities in our cohort, the following stand out: 41.5% of our patients exhibit ID, 7% have ADHD, and 6.5% experience epilepsy.

7.2.1.3 Selection of the regulatory regions of interest

For the selection of the regulatory regions of interest, we employed the candidate cis-regulatory elements (cCREs) Registry from ENCODE[347,348] (version 2, https://screen-v2.wenglab.org/). Launched in 2003, the ENCODE project aimed to create a comprehensive map of functional elements within the human genome. This encompasses not only genes, but also biochemical regions associated with gene regulation, such as transcription factor binding sites, open chromatin, and histone marks, as well as various transcript isoforms. The ENCODE registry v2 contained 926,535 human cCREs in GRCh38 and covered 839 human cell and tissue types[347].

Brain and GI tissues with Dnase-seq data available were selected, both from adult (n = 14) and embryonic tissue (n = 59).

ENCODE integrates high-quality DNase-seq and H3K4me3, H3K27ac, and CTCF ChIP-seq data. Thus, the cCREs within the registry of ENCODE consist of a subset of DNase hypersensitivity sites (DHSs), which defines open (active) chromatin regions, in combination

with both histone modifications and CTCF-binding data. As such, cCREs are divided into (i) promoter-like signatures (PLS), with enriched trimethylation on histone H3 lysine 4 (H3K4me3)[349]; (ii) enhancer-like motifs (ELS), defined by high acetylation levels on histone H3 lysine 27 (H3K27ac)[349]; (iii) DNase-H3K4me3 signatures (with unknown function), and (iv) CTCF-only regions (which demarcates chromatin boundaries[237]); in a cell-type agnostic manner. Moreover, ELS are further grouped in proximal signatures (pELS), inside the region defined by < 2000 bp downstream of the TSS, and distal (dELS), > 2000 bp upstream of the TSS) (Figure 22).

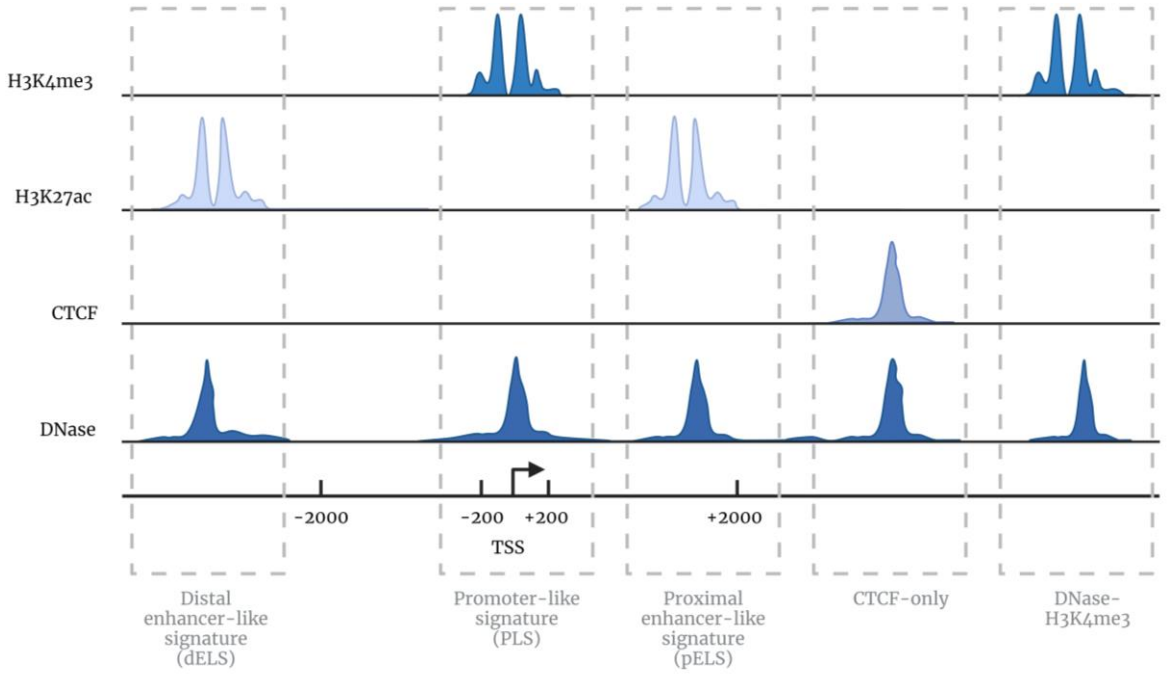


Figure 22. cCREs' classification in the ENCODE Encyclopedia. Created with BioRender.com.

For each tissue, cCREs labeled as "Low-Dnase" were excluded, as they are inactive in the given tissue[347].

From the total of 926,535 human cCREs, we selected those that exhibit activity in a greater number of the interrogated tissues. Thus, we selected cCREs that were active in 36 or more tissues (n = 85,394 cCREs).

7.2.1.4 Targeted sequencing

Selected regions were sequenced at the National Center for Genomic Analysis (CNAG) using the KAPA HyperChoice Target Enrichment custom probes.

Samples with sex discrepancies when compared to reported pedigrees were dropped and replaced, along with all other samples from the same trio. Moreover, samples which failed CNAG's quality control (DNA <100 ng / critical degradation (genomic quality number (GQN) < 3.3) were also removed and consequently substituted, leaving 71 trios from Santiago and 129 from Madrid.

Sequencing reads were aligned to GRCh38/hg38 using the Burrows-Wheeler Aligner[350].

SNVs and small indels (< 50 bp) were discovered using the Genome Analysis Toolkit (GATK)[351] HaplotypeCaller package version 3.4 (https://github.com/broadinstitute/gatk).

6.2.1.5. Data processing

7.2.1.4.1 Joint genotyping

Raw results were downloaded as single-sample gVCFs (genomic variant call format). While a variant call format (VCF) is a text file format that describes genetic variations detected in sequencing data, such as SNPs and indels, a gVCF contains a record for every position in the genome (or interval of interest), irrespective of whether a variant was detected at that site or not.

All individual gVCFs were combined into one multisample gVCF using GATK version 4.2.2.0. For this purpose, we imported single-sample gVCFs into GenomicsDB (as recommended by the GATK4 Best Practice Workflow for SNP and Indel calling[352]). In brief, this step transposes sample-centric variant information across genomic loci to make data more accessible to tools used in the next steps (https://gatk.broadinstitute.org/hc/en-us/articles/360057439331-GenomicsDBImport).

Then, SNVs and indels were jointly called across all samples with GenotypeGVCFs (https://gatk.broadinstitute.org/hc/en-us/articles/360037057852-GenotypeGVCFs), producing a final multi-sample VCF file (format VCFv4.2).

7.2.1.4.2 Variant quality score recalibration (VQSR)

Variant call accuracy was estimated using the GATK Variant Quality Score Recalibration (VQSR) method, following the recommendations provided in the GATK4 Best Practice Workflow for SNP and Indel calling[352].

To obtain further details about building the recalibration model and the covariates used for variant recalibration, consult appendix 10.3.

In summary, we developed the recalibration model based on call sets that have been validated to a very high degree of confidence. This model is subsequently applied to evaluate both known and novel variations identified within our own call set. The primary objective of this evaluation is to ascertain the likelihood that each call represents a genuine variant.

Moreover, the recalibration model takes into consideration the spatial distribution of these variants concerning specified annotations, by categorizing them into distinct clusters.

Subsequently, the clustering information is leveraged to assign variant quality score log odds (VQSLOD) scores to all identified variants, which represents the likelihood of a variant being a true variant rather than a false positive call. Notably, variants positioned closer to the central core of a cluster receive higher VQSLOD scores in comparison to outliers.

At the end of the recalibration process, variants are sorted by VQSLOD scores and grouped into different tranches. Tranches are “slices” of variants delimited by the user-specified threshold values. These thresholds refer to the sensitivity that can be obtained when applied to the call sets used for training the model. The concept is that the lowest tranche is highly specific but less sensitive, meaning there are few false positives but potentially many false negatives

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

(*i.e.*, missing calls). Each subsequent tranche introduces additional true positive calls along with an increasing number of false positive calls.

In the given context, we initially delineated threshold tranches at percentiles 90%, 95%, 98%, 98.9%, 99.99%, and 100% (refer to Figure 23), to obtain a broad overview of variant distribution. Subsequently, we opted for the 98.9% tranche, which involves applying a sensitivity level to our dataset that would allow the retrieval of 98.9% of true variants from the truth training sets (HapMap and Omni SNPs). The model then utilizes the corresponding value of VQSLOD as a threshold to filter the variants, categorizing them as PASS variants. Variants assigned to other tranches, such as the "VQSRTrancheSNP99.90to100.00" which correspond to the remaining 0.1% of the truth set in this example, are deemed false positives.

If we wanted an increased sensitivity, albeit at the potential cost of increased false positives, higher tranches could be considered. In contrast, if we wanted to be more precise, we could have taken the lowest tranche, which would only retrieve 95% of the truth sites but 5% of the false positives. However, our selection of the 98.9% tranche is justified by the following reasons: (i) the 98.9% tranche yields the highest number of true positives, and subsequent tranches, despite introducing additional true calls, escalate the number of false positives by a comparable, if not greater, magnitude; (ii) the ratio of transitions to transversions (Ti/Tv ratio), an approximate measure of quality (lower Ti/Tv ratios are associated with higher false positives)[352], does not exhibit a significant decrease in comparison to lower tranches (90-95%), although it shows a substantial improvement from the higher tranches (refer to Figure 23).

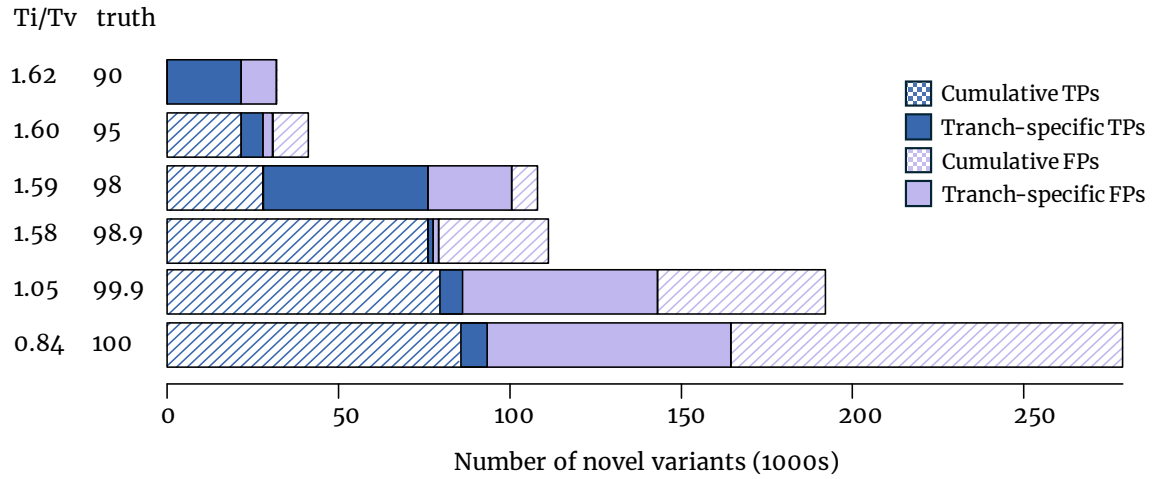


Figure 23. Tranches plot generated by the VQSR. The x-axis represents the count of putative novel variants, distinguishing between true-positive and false-positive variants. Meanwhile, the y-axis displays two quality metrics: the novel transition to transversion (Ti/Tv) ratio and the overall truth sensitivity, where TPs denote true positives (variants present in both our callset and the truth dataset), and FPs represent false positives (variants in our callset but absent in the truth dataset). This figure was created with the "–tranches-plot" parameter integrated in the VQSR workflow.

7.2.2 Analysis

A summary of the quality control and data cleaning procedure is depicted in Figure 24.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

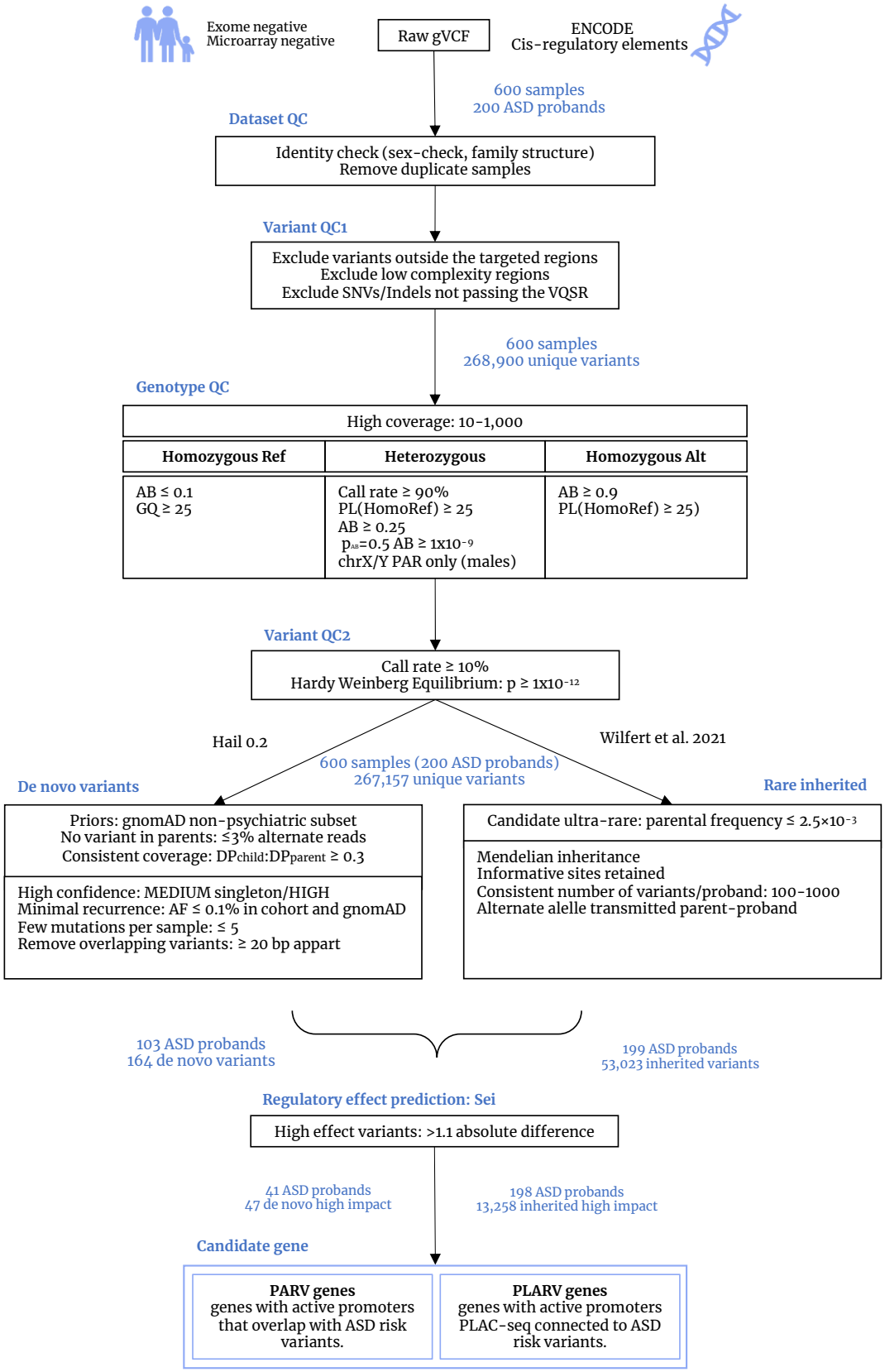


Figure 24. Workflow for variant quality control (QC) and filtering and variant/gene prioritization. *AB, allele balance; AF, allele frequency; bp, base pairs; DP, depth of coverage; PAR, pseudoautosomal regions; PL, phred-scaled likelihood.*

For variant and genotype QC, as well as optimization of *de novo* variant discovery, we utilized the filtering process established by Satterstrom *et al.*[40]. This process was previously optimized and implemented in an ASD cohort, and we adapted it for our dataset. For inherited variant discovery, we implemented the pipeline defined in Wilfert *et al.*[76].

7.2.2.1 Dataset Quality Control

The final VCF file, containing SNVs and indels for the 200 trios (600 samples), was loaded into Hail 0.2 (https:// hail.is/; https://github.com/hail-is/hail), an open-source framework designed for scalable exploration and analysis of data, with a specific focus on genomics.

Then, multiallelic sites were split into biallelic sites (resulting in 719,209 unique variants).

To verify the accuracy of the provided pedigree information, relatedness was assessed between each pair of samples using Hail's ibd() function.

The relatedness values were manually checked, and inferred pedigree structures were compared to reported pedigrees to search for discrepancies. Duplicate samples were identified by using an identity value higher than 0.8 (pi-hat > 0.8) and related samples were identified by using an identity value higher than 0.4 (pi-hat > 0.4)[353].

No obvious errors in reporting were detected.

7.2.2.2 Variant-level quality control

For the variant filtering steps, bcftools[354] (https://samtools.github.io/bcftools/bcftools.html) was employed unless otherwise specified.

After filtering the VCF file for the selected coordinates of interest to eliminate off-targets or amplified regions outside cCREs, low-complexity regions and variants that failed VQSR were excluded due to their high error rate or anomalous characteristics[41]. Indels were required to have a VQSLOD of ≥ -0.6549, while SNVs were required to possess a VQSLOD of ≥ 5.2528.

Following these steps, a total of 268,900 unique variants remained.

7.2.2.3 Genotype-level quality control

To ensure the quality of genotypes, various filters were implemented during the genotype quality control process.

Initially, we excluded calls with a depth of coverage (DP) < 10 or > 1,000 (as per the distribution of DP values). Moreover, we omitted any call if the GQ was < 25.

For homozygous reference calls, genotypes were then filtered if they had < 90% of the read depth supporting the reference allele.

Homozygous variant calls underwent filtering for genotypes with < 90% of the read depth supporting the alternate allele or a Phred-scaled likelihood (PL) of being homozygous reference < 25. PL is a representation of the probability of a particular genotype, it tells us how confident we are that the assigned genotype is correct[351]. In this context, a PL of being homozygous reference < 25 (standard cutoff), means that the algorithm has assigned a low likelihood to the possibility that the genotype is homozygous reference.

Heterozygous calls were filtered based on genotypes with < 90% of the read depth supporting either the reference or alternate allele, a PL of being homozygous reference < 25, <

25% of the read depth supporting the alternate allele (*i.e.*, an allele balance (AB) < 0.25), or a probability of the AB (calculated from a binomial distribution centered on 0.5) < 1 x 10^{9}. The probability of the AB refers to the likelihood of having a certain distribution of alleles (reference *versus* alternate allele). The binomial distribution centered on 0.5 assumes an equal probability of observing either the reference or alternate allele. By choosing the stringent criterion for this probability to be $1x10^{-9}$, we are considering only extremely imbalanced allele distributions as acceptable. This, in turn, reflects a high degree of confidence in the determination of allele balance.

We additionally omitted any heterozygous call in the X or Y non pseudoautosomal (PAR) regions (in a sample that imputed as male (using BEDTools[355], https://github.com/arq5x/bedtools2). This step enhances the reliability of the inferred calls and ensures data accuracy: males are not expected to be heterozygous in non-pseudoautosomal regions, since these regions are unique to males and lack a homologous counterpart on the X chromosome[356].

Following the implementation of these filters, variants with a call rate < 10% or a Hardy-Weinberg equilibrium (HWE) p-value <1 x 10^{12} were excluded. The HWE is a principle in population genetics, which states that "genotype frequencies in a population remain constant between generations in the absence of disturbance by outside factors"[357]. In this context, deviations from HWE may suggest issues such as genotyping errors, population stratification, selection, or other factors altering the genetic structure of the population. Thus, we excluded variants with a highly significant deviation from the expected genotype frequencies according to the HWE.

Following the application of these filters and the exclusion of sites that were no longer variants, the dataset comprised 267,157 distinct variants in 200 ASD trios. This dataset was then used as the starting point for the *de novo* and inherited workflows.

7.2.3 *De novo* variant detection

De novo variants were identified within the previously described 600-sample dataset.

Variant calling for DNVs utilized the de_novo() function implemented in Hail 0.2. To establish population allele frequencies for all the variants in our dataset, data were sourced from the non-psychiatric subset of gnomAD (https://gnomad.broadinstitute.org/), which includes samples not ascertained for neurological or psychiatric phenotypes (non-neuro subset). Subsequently, these frequencies were utilized as input priors.

Additional criteria for variant calling included the following: (i) homozygous reference genotypes in parents should exhibit no more than 3% of reads supporting the alternate allele, (ii) children's heterozygous calls were required to have a minimum of 30% of reads supporting the alternate allele, and (iii) the ratio of child DP to parental DP needed to be at least 0.3. This procedure resulted in the identification of 11,311 putative *de novo* variants.

To ensure the quality of *de novo* variants, we retained variants that were classified as high confidence by Hail or were of medium confidence and represented singletons in the dataset (resulting in the inclusion of 10,383 putative *de novo* variants). Subsequently, any variant with an allele frequency exceeding 0.1% across the samples in our dataset (of note, any variant

yielded such frequency) or in the non-neuro subset of gnomAD was excluded (leading to the exclusion of 9,739 putative *de novo* variants).

Moreover, samples were excluded if they had more than 5 *de novo* variants (3 samples excluded with 206, 171 and 101 *de novo* variants, respectively), based on z-scores ((data value-mean)/ standard deviation) ≥ 3.

Additionally, we filtered out variants that were less than 20 base pairs apart from each other to mitigate the likelihood of false positives (resulting in the exclusion of 2 putative *de novo* variants).

After applying these filters, the resultant list of high confidence *de novo* variants included 164 *de novo* variants from 103 probands.

7.2.4 Rare inherited variation

As with *de novo* variation, we used the dataset of 600 samples and 267,157 unique variants described above as a starting point to identify high confidence ultra-rare transmitted variants. For this purpose, we employed the pipeline described by Wilfert *et al.*[76] (https://github.com/EichlerLab/ultra_rare_transmitted).

In summary, the pipeline assesses Mendelian inheritance across all sites, and identifies the parent of origin of each variant call exclusively for informative sites (those specific to one parent and transmitted to the child). Subsequently, it examines the transmission pattern, enabling the identification of the corresponding recipient child. Following this, it assesses the genotypes across parents, distinguishing between heterozygous and homozygous reference/alternate states. Then, the extraction of ultra-rare variants is based on the criteria of parent heterozygous counts < 10 and homozygous alternate = 0 (per-family basis)[76].

Following this method, all sites that were heterozygous and observed only once in the parent population (parental frequency $\leq 2.5\times10^{-3}$) were designated as candidate ultra-rare private variants.

Subsequently, only variants which did not violate the rules of Mendelian inheritance were kept, resulting in a total of 84,311 variants.

Then, only informative sites were retained, leaving 55,903 variants. After excluding one outlier family with 2,880 transmitted variants (z-score = 13.19), we further filter the variant call set to include only those sites where the parent transmitted the alternate allele to the proband, resulting in a final set of 53,023 variants identified in 199 ASD probands.

7.2.5 Regulatory impact prediction

De novo and inherited variants were evaluated to determine their impact on regulatory activities using the Sei framework[348] (https://github.com/FunctionLab/sei-framework), installed with a local Anaconda environment.

Sei provides a comprehensive mapping of any given sequence to regulatory activities, classified in 40 distinct sequence classes. It further provides quantitative scores that represent changes in regulatory activities[348].

This mapping incorporates predictions based on 21,907 chromatin profiles, including those of TFs, histone marks, and chromatin accessibility profiles across a diverse set of cell types.

The sequence classes are defined based on 30 million sequences tiling the genome, thus covering a wide range of sequence activities. For ease of interpretation, sequence classes are grouped into categories such as P (Promoter), E (Enhancer), CTCF (CTCF-cohesin binding), TF (TF binding), PC (Polycomb-repressed), HET (Heterochromatin), TN (Transcription), and L (Low Signal) (Table 19, Figure 25)[348].

Table 19. Sequence classes defined in Sei.

Sequence class label	Sequence class name	Group
PC1	Polycomb / Heterochromatin	PC
L1	Low signal	L
TN1	Transcription	TN
TN2	Transcription	TN
L2	Low signal	L
E1	Stem cell	E
E2	Multi-tissue	E
E3	Brain / Melanocyte	E
L3	Low signal	L
E4	Multi-tissue	E
TF1	NANOG / FOXA1	TF
HET1	Heterochromatin	HET
E5	B-cell-like	E
E6	Weak epithelial	E
TF2	CEBPB	TF
PC2	Weak Polycomb	PC
E7	Monocyte / Macrophage	E
E8	Weak multi-tissue	E
L4	Low signal	L
TF3	FOXA1 / AR / ESR1	TF
PC3	Polycomb	PC
TN3	Transcription	TN
L5	Low signal	L
HET2	Heterochromatin	HET
L6	Low signal	L
P	Promoter	P
E9	Liver / Intestine	E
CTCF	CTCF-Cohesin	CTCF
TN4	Transcription	TN
HET3	Heterochromatin	HET
E10	Brain	E
TF4	OTX2	TF
HET4	Heterochromatin	HET
L7	Low signal	L
PC4	Polycomb / Bivalent stem cell Enh	PC
HET5	Centromere	HET
E11	T-cell	E
TF5	AR	TF
E12	Erythroblast-like	E
HET6	Centromere	HET

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

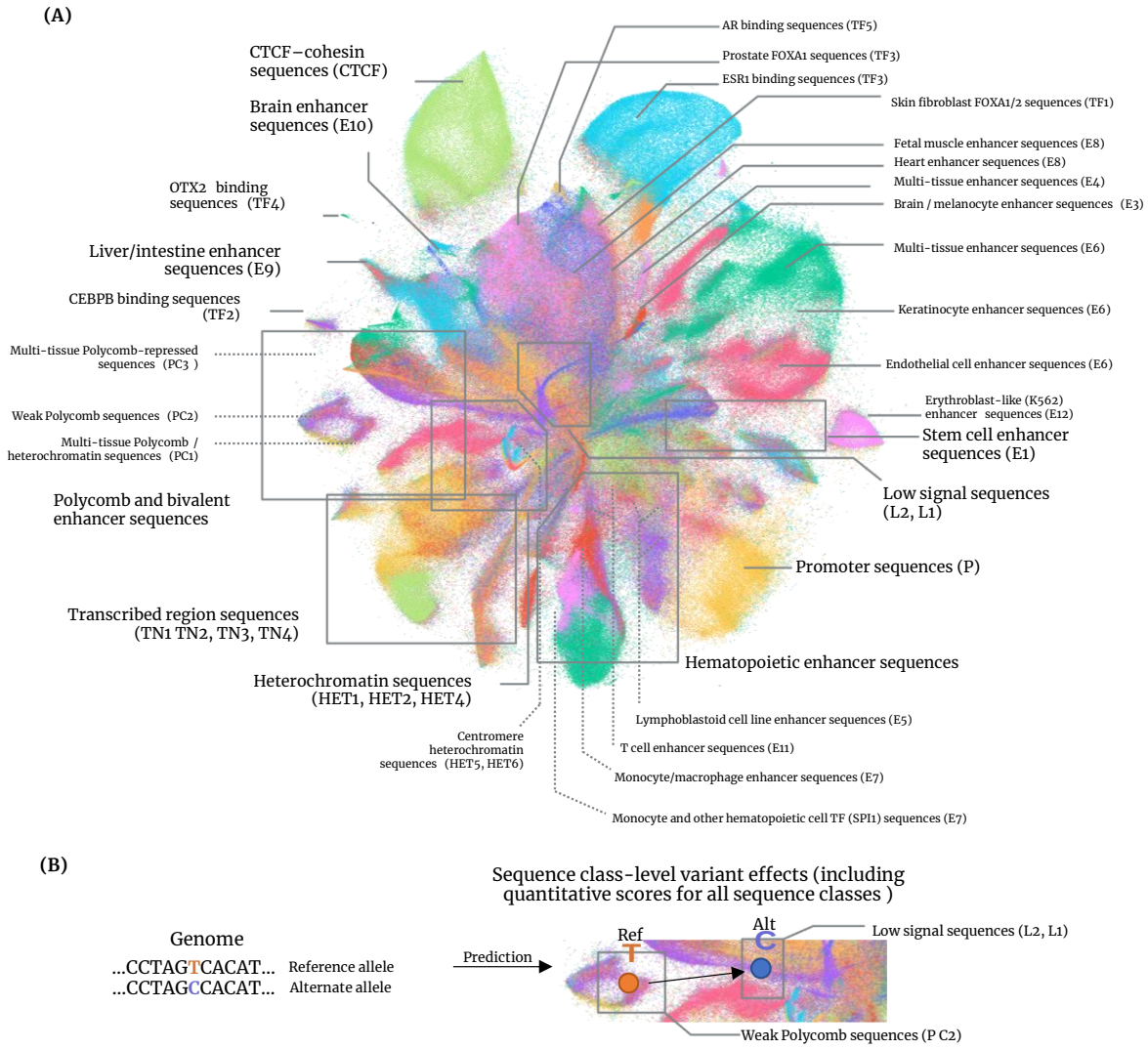


Figure 25. Sei, a deep convolutional network sequence model to globally assess the regulatory landscape of the genome. Panel A represents the visualization regarding the global regulatory landscape of the human genome sequences. Panel B represents the prediction of sequence class-level variant effects, quantified by changes in sequence class scores. In an illustrative sequence, a regulatory role initially assigned to the sequence class PC2 loses its activity with the alternate allele. Consequently, it is reassigned to the sequence class L1. This figure was re-created from Chen *et al.*[348] thanks to the repository the authors created, that includes the code to generate the illustrations from the manuscript (https://github.com/FunctionLab/sei-manuscript/tree/main). The code to generate this figure was adapted from this source and the figure was created with a jupyter notebook.

To enhance the interpretability of variant effects, the computation considers nucleosome-occupancy-adjusted scores. Normalizing for nucleosome occupancy is crucial in representing the predicted variant effects on histone marks. For instance, a LoF mutation near the TSS might decrease the H3K4me3 modification level while increasing nucleosome occupancy. This dual effect results in an overall increase in the observed quantity of H3K4me3[348].Therefore, the computation of the variant effect is expressed as the difference, nucleosome-occupancy-adjusted, between the alternate and reference scores of the raw sequence class scores[348].

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Within this context, the Sei framework was utilized to acquire sequence class score predictions for the final call set of *de novo*/inherited variants.

For every mutation, we predicted the sequence class scores for both the reference and alternate alleles and computed the sequence class-level variant effect as the predicted scores for the alternate allele subtracting the scores for the reference allele.

For mutations with a strong effect in a different sequence class than the originally assigned sequence class (absolute value higher than the original sequence class by > 1.1 difference[348]) we reassigned the mutation to the sequence class with the strongest effects.

Moreover, variants were annotated in terms of positive effect predictions (increase in the assigned activity of a sequence) or negative effect predictions (decrease in the activity).

These filtering yielded 47 *de novo* variants, in 44 ASD probands with changes in sequence class activity and 13,258 ultra-rare inherited variants in 198 probands.

7.2.6 Candidate gene elucidation

After applying the prioritization criteria for their impact on regulatory activities, we aimed to perform a candidate gene selection. For this purpose, we leveraged brain cell-type specific interactome maps to further confirm our non-coding regulatory regions and pinpoint genes likely to be regulated by these non-coding variants.

In their work, Nott and colleagues characterized the enhancer-promoter interactome of three different human brain cell types (namely neurons, microglia and oligodendrocytes) *in vivo*. In brief, they characterized open regions of chromatin with ATAC-seq and identified active chromatin regions and promoters with ChIP-seq analysis of the H3K27ac and H3K4me3 epigenetic marks, respectively. Following this rationale, they defined putative active promoters (ATAC-seq and H3K4me3 enrichment overlapping H3K27ac peaks within 2,000 bp to a nearest TSS) and enhancers (ATAC-seq plus H3K27ac enrichment in a region outside H3K4me3 peaks) in each cell type. Moreover, they established the relationship between promoters and distal regulatory regions through the application of proximity ligation-assisted ChIP-seq (PLAC-seq). In this method, proximity ligation precedes the enrichment for active promoters using H3K4me3 ChIP-seq. This allowed for the characterization of chromatin loops between active promoters and distal enhancer/superenhancers (*i.e.*, clusters of multiple enhancers)[109] (Figure 26).

To better understand the likely functional consequences of our variant call set, we defined putative ASD-risk genes according to whether they exhibit active promoters overlapping with any variant in the call set and/or are linked to active promoters through PLAC-seq defined chromatin interactions.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

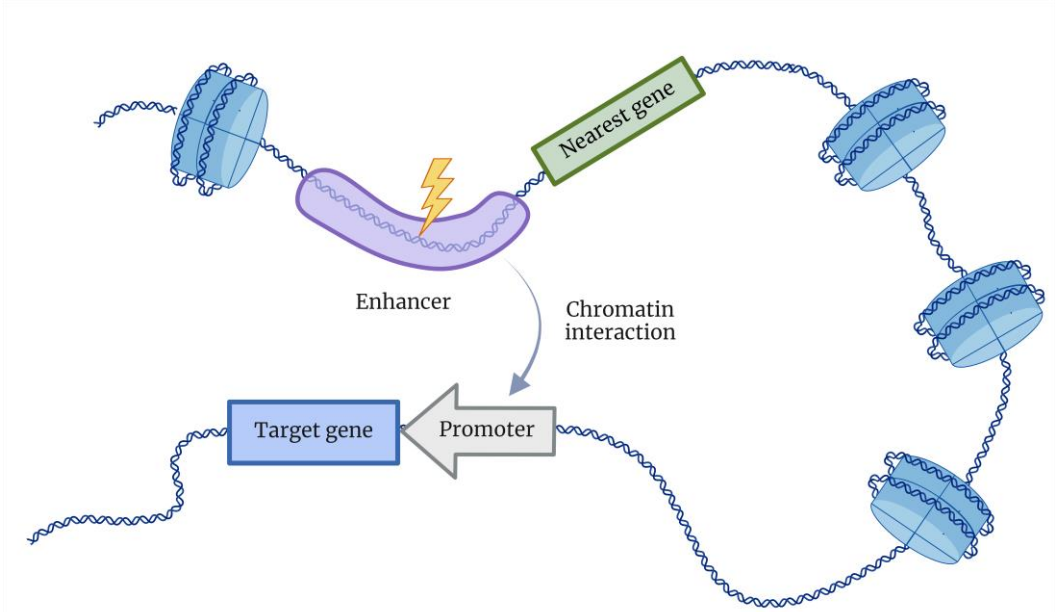


Figure 26. Long range chromatin interactions. Despite its genomic localization in close proximity to a nearest gene, one enhancer may act to regulate the transcription of a gene many kilobases away through chromatin looping. Variants affecting enhancers (or disrupting chromatin interactions) may lead to a shift in promoter preferences and consequently abolish the expression of the target gene. Re-created from Kim *et al.*[142] with BioRender.com.

Initially, the analysis involved determining if each variant intersected with an active promoter in each cell type. Variants that did overlap an active putative promoter were assigned to the corresponding gene, designating these as "genes containing promoters harboring ASD regulatory variants" (from now on, PARV genes).

Subsequently, each variant was extended by 2,500 bp upstream/downstream to correspond with the PLAC-seq bin size of 5,000 bp. These variant-extended regions were then intersected with PLAC-seq bins in each cell type. If a region overlapped one or more PLAC-seq bins distally linked to an active promoter, that region was then assigned to the corresponding "gene containing a promoter PLAC-linked to ASD regulatory variants" (from now on, PLARV genes).

7.2.7 Variant prioritization

We applied additional criteria to prioritize non-coding variants that were already categorized as having a significant regulatory effect, in regard to the putative ASD gene they are regulating (see previous sections 7.2.5 and 7.2.6), with the aim of considering them as potential causative factors for each affected individual.

For this purpose, we first curated a consensus list of known ASD risk genes from different sources. We collected 291 known ASD genes from the following sources: (i) high-confidence (score = 1) ASD genes collected in SFARI (accessed the 11/2023), (ii) dominant ASD genes reported in recent literature and included in the SPARK genes list (ref) (11/2023), (iii) unique genes from recent large ASD genetic studies with an FDR <0.1 (these included references [40,89,358]), and (iv) autosomal genes with exome-wide significance ($p < 2.5 \times 10^{-6}$) after combining p-values from enrichment of all *de novo* variants, transmission disequilibrium test (TDT) of

rare, inherited LoFs from unaffected parents to affected offspring, and increased rate of LoFs in cases vs population controls[201].

We then cross-referenced the final dataset of PLARV genes with the custom list of ASD-risk genes and selected those variants where the corresponding PLARV gene was present in the list.

The rationale for excluding PARV genes linked to the variants, as we will delve into later, is that, in certain instances, risk variants were associated with more distant active promoters rather than the nearest gene promoter. A connection with the gene was deemed validated only under functional evidence (*i.e.*, PLAC interactions with PLARV genes) and not genomic localization (variants nearby a PARV gene's TSS).

7.2.8 Protein-protein interaction (PPI) network

We conducted PPI network analyses and computed network statistics using Cytoscape v.3.7.2, an open source software platform for visualizing complex networks (https://cytoscape.org/).

We employed the multiple protein input option of the STRING database v.11. with default settings, except that we restricted interactions to those of high confidence (0.70). STRING constitutes a biological database encompassing both known and predicted protein-protein interactions. Its implementation is accessible within Cytoscape.

7.2.9 Cluster profiler

Refer to section "6.2.5.2 Enrichment analysis" in the previous chapter for specific details. The same parameters were employed here.

7.2.10 Transcription factors (TF) enrichment

The analysis of TFs with enriched binding sites in the base positions affected by *de novo* and inherited high regulatory impact variants was conducted using the Enrichment Analysis tool in ChIP-Atlas (https://chip-atlas.org/enrichment_analysis). The specified parameters and input files were as follows: Cell type Class = Neural, Experiment type = ChIP: TFs and others, Threshold for Significance = 50, Dataset A = BED file of the positions of the variants, Dataset B = random permutations of dataset A (x 1000).

7.2.11 Presence in topologically associated domains (TADs)

The interval list of TADs in human tissues and cells (BED format, hg19), was acquired from iPSC-derived neurons. This information was sourced from the Gene Expression Omnibus (GEO) database under the accession number GSE79965 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE79965).

To identify potential associations, prioritized variants were cross-referenced with TAD boundaries. This process aimed to detect instances where multiple variants occurred within the same TAD, providing insights into the spatial organization of genomic elements in the context of these prioritized genetic variations.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

7.2.12 Statistical significance tests

All statistical analyses were conducted in R.

7.3 RESULTS

7.3.1 Dataset

We analyzed targeted sequencing data in a cohort of 200 ASD trios, with a ratio of ~4:1 male probands to females (161 males, 39 females; male-to-female ratio = 4.13), in line with previous estimates[4].

Based on the evidence of GI and brain tissues playing a role in ASD's etiology[116], regions were selected according to their activity in these tissues. This yielded a total of 85,394 cCRES active in 25 brain-tissue samples and 48 GI-tissue samples, spanning a total of 21.35 Mb of the genome (*i.e.,* 0,68% from the total length, 0,70% from the non-coding length)[35].

The vast majority of selected cCREs corresponded to the ENCODE classification of distal/proximal enhancer-like signature (n = 61,822; 72.40%) whereas promoters were represented to a lesser extent (n = 20,306; 23.78%) (Figure 27), as expected in the basis of the one-to many relationship between promoters and enhancers[109].

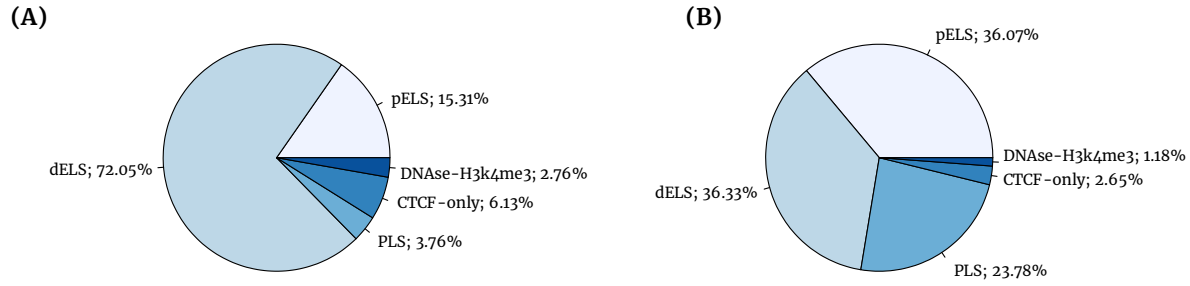


Figure 27. Pie chart depicting the classification of cCREs. Panel A illustrates the distribution of cCREs in the ENCODE registry, while Panel B displays the percentages within our dataset's selection.

7.3.1.1 *De novo* variants

From the trio-based data, we identified 164 rare *de novo* variants in the selected candidate regulatory regions (allele frequency ≤ 0.1% in our dataset and in the non-psychiatric of gnomAD), with 51,50% of cases carrying at least one such variant (103 ASD probands; $\bar{x}$ = 1.59 ± 0.84 variants/child). These variants were present inside an ELS cCRE (both distal and proximal) in 67.08% of the instances, whereas 31.10% were harbored in PLS. The remaining variants were located inside a CTCF-only cCRE (1.22%) and DNAse-H3K4me3 cCREs (0.61%) (Figure 28 A).

7.3.1.2 Inherited variants

To study ultra-rare inheritance, we included variants with an allele frequency ≤ 2.5×10^{-3} in the parent population and selected those transmitted to the child. After removing one outlier family with inconsistencies in respect to the average number of mutations per family, 53,023 inherited variants were detected in 199 cases ($\bar{x}$ = 266.45 ± 68.93 variants/child). These variants were present inside an ELS cCRE in 67.02% of the instances, whereas 29.45%

were harbored in PLS cCREs. The remaining variants were located inside a CTCF-only cCRE (2.39%) and DNAse-H3K4me3 cCREs (1.14%) (Figure 28 B).

Of note, we observed 26,960 informative sites with a mother-proband inheritance, and 26,063 informative sites with a father-proband inheritance (50.85% *versus* 49.15% of the instances, respectively). The observed differences between these two inheritance patterns were statistically significant ($p = 9.97 \times 10^{-5}$; two-sided binomial exact test), consistent with the maternal transmission bias observed for large and small CNVs and inherited private truncating SNVs[72,232,359,360]. This bias remained significant even when restricting to autosomal chromosomes ($p = 7.8 \times 10^{-3}$, two-sided binomial exact test).

Additionally, when restricted to the X chromosome, most of the variants were maternally inherited (97.98% with maternal inheritance pattern *versus* 2.02% paternal inheritance pattern), as would be expected. Intriguingly, all variants exhibiting an X-linked maternal inheritance pattern were consistently transmitted to an affected son ($p = 2.2 \times 10^{-16}$; two-sided binomial exact test) and in any case, to a daughter.

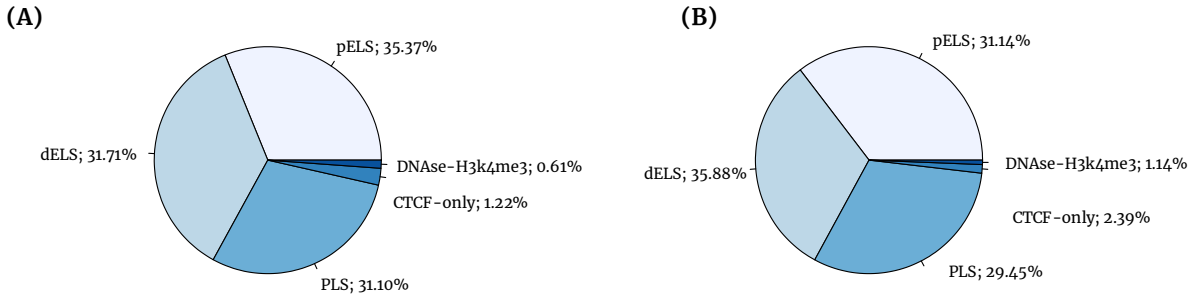


Figure 28. Pie chart representation of *de novo* and inherited variants based on cCREs' classification. Panel A illustrates the classification for *de novo* variants, while Panel B depicts the classification for inherited variants.

7.3.1.3 Variant enrichment in cCREs (candidate cis-Regulatory Elements)

To assess potential statistical enrichments or depletions in the number of *de novo* and inherited variants per cCRE category, we conducted two-sample proportions tests.

Comparative analysis with the proportion of each cCRE in the original selection (Figure 27 B) revealed an enrichment of *de novo* and inherited variants in PLS cCREs (p-value for proportion test = 0.02 and $< 2.2 \times 10^{-16}$, respectively) (Table 20). Additionally, inherited variants were depleted in CTCF cCREs (p-value = 1.46×10^{-3} for the proportion test) and enriched in pELS cCREs (p-value = $< 2.2 \times 10^{-16}$ for the proportion test).

Table 20. Enrichment of cCREs in *de novo* and inherited variants.

cCRE	Original selection	*De novo*	p-value	Inherited	p-value
dELS	31,023	52	0.25	19,027	0.09
pELS	30,799	58	0.92	16,510	**$< 2.2 \times 10^{-16}$**
PLS	20,306	51	**0.02**	15,616	**$< 2.2 \times 10^{-16}$**
CTCF	2,260	2	0.37	1,265	**1.46×10^{-3}**
DNase-H3K4me3	1,006	1	0.75	605	0.55

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

7.3.2 Effect of genetic variants in gene regulation

In order to prioritize variants for their potential impact on gene regulatory activities, we used Sei, a new deep learning sequence model that enables the interpretation of genetic variants.

7.3.2.1 *De novo* variants

De novo variant class scores ranged from 0.07 to 6.68, with 28.66% (n = 47) of the variants yielding absolute differences higher than the stipulated class score > 1.1 difference to consider a variant of high regulatory impact (Appendix 10.2, Supplementary Table 4). These variants were assigned to 10 different regulatory classes in Sei.

Of note, the 47 variants were present in 44 probands, with 3 of them harboring 2 different variants in separate chromosomes ($\bar{x}$ = 1.07 ± 0.25 *de novo* variants/child).

Half of these variants (53.19%) represented the class CTCF, which demarcates topological loop boundaries[237]. Out of the 25 variants assigned to the regulatory class "CTCF CTCF-Cohesin", 24 were classified in ENCODE as PLS/dELS/pELS-CTCF bound, and one variant was categorized as CTCF-only.

For the other regulatory classes, variants were inside ENCODE cCREs defined as PLS/dELS/pELS (CTCF bound or not) without following any apparent trend (*i.e.*, no regulatory class was significantly correlated with a specific type of cCRE) (data not shown).

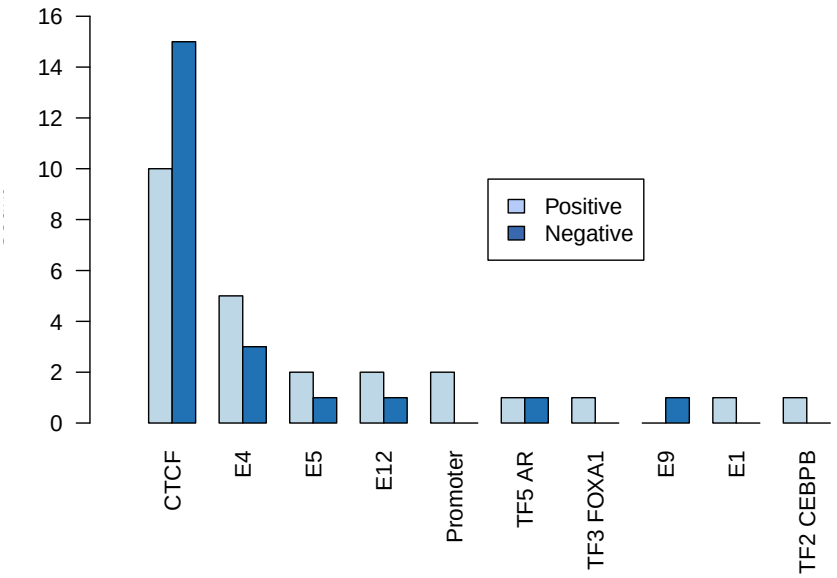


Figure 29. Counts of *de novo* variants assigned to each regulatory class. Regulatory effect scores are categorized into positive and negative scores. No significant differences in positive / negative scores were detected.

7.3.2.2 Inherited variants

Inherited variant class scores ranged from 0.006 to 41.41, with 25.00% (n = 13,258) of the variants yielding absolute differences higher than > 1.1 (Appendix 10.2, Supplementary Table 6) in 30 different regulatory classes.

The average number of inherited variants was $\bar{x}$ = 66.96 ± 17.41, and the maximum observed number of inherited variants in a proband was 182.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

The CTCF class was the most widely represented in our dataset (40.62% of the variants). In line with the results from *de novo* variants, only 162 variants were present in a CTCF-only cCRE, while the rest were classified in ENCODE as PLS/dELS/pELS/DNase-H3K4me3 CTCF bound or not. Moreover, there were no specific correlations between ENCODE's cCRE classification restricted to any Sei-assigned class groups, except for CRCF-only cCREs (p-value = 1.79 x 10^{-6} for the proportion test) (Appendix 10.2, Supplementary Table 5).

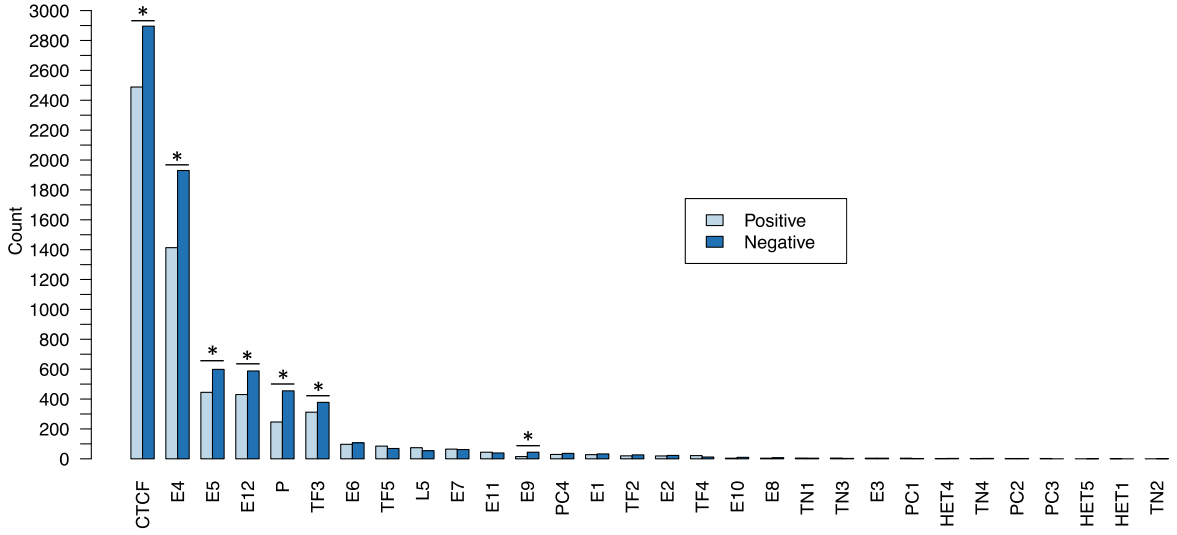


Figure 30. Counts of inherited variants assigned to each regulatory class. Regulatory effect scores are categorized into positive and negative scores. *Significant differences in the number of variants with positive or negative scores (two-sided binomial test).

7.3.2.3 Differences in gene regulation for inherited and *de novo* variants

De novo variants did not yield any significant differences in terms of positive / negative scores neither for the CTCF category (10 variants gained affinity for CTCF binding, and 15 lost this affinity; p-value = 0.42, two-sided binomial test), nor for the rest (Table 21, Figure 29).

For inherited variants, however, we found a significant enrichment in negative scores within the CTCF regulatory class (*i.e.*, a significant decrease in the affinity for CTCF; p-value = 1.55 x 10^{-8}, one-sided binomial test) (Table 21, Figure 30). Besides the CTCF class, 6 additional regulatory classes yielded a significant enrichment in negative scores (Table 21).

Table 21. Enrichment of variant scores within each of the regulatory classes, divided in positive and negative effects.

Regulatory class	*De novo*		p-value	Inherited		p-value
	+	-		+	-	
CTCF CTCFCohesin	10	15	0.42	2489	2896	**1.55 x 10^{-8}**
E4 Multi-tissue	5	3	0.73	1413	1930	**<2.2 x 10^{-16}**
E5 Bcelllike	2	1	1.00	445	599	**1.05 x 10^{-6}**
E12 Erythroblastlike	2	1	1.00	430	588	**4.11 x 10^{-7}**
P Promoter	2	0	0.50	246	455	**1.24 x 10^{-15}**
TF3 FOXA1 / AR / ESR1	1	0	1.00	312	378	**6.64 x 10^{-3}**
E6 Weak epithelial	-	-	-	97	107	0.53

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

TF5 AR	1	1	1.00	85	69	0.23
L5 Low signal	-	-	-	74	54	0.09
E7 Monocyte / Macrophage	-	-	-	65	62	0.86
E11 Tcell	-	-	-	44	39	0.66
E9 Liver / Intestine	0	1	1.00	15	44	**1.02×10^{-4}**
PC4 Polycomb / Bivalent stem cell Enh	-	-	-	29	36	0.46
E1 Stem cell	1	0	1.00	28	33	0.61
TF2 CEBPB	1	0	1.00	20	26	0.46
E2 Multi-tissue	-	-	-	19	23	0.64
TF4 OTX2	-	-	-	21	12	0.16
E10 Brain	-	-	-	4	10	0.18
E8 Weak multi-tissue	-	-	-	4	7	0.55
TN1 Transcription	-	-	-	5	4	1.00
TN3 Transcription	-	-	-	5	3	0.73
E3 Brain / Melanocyte	-	-	-	4	4	1.00
PC1 Polycomb / Heterochromatin	-	-	-	4	2	0.69
HET4 Heterochromatin	-	-	-	2	3	1.00
TN4 Transcription	-	-	-	2	3	1.00
PC2 Weak Polycomb	-	-	-	2	2	1.00
PC3 Polycomb	-	-	-	2	0	0.50
HET5 Centromere	-	-	-	0	1	1.00
HET1 Heterochromatin	-	-	-	1	0	1.00
TN2 Transcription	-	-	-	0	1	1.00

7.3.3 Candidate gene association

To assess the enrichment of genetic variations linked to complex traits within cell type-specific regulatory regions, Nott and collaborators conducted a linkage disequilibrium score (LDSC) regression analysis of heritability. In their LDSC approach, employing GWAS summary statistics from Grove *et al.*[37], they revealed that the SNP heritability of ASD was significantly enriched in regulatory elements specific to neurons, particularly in neuronal enhancers. In light of these results, our analysis will predominantly be centered on neuronal-specific regulatory regions.

7.3.3.1 *De novo*

From the total of 47 high regulatory impact *de novo* variants, 29 fell inside a putative promoter (61.70% of the instances), all of them corresponding to a different PARV gene (Figure 31A).

We further defined PLAC interactions between active promoters and *de novo* variants and identified 70 PLARV genes linked to these variants across the three cell types. Forty-two of the total number of PLARV genes were identified in neurons, of which 19 were not detected in any other cell type (Figure 31 B).

When comparing PLARV and PARV genes, we found that in most instances (63.3%) genes with promoters harboring ASD regulatory variants were PLAC-connected to a region affecting the same gene. In other words, the gene influenced by the regulatory variant is often the "closest gene" in proximity to that variant (Figure 31A).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

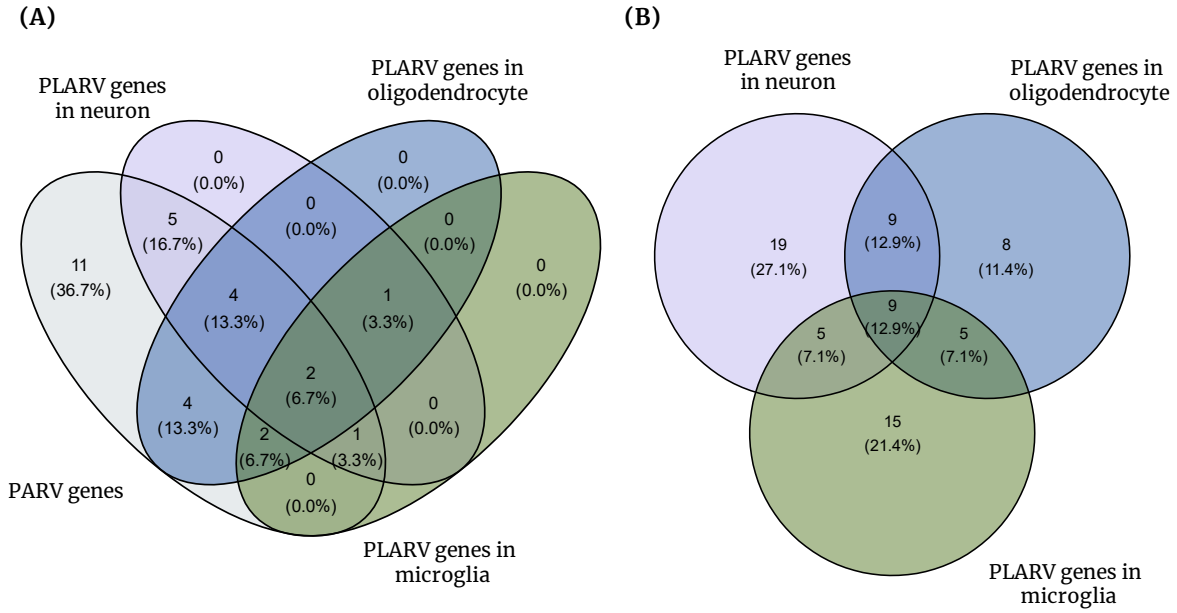


Figure 31. Venn diagram for *de novo* PARV and PLARV genes. (a) Variants falling inside a promoter (n = 29) are here represented in terms of the associated PARV and/or PLARV gene. Genes containing promoters harboring ASD regulatory variants are represented in the "PARV genes" category. Genes that are PLAC-linked to ASD regulatory variants are represented in the "PLARV gene" categories, one for each cell type analyzed here (neuron, oligodendrocytes, and microglia). Note that the purpose of this graph is to represent the overlap between PARV and PLARV genes, and thus, PLARV genes that do not overlap with PARV genes are here excluded. (b) PLARV genes ascribed to variants falling, or not, inside a promoter in each of the three different cell types.

However, we found that they did not overlap in 36.7% of the instances, which means that the "closest gene" is not compulsory the variant-regulated gene. An example is the *ETFBKMT* and *AMN1* loci, which have interactions to a shared promoter among the three cell types and is present nearby *KIA1551*'s TSS (Figure 32).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

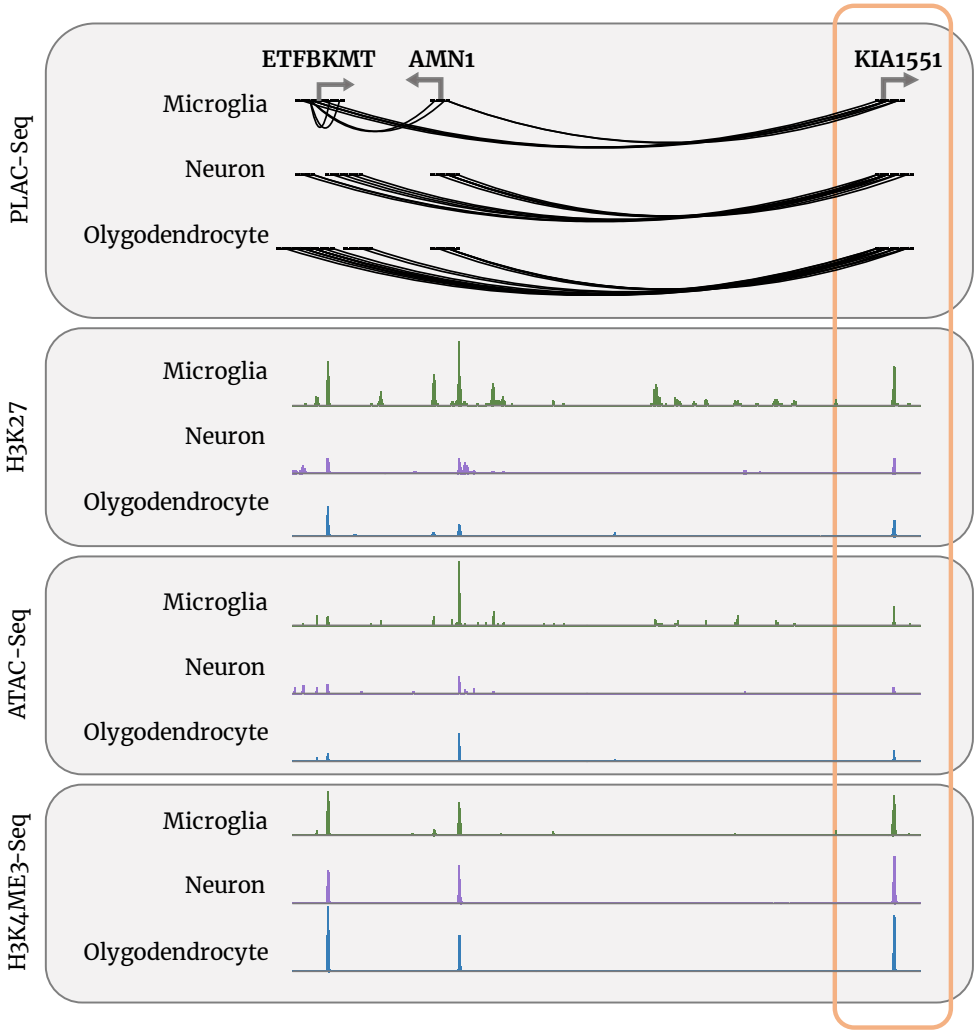


Figure 32. UCSC browser of ATAC-seq, H3K27ac and H3K4me3 ChIP-seq and PLAC-seq loops at the *ETFBKMT* and *AMN1* loci. A common promoter among the three cell types (highlighted in orange) is associated with *ETFBKMT* and *AMN1* expression although its genomic localization is next to *KIA1551*. This image is an edition of a screenshot of the UCSC browser track that includes PLAC-Seq data from Nott *et al.*[109]. UCSC browser session: https:// genome.ucsc.edu/s/nottalexi/glassLab_BrainCellTypes_hg19.

Moreover, out of the 18 variants falling outside promoters, 13 were PLAC-linked to a promoter, and thus assigned to the corresponding PLARV gene (72.22% of the instances) (Appendix 10.2, Supplementary Table 4).).

This yields a total of 42 variants (89.36%, out of the final list of high regulatory impact *de novo* variants) located within promoters and/or within regulatory regions PLAC-linked to promoters, and 37 variants with at least one PLARV associated gene (78.72%).

7.3.3.2 Inherited

From the total of 13,258 high regulatory impact ultra-rare inherited variants, 6,315 fell inside a putative promoter (47.63% of the instances), corresponding to 4,500 different PARV genes (Figure 33A; Appendix 10.2, Supplementary Table 6). In 71,18% of the instances, PARV genes were identified by a unique high regulatory impact variant inside their promoters.

However, some PARV genes harbored up to 8 different variants in their promoters (namely, *RNF152* and *MIDN*).

We further defined PLAC interactions between active promoters and ultra-rare inherited variants and identified 5,798 PLARV genes linked to these variants across the three cell types. Out of the total number of PLARV genes, 3,172 were identified in microglia, of which 1,389 were not detected in any other cell type (Figure 33 B). In regard to neurons, we identified more PLARV genes (n = 3,176) but only 1,191 were neuron specific.

When comparing PLARV and PARV genes, we found that in some instances (79.10%) genes with promoters harboring ASD regulatory variants were PLAC-connected to a region influencing the same gene. However, we found that they did not overlap in 20.90% of the instances (Figure 33A).

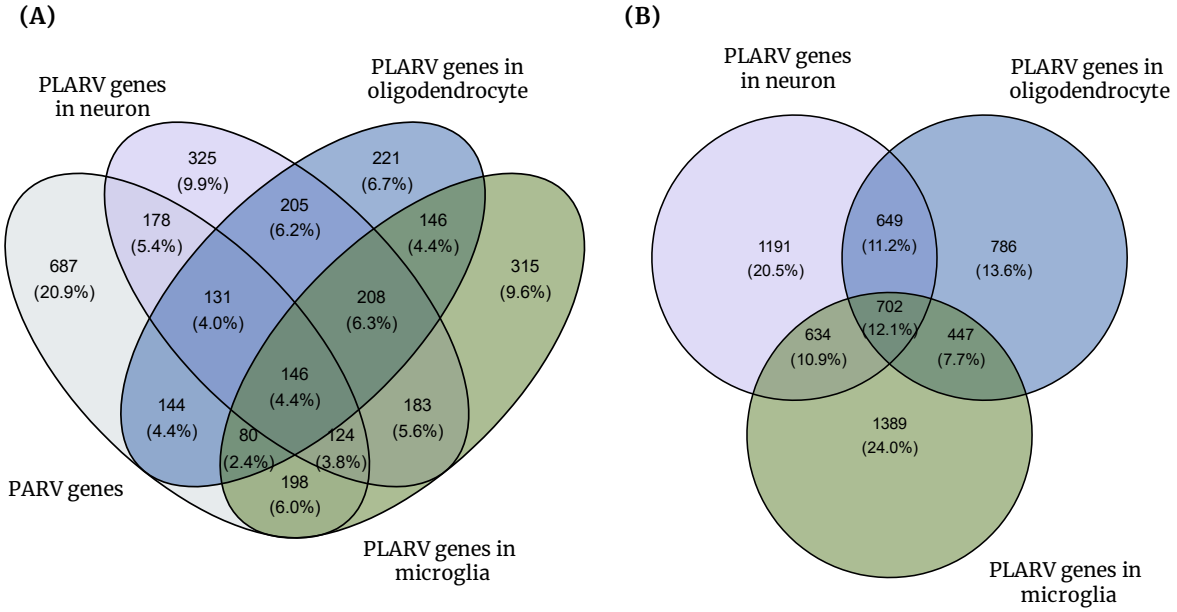


Figure 33. Venn diagram for inherited PARV and PLARV genes. (a) Variants falling inside a promoter (n = 6,315) are here represented in terms of the associated PARV and/or PLARV gene. (b) PLARV genes ascribed to variants falling, or not, inside a promoter in each of the three different cell types. For further explanation see caption in Figure 31.

Moreover, out of the 6,943 variants falling outside promoters, 5,510 were PLAC-linked to a promoter in at least one cell type, and thus assigned to the corresponding PLARV gene (*i.e.*, 79.36% of the variants without a PARV gene were associated with a PLARV gene) (Appendix 10.2, Supplementary Table 6).

This yields a total of 11,825 variants (89.19%, out of the final list of high regulatory impact inherited variants) located within promoters and/or within regulatory regions PLAC-linked to promoters, and 11,203 with at least one PLARV associated gene (84,50%).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

7.3.4 Gene function

When analyzing 69 PLARV genes associated with the final list of *de novo* high regulatory impact variants, we did not get any significant enrichment for any GO category. However, top BP (based on the gene ratio) included regulation of protein stability, dendrite development and chromatin remodeling, in line with previous evidence (see section 3.4 in the Introduction).

Nonetheless, when focusing on the final list of PLARV genes associated with the inherited variants, we did find significant enrichments in MF categories relevant for ASD such as ubiquitin protein ligase binding (q-value = 1.12×10^{-9}), DNA-binding transcription factor binding (q-value = 6.51×10^{-8}) and transcription coactivator activity (q-value = 2.21×10^{-5}), that remained significant even when independently assessed in each cell type. The same observation was noted for CC categories like nuclear envelope (q-value = 1.13×10^{-14}), ubiquitin ligase complex (q-value = 8.61×10^{-9}), neuron to neuron synapse (q-value = 3.12×10^{-8}) and spliceosomal complex (q-value = 3.19×10^{-6}) (data not shown).

Furthermore, we found significant enrichments in BP categories relevant for ASD such as RNA splicing (q-value = 4.33×10^{-10}), regulation of neuron projection development (q-value = 1.67×10^{-7}), axonogenesis (q-value = 1.67×10^{-7}), and histone modification (q-value = 1.52×10^{-6}). Top enriched terms remained significant even when analyzing the three cell types separately (Figure 34).

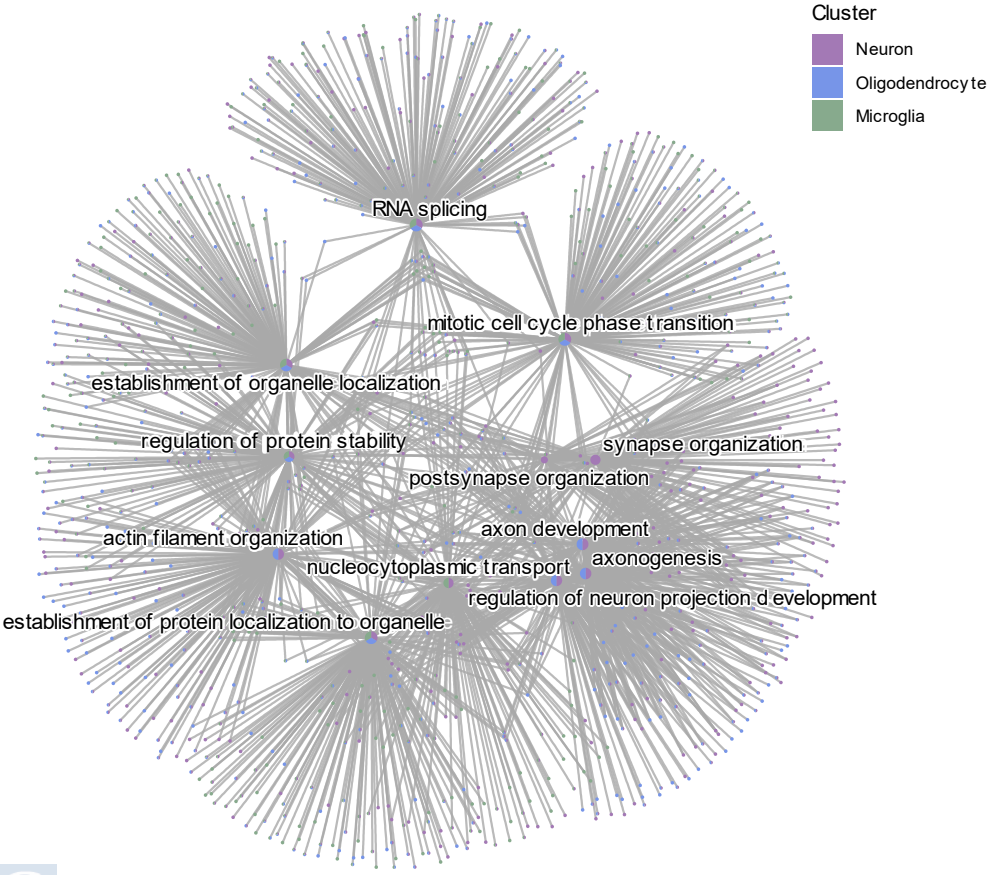


Figure 34. Biological processes significantly enriched within PLARV genes from the inherited variant list across the three cell types. The graphic depicts top 5 enriched terms for each gene cluster (*i.e.*, PLARV genes in

each cell type). Circles represent the number of genes, with larger circles including 160 genes and smaller ones including 71.

7.3.5 TF enrichment

Given the enrichment in categories related to TF binding in the GO analysis, and the fact that the most represented sequence class in our dataset pertains to CTCF binding, we aimed to characterize further enrichment in TF binding in our final list of *de novo* and inherited variants (Appendix 10.2, Supplementary Tables 7, 8). Of note, at the time this analysis was performed, ENCODE's version 2 browser could not be accessed, so we assessed TF enrichment with data from other sources (see Methods).

As anticipated, CTCF demonstrated a significant association, and thus, our previous findings were corroborated.

7.3.6 Variant prioritization and diagnostic yield

Although the primary aim of this chapter was not to provide a clinical diagnosis, limited by the incapability to robustly establish a genotype-phenotype relationship in the methodology discussed here, we sought to leverage the information from variants ascribed to PLARV genes.

To avoid a huge burden of interpretation, under emerging recommendations for clinical interpretation of the non-coding genome[346], we have focused on variants located in non-coding regions of the genome only if (i) they are within regulatory elements with functionally validated connections to target genes as proven by the Hi-C analysis here leveraged, and (ii) those genes are known to be associated with ASD risk, as compiled in our custom ASD risk gene list (Appendix 10.2, Supplementary Table 9).

By following this approach, we found: one patient with a *de novo* variant PLAC-linked to *PSMD1* specifically in neuron and one patient with a *de novo* variant harbored in *POGZ*'s promoter (Appendix 10.2, Supplementary Table 4). Both patients have a co-diagnosis of ID (data not shown).

Moreover, we found 489 inherited variants with interactions to 124 different PLARV genes included in our custom list of ASD-associated genes. These were present in 185 different patients, with an average of 2.64 ± 1.69 variants per child (Appendix 10.2, Supplementary Table 10).

Interestingly, we found one patient who harbored 10 different variants (which was the maximum observed number of variants per patient) and had one affected sibling (who has not been sequenced). All 10 variants were present in different TADs and corresponded to 10 different PLARV genes (most of them were detected specifically in neuron (Appendix 10.2, Supplementary Table 10)).

Additionally, one patient with a Tourette's syndrome diagnosis harbored 9 inherited variants across various genes. Moreover, one patient harbored 8 variants, while four other patients harbored 7 variants. Among these five patients, three were diagnosed with ADHD, and one had ID. All variants harbored in each patient were present in different TADs (data not shown).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

7.4 DISCUSSION

In this chapter, we have scrutinized targeted sequencing data from 200 ASD trios, where *de novo* LGD mutations and large CNVs, crucial in regulation of gene expression, had eluded detection in prior WES and microarray analyses. Moreover, we have applied the Sei framework, which is the most comprehensive chromatin-level sequence model to date. To our knowledge, this is the first time it has been applied to an ASD cohort, and we anticipate that further fruitful findings will result from larger cohorts.

Discussions about the limitations in deciphering the non-coding genome have been presented in earlier sections. Specifically, the absence of a non-coding counterpart to the triplet code, found in protein coding regions, has been a crucial obstacle for predicting the impact of mutations on gene function within the non-coding genome. Nevertheless, in this study, we have integrated data from ATAC-Seq and PLAC-Seq (which incorporates ChIP-Seq) to annotate variants based on four non-coding annotations: (i) regions of open chromatin, where DNA is exposed, thus allowing protein binding; (ii) regions of active chromatin, where epigenetic marks suggest transcription of a nearby gene; (iii) transcription factor binding sites, and (iv) prediction of the regulatory gene target using proximity to the variant and/or physical interactions genome-wide.

Importantly, in terms of general conclusions: we have identified that 28% of *de novo* variants and 25% of inherited variants exhibit a high regulatory potential in patients with negative results from WES and microarray analysis, as assessed by Sei.

By incorporating PLAC-Seq data, we were able to functionally assign a gene for ~80%/85% of *de novo* and inherited variants, respectively. Notably, when a variant was in close genomic proximity to a PARV gene, these associations were not corroborated by functional annotations in as much as 36.7% of the instances, in line with previous findings[110] (*i.e.*, PARV and PLARV genes did not overlap, as illustrated in Figures 31A, 32, and 33A).

Finally, while initiatives like ENCODE provide invaluable insights into defining regulatory elements of the genome, caution should be exercised when attempting to prioritize a specific type of regulatory element based on a priori hypotheses of heightened influence on gene regulation: many of the sequence classes found to be affected in ASD in this study were not significantly enriched in a specific type of cCRE signature (Appendix 10.2, Supplementary Table 5).

7.4.1 Exploring the sex bias in regulatory variation within ASD

In our cohort, the male-to-female ratio was 4.0, consistent with previous findings[4]. This ratio persisted even when prioritizing high-impact regulatory variants, with a ratio of 3.7 within *de novo* variants and 4.03 within inherited variants.

However, in evaluating transmission bias for ultra-rare inherited variants, we detected a statistically significant transmission bias favoring inheritance from the mother that remained significant even when the analysis was limited to autosomes (Appendix 10.2, Supplementary Table 6).

The debate surrounding the excess of germline DNMs arising on the maternal chromosome and the transmission bias favoring maternal inheritance has been a longstanding and essential

aspect of ASD etiology[61,72,86,96]. Theoretically, if females exhibit reduced vulnerability to ASD, as evidenced by the consistently observed female-to-male ratio of 4:1 in this and most ASD studies, and individuals with ASD experience reduced fecundity, basic genetic principles predict that mothers are more likely sources of risk alleles than fathers.

Initial observations relating to the observed sex bias indicated a trend of excess maternal *de novo* CNVs, although the pattern did not achieve statistical significance. Notably, females with ASD exhibited a higher number of *de novo* CNVs compared to males, and these genomic imbalances were both larger and impacted a greater number of genes[361]. Later, some of the first statistically significant genetic evidence of a transmission bias from mothers to their sons was observed for protein-coding SNVs in the study conducted by Krumm *et al.*[362].

On the contrary, one study identified a significant paternal-origin effect for CRE-SVs[96]. One possible explanation offered for the observation that mutations in regulatory regions follow an opposite pattern to what is predominantly seen in coding mutations, is that paternally inherited regulatory mutations, due to their less damaging impact, may require a greater number to manifest in disease. Crucially, this has given rise to an alternative hypothesis, suggesting that regulatory mutations may adhere to distinct principles compared to coding mutations.

Nevertheless, in line with our results, instances of the transmission of non-coding putative regulatory DNA from the mother to the male proband have been noted in genes such as *DSCAM* and *TRIO*, and both of them have been observed to be disrupted in ASD[232].

Taken together, our results represent, to the best of our knowledge, the first demonstration, in a non-biased global study of the non-coding genome, that parent-of-origin effects may follow similar rules as coding variation. However, these results should be cautiously interpreted and the comprehensive contribution of a gender bias in non-coding variation to ASD needs to be determined through much larger studies. Still, this underscores that influences of sex bias on genetic risk for ASD are more intricate than previously understood, and the allelic spectrum of variants varies between the maternal and paternal genomes.

7.4.2 Sequence class dysregulation

Probably, the most important observation in this study is the implication of a global dysregulation of CTCF.

CTCF is reported as a high confidence gene (score = 1S) in SFARI, associated with Tourette syndrome[363], ID[364], and, more recently, with ASD[58,61,72]. In fact, it was identified by Zhou *et al.*[73] as a gene reaching exome-wide significance in their analysis of rare *de novo* and inherited coding variants in 42,607 ASD cases, ($p < 2.5 \times 10^{-6}$) and thus included in our list of high risk ASD genes (Appendix 10.2, Supplementary Table 9).

Interestingly, in our random selection of cCREs, only 2.6% were classified as CTCF in ENCODE (Figure 27 B), and this proportion drastically decreased when we filtered for *de novo* and inherited mutations. Moreover, we observed a significant depletion of CTCF cCREs in the final list of inherited variants (Table 20).

However, when prioritizing variants with a high regulatory impact, not only did their proportion increase, but they became the regulatory class most represented amongst the detected mutations. Specifically, an increased number of mutations was associated with a

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

reduction (and not an increment) in the CTCF regulatory score, and this difference attained statistical significance for inherited variants ($p = 3.11 \times 10^{-8}$) (Table 21). Although *de novo* variants, constrained by a limited statistical power in terms of numbers, did not reach significance, they showed a similar trend.

Furthermore, these results were corroborated from the TF enrichment analysis (Appendix 10.2, Supplementary Tables 7, 8).

With this regard, it is important to note that the targeted sequencing approach followed in this chapter only covers 0.70% of the genome. Moreover, the selection of regulatory elements was performed in an arbitrary fashion, as it did not involve a priori hypotheses about which elements were most likely to influence gene expression. We solely restricted the regulatory elements to those active in tissues relevant to ASD. Thus, if we were to follow a WGS strategy, the results presented here suggest that we would likely obtain the same outcomes, supported with greater statistical power.

As further evidence, if we consider the top 10% of mutations with the highest class scores, they exclusively belong to the CTCF class, both for inherited and for *de novo* variants (Appendix 10.2, Supplementary Tables 4, 6).

Moreover, CTCF was one of the most frequently disrupted TF motifs in ASD cases *versus* controls in a single-cell analysis of gene expression and chromatin accessibility[237], and in this study. This suggested a potential mechanistic impact on the chromatin architecture: deletion of CTCF binding sites or lost on CTCF binding affinity can lead to formation of new TADs, which can cause genes to contact enhancers normally spatially too far away to do so (a process known as "enhancer hijacking")[237].

Of note, while writing this discussion, Nakamura *et al.*[365], published an analysis of promoter DNV from WGS data in a cohort of 5,044 ASD patients and 4,095 siblings. Interestingly, among the findings in their analyses, they specifically identified the enrichment of ASD-gene TAD promoters at CTCF-bound regions, further corroborating our results.

In addition, Nakamura *et al.*[365], found a specific association of DNV in TADs containing ASD risk genes and identified specific TADs with enrichment of promoter DNVs. Thus, we aimed to assess whether the variants in each of the patients with elevated numbers of high impact inherited variants were located within the same and/or adjacent TADs. If this were to be corroborated, we could hypothesize that the loss of CTCF binding affinity in TADs relevant to ASD might collectively contribute to disease risk by simultaneously affecting the expression of pertinent genes. On the contrary, we found that these variants were present in different TADs. Still, these findings prove interesting as they offer mechanistic insights into how non-coding regulatory variation may be influencing ASD risk.

Overall, it seems reasonable to assume that a substantial part of the non-coding genome in ASD is dysregulated in terms of its 3D-structure, and that further characterization of chromatin loop alterations (for example, by Hi-C analysis) will prove fruitful in further unraveling the missing heritability in ASD.

Yet, it has been reported that CTCF contributes not only to the formation of TAD boundaries but also to the stabilization of promoter-enhancer interactions. Thus, it may also be possible that TAD boundaries remain unaffected, but changes in promoter targeting by

enhancers (which is influenced by CTCF-binding) are responsible for the changes in gene expression leading to disease[366,367].

7.4.3 Biological, molecular, and cellular consequences of regulatory variants

The determination that non-coding variants predominantly target the same functional processes as protein-coding genes associated with ASD strongly supports the idea that non-coding variants may have causal roles in ASD development[61,201]. We thus sought to assess whether this holds true within our variant dataset. To do so, we conducted a burden analysis across all the gene sets listed in Gene Ontology. This analysis aimed to identify gene pathways enriched in genes affected by high-impact regulatory variants.

We found a significant enrichment of variants in pathways involved in "chromatin organization", "RNA processing translation" and "synaptic transmission" among others, which is largely consistent with previous findings (Figure 34). The association signal for the identified genes originates from rare inherited variation. While not achieving statistical significance for *de novo* variants, our results indicate that these variants impact BP in a manner consistent with those influenced by rare inherited variation. Additionally, both *de novo* and inherited genes contribute to a shared protein-protein interaction network with a total of 1,163 interconnected protein products, enriched for expression in the central nervous system ($p < 1 \times 10^{-6}$) (data not shown).

However, the key question remains as follows: whether variation found in the non-coding genome is targeting the same set of genes that have been implicated in ASD by WES and CNV studies or a unique group of genes missed by previous efforts. When contrasted to our custom list of high-risk ASD genes, we have found that the majority of high-impact regulatory variants, whether *de novo* or inherited, are PLAC-linked to genes not previously associated with ASD. Still, GO enrichments point to a scenario where coding and non-coding variation synergize in already characterized ASD relevant pathways. This aligns with recent observations indicating that rare inherited variation predominantly affects previously unknown risk genes impacting similar pathways to those influenced by known genes[73,74,76,89].

Overall, this convergence supports the idea of a causal contribution of non-coding regulatory mutations to ASD etiology.

7.4.4 Preliminary genotype-phenotype associations

Although it was not the primary aim of this study, we find it beneficial to highlight the following scenarios in regard to theoretical genotype-phenotype correlations. Nonetheless, we acknowledge limitations in our conclusions, as functional characterization of the evidence discussed in this section will be necessary to robustly establish evident interactions between non-coding variation and phenotype consequences.

Firstly, when analyzing high-impact inherited regulatory variations, and before prioritizing variants associated with PLARV genes already established to have a functional link to ASD, we identified two genes, *RNF152* and *MIDN*, each with eight different variants (Appendix 10.2, Supplementary Table 6). Moreover, all these variants were also PLAC-linked to the

USC DE SANTIAGO DE COMPOSTELA

corresponding gene, and we further characterize 3 variants outside *MIDN's* promoter but PLAC-linked to the gene.

RNF152 has been characterized to catalyze the ubiquitination of Rheb, a small GTPase enriched on the lysosomal membrane, where it activates mTORC signaling[368]. The ubiquitination of Rheb suppresses mTORC1 activation, which, in turn, has been found to be aberrantly activated in ASD patients[369]. In this regard, we hypothesize that the elevated number of regulatory variants, collectively affecting *RNF152* gene expression, leads to reduced levels of its protein product, thereby resulting in overactivation of the mTOR signaling pathway.

MIDN, on the other hand, has been linked to Parkinson's disease and regulates neurite outgrowth[370]. Moreover, it has been associated with female ASD in a WES[371]. Although most of our variants are found in male patients, these evidence points to a probable true association with ASD.

While these two genes are not currently listed in the SFARI database, we present additional evidence supporting their potential implications in ASD etiology.

Furthermore, several lines of evidence point to more severe phenotypes relating to higher burdens of genomic load. In particular, earlier studies have identified an inverse correlation between the IQ of individuals with ASD and the burden of specific regulatory variants[201].

In this light, it is interesting that our analysis uncovered a patient with 10 distinct high-impact regulatory variants with PLARV genes included in our list of ASD-risk genes (Appendix 10.2, Supplementary Table 10). This individual stands out as the sole case within our sample cohort who has a documented sibling with ASD. Considering that multiplex families often carry an elevated burden of inherited mutations and given the negative results for both exome and microarray analyses in this family, we propose that there is a slight possibility that the cumulative effects of the high-risk inherited variants described here are influencing disease risk and contributing to the observed recurrence in this family.

However, cautiousness should be taken when interpreting these results. First, it would be ideal to confirm, by sequencing methodologies, that these variants are also present in the affected sibling, and they are absent in a suitable control cohort. If so, functional assays should be performed in order to properly define their mode of action. Given that they are situated within different TADs, the likelihood of disrupted promoter-enhancer interactions within a specific genomic region, as observed elsewhere[365], and thus the possibility of them collectively functioning as a mutational hotspot, is ruled out.

Moreover, the second patient with the highest number of variants in PLARV genes which are known ASD-risk genes, harbors two different variants PLAC-linked to *FAM98C* (Appendix 10.2, Supplementary Table 10). *FAM98C* is classified with SFARI score 2. Initially, the identification of a PTV in the gene linked it to ASD[69]. Furthermore, a recent TADA analysis definitively identified *FAM98C* as an ASD candidate gene with a FDR < 0.1[75]. Our analysis provides further evidence to associate the gene with disease risk.

7.4.5 Limitations and future perspectives

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

The primary limitation in this study is the absence of a sibling/control cohort. Future perspectives involve retrieving additional data from a control cohort to address this. Ideally, access to larger cohorts of ASD patients would enhance statistical power, allowing for the expansion of current conclusions drawn from inherited variants to DNVs, which have been limited in number.

Moreover, the main conclusions gathered were made possible thanks to the utilization of PLAC-Seq data from key brain cell subpopulations. However, our initial selection of cCREs was based on activity in the brain, as well as in GI tissue. While we could not draw any conclusions regarding the latter, major goals will be to extend these approaches to all diseased tissues when data becomes accessible.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CONCLUDING REMARKS

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

8 CONCLUDING REMARKS

1. The results obtained in the colocalization study provide evidence of the involvement of *MAPT* at the brain level and *CRHR1* at a general level as causal genes underlying GWAS signals.

2. To date, *CRHR1* has not been reported as an ASD risk gene, and *MAPT* is categorized with only "suggestive evidence" of association with ASD in SFARI, representing the lowest confidence category. In this study, we present additional evidence supporting their implications in disease etiology. Additionally, the proximity of both genes within adjacent loci points to their implication in disease onset and suggests that future GWAS with larger sample sizes may also attain genome-wide significance for this region, establishing a robust association with ASD.

3. We have developed an *in silico* validation workflow for splice-altering variants applicable to WES data. While functional validation is essential for a robust determination of the biological impacts of these variants, data-driven validation methodologies, like the one presented here, could potentially be employed in the future for prioritizing variants in diagnosis and functional studies.

4. Genes harboring *in silico* validated splicing variants encompass both previously implicated autism-associated genes and newly identified ones. While our analysis has revealed a non-overlapping trend in terms of biological implications for splicing and coding variants, the biological pathways converge with those previously associated with ASD.

5. We have established a framework for assessing the functional annotation of regulatory elements in an unbiased manner. This framework incorporates filtering and quality control steps for the discovery of *de novo* and ultra-rare inherited variants (with well-established roles in ASD). Additionally, it involves the *in silico* assessment of their predicted functional impacts on gene regulation and the identification of candidate genes through the evaluation of PLAC-Seq data.

6. In patients where no apparent genetic cause has been identified, a noteworthy burden of predicted high-impact regulatory ultra-rare inherited variants has been observed.

7. The unbiased selection of candidate regulatory elements from ENCODE has proven successful in identifying regulatory consequences for non-coding variants, as no specific regulatory element is enriched for any particular regulatory function. This approach ensures more generalizable conclusions and warrants the possibility of further replication in larger cohorts with WGS data.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

8. We have detected a statistically significant transmission bias favoring inheritance from the mother that remained significant even when the analysis was limited to autosomes. This suggests that non-coding variation adheres to the same principles compared to coding mutations.

9. We have identified a global dysregulation of CTCF in ASD. The majority of inherited regulatory variants were linked to a decreased affinity for CTCF-binding. While *de novo* variants, constrained by limited statistical power regarding their numbers, did not reach significance, they exhibited a similar trend. This underscores a potential scenario in the ASD regulatory landscape, wherein a substantial portion of the non-coding genome in ASD exhibits dysregulation in terms of its 3D conformation and/or long range chromatin interactions.

10. Non-coding variation has predominantly been associated with previously unknown genes, although these genes converge in pathways already linked to disease. Additionally, our analysis indicates several potential phenotype-genotype correlations, providing further support for new ASD risk genes.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

BIBLIOGRAPHY

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

9 BIBLIOGRAPHY

1. Kanner L. Autistic disturbances of affective contact. Nerv Child. 1943;2:217-250.
2. American Psychiatric Association, American Psychiatric Association, eds. Diagnostic and Statistical Manual of Mental Disorders: DSM-5. 5th ed. American Psychiatric Association. 2013.
3. Elsabbagh M, Divan G, Koh YJ, et al. Global prevalence of autism and other pervasive developmental disorders. Autism Res Off J Int Soc Autism Res. 2012;5(3):160-179. doi:10.1002/aur.239
4. Zeidan J, Fombonne E, Scorah J, et al. Global prevalence of autism: A systematic review update. Autism Res. 2022;15(5):778-790. doi:10.1002/aur.2696
5. Lord C, Brugha TS, Charman T, et al. Autism spectrum disorder. Nat Rev Dis Primer. 2020;6(1):1-23. doi:10.1038/s41572-019-0138-4
6. Murray, CJL. The Global Burden of Disease Study at 30 years. Nat Med. 2022;28: 2019–2026. doi.org/10.1038/s41591-022-01990-1
7. Al-Beltagi M. Autism medical comorbidities. World J Clin Pediatr. 2021;10(3):15-28. doi:10.5409/wjcp.v10.i3.15
8. Barlattani T, D'Amelio C, Cavatassi A, et al. Autism spectrum disorders and psychiatric comorbidities: a narrative review. J Psychopathol. 2023;29:3-24.doi:10.36148/2284-0249-N281
9. Thurm A, Farmer C, Salzman E, Lord C, Bishop S. State of the Field: Differentiating Intellectual Disability From Autism Spectrum Disorder. Front Psychiatry. 2019;10:526. doi:10.3389/fpsyt.2019.00526
10. Bhat AN. Is Motor Impairment in Autism Spectrum Disorder Distinct From Developmental Coordination Disorder? A Report From the SPARK Study. Phys Ther. 2020;100(4):633-644. doi:10.1093/ptj/pzz190
11. Sullivan MO, Gallagher L, Heron EA. Gaining Insights into Aggressive Behaviour in Autism Spectrum Disorder Using Latent Profile Analysis. J Autism Dev Disord. 2019;49(10):4209-4218. doi:10.1007/s10803-019-04129-3
12. Sammels O, Karjalainen L, Dahlgren J, Wentz E. Autism Spectrum Disorder and Obesity in Children: A Systematic Review and Meta-Analysis. Obes Facts. 2022;15(3):305-320. doi:10.1159/000523943
13. Micai M, Fatta LM, Gila L, et al. Prevalence of co-occurring conditions in children and adults with autism spectrum disorder: A systematic review and meta-analysis. Neurosci Biobehav Rev. 2023;155:105436. doi:10.1016/j.neubiorev.2023.105436
14. Percy AK. Rett Syndrome: Exploring the Autism Link. Arch Neurol. 2011;68(8):985-989. doi:10.1001/archneurol.2011.149

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

15. Emberti Gialloreti L, Mazzone L, Benvenuto A, et al. Risk and Protective Environmental Factors Associated with Autism Spectrum Disorder: Evidence-Based Principles and Recommendations. J Clin Med. 2019;8(2):217. doi:10.3390/jcm8020217
16. Pugsley K, Scherer SW, Bellgrove MA, Hawi Z. Environmental exposures associated with elevated risk for autism spectrum disorder may augment the burden of deleterious de novo mutations among probands. Mol Psychiatry. 2022;27(1):710-730. doi:10.1038/s41380-021-01142-w
17. Ronald A, Hoekstra R. Progress in Understanding the Causes of Autism Spectrum Disorders and Autistic Traits: Twin Studies from 1977 to the Present Day. Am J Med Genet B Neuropsychiatr Genet. 2011;156B(3):255-274. doi:10.1002/ajmg.b
18. Mayhew AJ, Meyre D. Assessing the Heritability of Complex Traits in Humans: Methodological Challenges and Opportunities. Curr Genomics. 2017;18(4):332-340. doi:10.2174/1389202918666170307161450
19. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era — concepts and misconceptions. Nat Rev Genet. 2008;9(4):255-266. doi:10.1038/nrg2322
20. Folstein S, Rutter M. Infantile autism: a genetic study of 21 twin pairs. J Child Psychol Psychiatry. 1977;18(4):297-321. doi:10.1111/j.1469-7610.1977.tb00443.x
21. Ronald A, Hoekstra RA. Autism spectrum disorders and autistic traits: A decade of new twin studies. Am J Med Genet B Neuropsychiatr Genet. 2011;156(3):255-274. doi:10.1002/ajmg.b.31159
22. Tick B, Bolton P, Happé F, Rutter M, Rijsdijk F. Heritability of autism spectrum disorders: a meta-analysis of twin studies. J Child Psychol Psychiatry. 2016;57(5):585-595. doi:10.1111/jcpp.12499
23. Sandin S, Lichtenstein P, Kuja-Halkola R, Larsson H, Hultman CM, Reichenberg A. The familial risk of autism. JAMA. 2014;311(17):1770-1777. doi:10.1001/jama.2014.4144
24. Sandin S, Lichtenstein P, Kuja-Halkola R, Hultman C, Larsson H, Reichenberg A. The Heritability of Autism Spectrum Disorder. JAMA. 2017;318(12):1182. doi:10.1001/jama.2017.12141
25. Bai D, Yip BHK, Windham GC, et al. Association of Genetic and Environmental Factors With Autism in a 5-Country Cohort. JAMA Psychiatry. 2019;76(10):1035-1043. doi:10.1001/jamapsychiatry.2019.1411
26. Emberti Gialloreti L, Curatolo P. Autism Spectrum Disorder: Why Do We Know So Little?. Front Neurol. 2018;9:670. doi:10.3389/fneur.2018.00670
27. Havdahl A, Niarchou M, Starnawska A, Uddin M, van der Merwe C, Warrier V. Genetic contributions to autism spectrum disorder. Psychol Med. 2021;51(13):2260-2273. doi:10.1017/S0033291721000192
28. Andrews SV, Sheppard B, Windham GC, et al. Case-control meta-analysis of blood DNA methylation and autism spectrum disorder. Mol Autism. 2018;9:40. doi:10.1186/s13229-018-0224-6
29. Wong CCY, Smith RG, Hannon E, et al. Genome-wide DNA methylation profiling identifies convergent molecular signatures associated with idiopathic and syndromic autism in

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

post-mortem human brain tissue. Hum Mol Genet. 2019;28(13):2201-2211. doi:10.1093/hmg/ddz052
30. Joubert BR, den Dekker HT, Felix JF, et al. Maternal plasma folate impacts differential DNA methylation in an epigenome-wide meta-analysis of newborns. Nat Commun. 2016;7:10577. doi:10.1038/ncomms10577
31. De Rubeis S, Buxbaum JD. Genetics and genomics of autism spectrum disorder: embracing complexity. Hum Mol Genet. 2015;24(R1):R24-R31. doi:10.1093/hmg/ddv273
32. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-753. doi:10.1038/nature08494
33. Goswami C, Chattopadhyay A, Chuang EY. Rare variants: data types and analysis strategies. Ann Transl Med. 2021;9(12). doi:10.21037/atm-21-1635
34. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019;20(1):246. doi:10.1186/s13059-019-1828-7
35. Piovesan A, Pelleri MC, Antonaros F, Strippoli P, Caracausi M, Vitale L. On the length, weight and GC content of the human genome. BMC Res Notes. 2019;12:106. doi:10.1186/s13104-019-4137-z
36. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019;176(3):535-548.e24. doi:10.1016/j.cell.2018.12.015
37. Grove J, Ripke S, Als TD, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019;51(3):431-444. doi:10.1038/s41588-019-0344-8
38. Gaugler T, Klei L, Sanders SJ, et al. Most genetic risk for autism resides with common variation. Nat Genet. 2014;46(8):881-885. doi:10.1038/ng.3039
39. Huguet G, Benabou M, Bourgeron T. The Genetics of Autism Spectrum Disorders. Trends Cogn Sci. 2016;101-129. doi:10.1007/978-3-319-27069-2_11
40. Satterstrom FK, Kosmicki JA, Wang J, et al. Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell. 2020;180(3):568-584.e23. doi:10.1016/j.cell.2019.12.036
41. An JY, Lin K, Zhu L, et al. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. Science. 2018;362(6420):eaat6576. doi:10.1126/science.aat6576
42. Doan RN, Lim ET, De Rubeis S, et al. Recessive gene disruptions in autism spectrum disorder. Nat Genet. 2019;51(7):1092-1098. doi:10.1038/s41588-019-0433-8
43. Dawn Teare M, Barrett JH. Genetic linkage studies. The Lancet. 2005;366(9490):1036-1044. doi:10.1016/S0140-6736(05)67382-5
44. International Molecular Genetic Study of Autism Consortium. A full genome screen for autism with evidence for linkage to a region on chromosome 7q. Hum Mol Genet. 1998;7(3):571-578. doi:10.1093/hmg/7.3.571
45. Weiss LA, Arking DE, Gene Discovery Project of Johns Hopkins & the Autism Consortium, Daly MJ, Chakravarti A. A genome-wide linkage and association scan reveals novel loci for autism. Nature. 2009;461(7265):802-808. doi:10.1038/nature08490

46. Alarcón M, Abrahams BS, Stone JL, et al. Linkage, Association, and Gene-Expression Analyses Identify CNTNAP2 as an Autism-Susceptibility Gene. Am J Hum Genet. 2008;82(1):150-159. doi:10.1016/j.ajhg.2007.09.005
47. Abrahams BS, Geschwind DH. Advances in autism genetics: on the threshold of a new neurobiology. Nat Rev Genet. 2008;9(5):341-355. doi:10.1038/nrg2346
48. Powder KE. Quantitative Trait Loci (QTL) Mapping. Methods Mol Biol Clifton NJ. 2020;2082:211-229. doi:10.1007/978-1-0716-0026-9_15
49. Sanders SJ, Neale BM, Huang H, et al. Whole genome sequencing in psychiatric disorders: the WGSPD consortium. Nat Neurosci. 2017;20(12):1661-1668. doi:10.1038/s41593-017-0017-9
50. SPARK. Accessed January 11, 2024. https://sparkforautism.org/
51. Geschwind DH, Sowinski J, Lord C, et al. The Autism Genetic Resource Exchange: A Resource for the Study of Autism and Related Neuropsychiatric Conditions. Am J Hum Genet. 2001;69(2):463-466.
52. Lajonchere CM. Changing the Landscape of Autism Research: The Autism Genetic Resource Exchange. Neuron. 2010;68(2):187-191. doi:10.1016/j.neuron.2010.10.009
53. MSSNG. Accessed January 11, 2024. https://research.mss.ng/
54. The Autism Genome Project Consortium (AGP). Nature Index. Published December 26, 2023. Accessed January 11, 2024. https://www.nature.com/nature-index/institution-outputs/canada/the-autism-genome-project-consortium-agp/537af26d140ba07f3d000006
55. Fischbach GD, Lord C. The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors. Neuron. 2010;68(2):192-195. doi:10.1016/j.neuron.2010.10.006
56. Buxbaum JD, Bolshakova N, Brownfeld JM, et al. The Autism Simplex Collection: an international, expertly phenotyped autism sample for genetic and phenotypic analyses. Mol Autism. 2014;5(1):34. doi:10.1186/2040-2392-5-34
57. O'Roak BJ, Deriziotis P, Lee C, et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet. 2011;43(6):585-589. doi:10.1038/ng.835
58. Iossifov I, O'Roak BJ, Sanders SJ, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014;515(7526):216-221. doi:10.1038/nature13908
59. Sanders SJ, Murtha MT, Gupta AR, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature. 2012;485(7397):237-241. doi:10.1038/nature10945
60. Neale BM, Kou Y, Liu L, et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature. 2012;485(7397):242-245. doi:10.1038/nature11011
61. Yuen RKC, Merico D, Cao H, et al. Genome-wide characteristics of de novo mutations in autism. NPJ Genomic Med. 2016;1:160271-1602710. doi:10.1038/npjgenmed.2016.27
62. Kong A, Frigge ML, Masson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012;488(7412):471-475. doi:10.1038/nature11396

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

63. Nguyen HT, Bryois J, Kim A, et al. Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. Genome Med. 2017;9:114. doi:10.1186/s13073-017-0497-y
64. He X, Sanders SJ, Liu L, et al. Integrated Model of De Novo and Inherited Genetic Variants Yields Greater Power to Identify Risk Genes. PLoS Genet. 2013;9(8):e1003671. doi:10.1371/journal.pgen.1003671
65. Weiner DJ, Wigdor EM, Ripke S, et al. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. Nat Genet. 2017;49(7):978-985. doi:10.1038/ng.3863
66. Trost B, Engchuan W, Nguyen CM, et al. Genome-wide detection of tandem DNA repeats that are expanded in autism. Nature. 2020;586(7827):80-86. doi:10.1038/s41586-020-2579-z
67. O'Roak BJ, Stessman HA, Boyle EA, et al. Recurrent de novo mutations implicate novel genes underlying simplex autism risk. Nat Commun. 2014;5:5595. doi:10.1038/ncomms6595
68. Iossifov I, Ronemus M, Levy D, et al. De Novo Gene Disruptions in Children on the Autistic Spectrum. Neuron. 2012;74(2):285-299. doi:10.1016/j.neuron.2012.04.009
69. De Rubeis S, He X, Goldberg AP, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014;515(7526):209-215. doi:10.1038/nature13772
70. Sanders SJ, Murtha MT, Gupta AR, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature. 2012;485(7397):237-241. doi:10.1038/nature10945
71. Sanders SJ, He X, Willsey AJ, et al. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron. 2015;87(6):1215-1233. doi:10.1016/j.neuron.2015.09.016
72. Krumm N, Turner TN, Baker C, et al. Excess of rare, inherited truncating mutations in autism. Nat Genet. 2015;47(6):582-588. doi:10.1038/ng.3303
73. Zhou X, Feliciano P, Shu C, et al. Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. Nat Genet. 2022;54(9):1305-1319. doi:10.1038/s41588-022-01148-2
74. Fu JM, Satterstrom FK, Peng M, et al. Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. Nat Genet. 2022;54(9):1320-1331. doi:10.1038/s41588-022-01104-0
75. Ruzzo EK, Pérez-Cano L, Jung JY, et al. Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks. Cell. 2019;178(4):850-866.e26. doi:10.1016/j.cell.2019.07.015
76. Wilfert AB, Turner TN, Murali SC, et al. Recent ultra-rare inherited variants implicate new autism candidate risk genes. Nat Genet. 2021;53(8):1125-1134. doi:10.1038/s41588-021-00899-8
77. Samocha KE, Robinson EB, Sanders SJ, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014;46(9):944-950. doi:10.1038/ng.3050

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

78. Samocha KE, Kosmicki JA, Karczewski KJ, et al. Regional missense constraint improves variant deleteriousness prediction. Published online June 12, 2017;148353. doi:10.1101/148353
79. Grantham R. Amino acid difference formula to help explain protein evolution. Science. 1974;185(4154):862-864. doi:10.1126/science.185.4154.862
80. Huber CD, Kim BY, Lohmueller KE. Population genetic models of GERP scores suggest pervasive turnover of constrained sites across mammalian evolution. PLoS Genet. 2020;16(5):e1008827. doi:10.1371/journal.pgen.1008827
81. Richards S, Aziz N, Bale S, et al. Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med Off J Am Coll Med Genet. 2015;17(5):405-424. doi:10.1038/gim.2015.30
82. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-443. doi:10.1038/s41586-020-2308-7
83. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLOS Genet. 2013;9(8):e1003709. doi:10.1371/journal.pgen.1003709
84. Yuen RK, Merico D, Bookman M, et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat Neurosci. 2017;20(4):602-611. doi:10.1038/nn.4524
85. Jiang Y hui, Yuen RKC, Jin X, et al. Detection of Clinically Relevant Genetic Variants in Autism Spectrum Disorder by Whole-Genome Sequencing. Am J Hum Genet. 2013;93(2):249-263. doi:10.1016/j.ajhg.2013.06.012
86. Dong S, Walker MF, Carriero NJ, et al. De novo insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder. Cell Rep. 2014;9(1):16-23. doi:10.1016/j.celrep.2014.08.068
87. Feliciano P, Zhou X, Astrovskaya I, et al. Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. NPJ Genomic Med. 2019;4:19. doi:10.1038/s41525-019-0093-8
88. Kasherman MA, Premarathne S, Burne THJ, Wood SA, Piper M. The Ubiquitin System: a Regulatory Hub for Intellectual Disability and Autism Spectrum Disorder. Mol Neurobiol. 2020;57(5):2179-2193. doi:10.1007/s12035-020-01881-x
89. Cirnigliaro M, Chang TS, Arteaga SA, et al. The contributions of rare inherited and polygenic risk to ASD in multiplex families. Proc Natl Acad Sci. 2023;120(31):e2215632120. doi:10.1073/pnas.2215632120
90. Weiss MM, Hermsen MA, Meijer GA, et al. Comparative genomic hybridisation. Mol Pathol. 1999;52(5):243-251. doi: 10.1136/mp.52.5.243
91. Sebat J, Lakshmi B, Malhotra D, et al. Strong Association of De Novo Copy Number Mutations with Autism. Science. 2007;316(5823):445-449. doi:10.1126/science.1138659

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

92. Sanders SJ, Ercan-Sencicek AG, Hus V, et al. Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism. Neuron. 2011;70(5):863-885. doi:10.1016/j.neuron.2011.05.002
93. Marshall CR, Noor A, Vincent JB, et al. Structural Variation of Chromosomes in Autism Spectrum Disorder. Am J Hum Genet. 2008;82(2):477-488. doi:10.1016/j.ajhg.2007.12.009
94. Pinto D, Pagnamenta AT, Klei L, et al. Functional Impact of Global Rare Copy Number Variation in Autism Spectrum Disorder. Nature. 2010;466(7304):368-372. doi:10.1038/nature09146
95. Werling DM, Brand H, An JY, et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. Nat Genet. 2018;50(5):727-736. doi:10.1038/s41588-018-0107-y
96. Brandler WM, Antaki D, Gujral M, et al. Paternally inherited cis-regulatory structural variants are associated with autism. Science. 2018;360(6386):327-331. doi:10.1126/science.aan2261
97. Levy D, Ronemus M, Yamrom B, et al. Rare De Novo and Transmitted Copy-Number Variation in Autistic Spectrum Disorders. Neuron. 2011;70(5):886-897. doi:10.1016/j.neuron.2011.05.015
98. Leppa VM, Kravitz SN, Martin CL, et al. Rare Inherited and De Novo CNVs Reveal Complex Contributions to ASD Risk in Multiplex Families. Am J Hum Genet. 2016;99(3):540-554. doi:10.1016/j.ajhg.2016.06.036
99. Zhang Y, Liu X, Guo R, et al. Biological implications of genetic variations in autism spectrum disorders from genomics studies. Biosci Rep. 2021;41(7):BSR20210593. doi:10.1042/BSR20210593
100. Uffelmann E, Huang QQ, Munung NS, et al. Genome-wide association studies. Nat Rev Methods Primer. 2021;1(1):1-21. doi:10.1038/s43586-021-00056-9
101. Altshuler D, Donnelly P, The International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437(7063):1299-1320. doi:10.1038/nature04226
102. Ma D, Salyakina D, Jaworski JM, et al. A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. Ann Hum Genet. 2009;73(Pt 3):263-273. doi:10.1111/j.1469-1809.2009.00523.x
103. Devlin B, Melhem N, Roeder K. Do common variants play a role in risk for autism? Evidence and theoretical musings. Brain Res. 2011;1380:78-84. doi:10.1016/j.brainres.2010.11.026
104. The Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. Mol Autism. 2017;8(1):21. doi:10.1186/s13229-017-0137-9
105. Anney R, Klei L, Pinto D, et al. Individual common variants exert weak effects on the risk for autism spectrum disorders. Hum Mol Genet. 2012;21(21):4781-4792. doi:10.1093/hmg/dds301
106. Wang K, Zhang H, Ma D, et al. Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature. 2009;459(7246):528-533. doi:10.1038/nature07999

107. Barton NH, Etheridge AM, Véber A. The infinitesimal model: Definition, derivation, and implications. Theor Popul Biol. 2017;118:50-73. doi:10.1016/j.tpb.2017.06.001
108. Geschwind DH, State MW. Gene hunting in autism spectrum disorder: on the path to precision medicine. Lancet Neurol. 2015;14(11):1109-1120. doi:10.1016/S1474-4422(15)00044-7
109. Nott A, Holtman IR, Coufal NG, et al. Brain cell type-specific enhancer-promoter interactome maps and disease risk association. Science. 2019;366(6469):1134-1139. doi:10.1126/science.aay0793
110. Sanders SJ, Neale BM, Huang H, et al. Whole Genome Sequencing in Psychiatric Disorders: the WGSPD Consortium. Nat Neurosci. 2017;20(12):1661-1668. doi:10.1038/s41593-017-0017-9
111. Leeuw CA de, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLOS Comput Biol. 2015;11(4):e1004219. doi:10.1371/journal.pcbi.1004219
112. Alonso-Gonzalez A, Calaza M, Rodriguez-Fontenla C, Carracedo A. Novel Gene-Based Analysis of ASD GWAS: Insight Into the Biological Role of Associated Genes. Front Genet. 2019;10:733. doi:10.3389/fgene.2019.00733
113. Mai J, Lu M, Gao Q, Zeng J, Xiao J. Transcriptome-wide association studies: recent advances in methods, applications and available databases. Commun Biol. 2023;6(1):1-10. doi:10.1038/s42003-023-05279-y
114. Hu Y, Li M, Lu Q, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. Nat Genet. 2019;51(3):568-576. doi:10.1038/s41588-019-0345-7
115. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369(6509):1318-1330. doi:10.1126/science.aaz1776
116. Rodriguez-Fontenla C, Carracedo A. UTMOST, a single and cross-tissue TWAS (Transcriptome Wide Association Study), reveals new ASD (Autism Spectrum Disorder) associated genes. Transl Psychiatry. 2021;11(1):1-11. doi:10.1038/s41398-021-01378-8
117. van Calker D, Serchov T. The “missing heritability”—Problem in psychiatry: Is the interaction of genetics, epigenetics and transposable elements a potential solution? Neurosci Biobehav Rev. 2021;126:23-42. doi:10.1016/j.neubiorev.2021.03.019
118. Gandal MJ, Leppa V, Won H, Parikshak NN, Geschwind DH. The road to precision psychiatry: translating genetics into disease mechanisms. Nat Neurosci. 2016;19(11):1397-1407. doi:10.1038/nn.4409
119. O’Roak BJ, Vives L, Fu W, et al. Multiplex Targeted Sequencing Identifies Recurrently Mutated Genes in Autism Spectrum Disorders. Science. 2012;338(6114):1619-1622. doi:10.1126/science.1227764
120. Zhang SF, Gao J, Liu CM. The Role of Non-Coding RNAs in Neurodevelopmental Disorders. Front Genet. 2019;10:1033. doi:10.3389/fgene.2019.01033
121. D’haene E, Vergult S. Interpreting the impact of noncoding structural variation in neurodevelopmental disorders. Genet Med. 2021;23(1):34-46. doi:10.1038/s41436-020-00974-1

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

122. Trost B, Thiruvahindrapuram B, Chan AJS, et al. Genomic architecture of autism from comprehensive whole-genome sequence annotation. Cell. 2022;185(23):4409-4427.e18. doi:10.1016/j.cell.2022.10.009
123. Tena JJ, Santos-Pereira JM. Topologically Associating Domains and Regulatory Landscapes in Development, Evolution and Disease. Front Cell Dev Biol. 2021;9:702787. doi:10.3389/fcell.2021.702787
124. Hicks SD, Middleton FA. A Comparative Review of microRNA Expression Patterns in Autism Spectrum Disorder. Front Psychiatry. 2016;7. doi:10.3389/fpsyt.2016.00176
125. Barrett LW, Fletcher S, Wilton SD. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. Cell Mol Life Sci. 2012;69(21):3613-3634. doi:10.1007/s00018-012-0990-9
126. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet. 2018;19(5):286-298. doi:10.1038/nrg.2017.115
127. Steri M, Idda ML, Whalen MB, Orrù V. Genetic variants in mRNA untranslated regions. WIREs RNA. 2018;9(4). doi:10.1002/wrna.1474
128. Mayr C. Regulation by 3′-Untranslated Regions. Annu Rev Genet. 2017;51(1):171-194. doi:10.1146/annurev-genet-120116-024704
129. LoParo D, Waldman ID. The oxytocin receptor gene (OXTR) is associated with autism spectrum disorder: a meta-analysis. Mol Psychiatry. 2015;20(5):640-646. doi:10.1038/mp.2014.77
130. Harrison AJ, Gamsiz ED, Berkowitz IC, Nagpal S, Jerskey BA. Genetic variation in the oxytocin receptor gene is associated with a social phenotype in autism spectrum disorders. Am J Med Genet B Neuropsychiatr Genet. 2015;168(8):720-729. doi:10.1002/ajmg.b.32377
131. Campbell DB, Datta D, Jones ST, et al. Association of oxytocin receptor (OXTR) gene variants with multiple phenotype domains of autism spectrum disorder. J Neurodev Disord. 2011;3(2):101-112. doi:10.1007/s11689-010-9071-2
132. Göpferich M, George NO, Muelas AD, et al. Single Cell 3'UTR Analysis Identifies Changes in Alternative Polyadenylation throughout Neuronal Differentiation and in Autism. Neuroscience. 2020; doi:10.1101/2020.08.12.247627
133. Turner TN, Coe BP, Dickel DE, et al. Genomic Patterns of De Novo Mutation in Simplex Autism. Cell. 2017;171(3):710-722.e12. doi:10.1016/j.cell.2017.08.047
134. Nagarajan R, Hogart A, Gwye Y, Martin MR, LaSalle JM. Reduced MeCP2 Expression is Frequent in Autism Frontal Cortex and Correlates with Aberrant MECP2 Promoter Methylation. Epigenetics. 2006;1(4):172-182. doi:10.4161/epi.1.4.3514
135. Tansey KE, Hill MJ, Cochrane LE, Gill M, Anney RJ, Gallagher L. Functionality of promoter microsatellites of arginine vasopressin receptor 1A (AVPR1A): implications for autism. Mol Autism. 2011;2(1):3. doi:10.1186/2040-2392-2-3
136. Padhi EM, Hayeck TJ, Mannion B, et al. De Novo Mutation in an Enhancer of EBF3 in Simplex Autism. Genomics. 2020; doi:10.1101/2020.08.28.270751
137. Zhao YT, Kwon DY, Johnson BS, et al. Long genes linked to autism spectrum disorders harbor broad enhancer-like chromatin domains. Genome Res. 2018;28(7):933-942. doi:10.1101/gr.233775.117

138. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol. 2018;19(10):621-637. doi:10.1038/s41580-018-0028-8
139. Mathelier A, Shi W, Wasserman WW. Identification of altered cis-regulatory elements in human disease. Trends Genet. 2015;31(2):67-76. doi:10.1016/j.tig.2014.12.003
140. Perenthaler E, Yousefi S, Niggl E, Barakat TS. Beyond the Exome: The Non-coding Genome and Enhancers in Neurodevelopmental Disorders and Malformations of Cortical Development. Front Cell Neurosci. 2019;13:352. doi:10.3389/fncel.2019.00352
141. Chen X, Zhou J, Zhang R, et al. Tissue-specific enhancer functional networks for associating distal regulatory regions to disease. Cell Syst. 2021;12(4):353-362.e6. doi:10.1016/j.cels.2021.02.002
142. Kim IB, Lee T, Lee J, et al. Non-coding de novo mutations in chromatin interactions are implicated in autism spectrum disorder. Mol Psychiatry. 2022;27(11):4680-4694. doi:10.1038/s41380-022-01697-2
143. Lekka E, Hall J. Noncoding RNAs in disease. FEBS Lett. 2018;592(17):2884-2900. doi:10.1002/1873-3468.13182
144. Hicks SD, Ignacio C, Gentile K, Middleton FA. Salivary miRNA profiles identify children with autism spectrum disorder, correlate with adaptive behavior, and implicate ASD candidate genes involved in neurodevelopment. BMC Pediatr. 2016;16(1):52. doi:10.1186/s12887-016-0586-x
145. Sarachana T, Zhou R, Chen G, Manji HK, Hu VW. Investigation of post-transcriptional gene regulatory networks associated with autism spectrum disorders by microRNA expression profiling of lymphoblastoid cell lines. Genome Med. 2010;2(4):23. doi:10.1186/gm144
146. Luo T, Liu P, Wang X, et al. Effect of the autism-associated lncRNA Shank2-AS on architecture and growth of neurons. J Cell Biochem. 2019;120(2):1754-1762. doi:10.1002/jcb.27471
147. Graf J, Kretz M. From structure to function: Route to understanding lncRNA mechanism. BioEssays. 2020;42(12):2000027. doi:10.1002/bies.202000027
148. Willingham AT, Orth AP, Batalov S, et al. A Strategy for Probing the Function of Noncoding RNAs Finds a Repressor of NFAT. Science. 2005;309(5740):1570-1573. doi:10.1126/science.1115901
149. Cesana M, Cacchiarelli D, Legnini I, et al. A Long Noncoding RNA Controls Muscle Differentiation by Functioning as a Competing Endogenous RNA. Cell. 2011;147(2):358-369. doi:10.1016/j.cell.2011.09.028
150. Chen G, Wang Z, Wang D, et al. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res. 2013;41(Database issue):D983-D986. doi:10.1093/nar/gks1099
151. Ziats MN, Rennert OM. Aberrant Expression of Long Noncoding RNAs in Autistic Brain. J Mol Neurosci. 2013;49(3):589-593. doi:10.1007/s12031-012-9880-8
152. Wang Y, Zhao X, Ju W, et al. Genome-wide differential expression of synaptic long noncoding RNAs in autism spectrum disorder. Transl Psychiatry. 2015;5(10):e660-e660. doi:10.1038/tp.2015.144

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

153. Cogill SB, Srivastava AK, Yang MQ, Wang L. Co-expression of long non-coding RNAs and autism risk genes in the developing human brain. BMC Syst Biol. 2018;12(S7):91. doi:10.1186/s12918-018-0639-x
154. Constantin L. The Role of MicroRNAs in Cerebellar Development and Autism Spectrum Disorder During Embryogenesis. Mol Neurobiol. 2017;54(9):6944-6959. doi:10.1007/s12035-016-0220-9
155. Wu YE, Parikshak NN, Belgard TG, Geschwind DH. Genome-wide, integrative analysis implicates microRNA dysregulation in autism spectrum disorder. Nat Neurosci. 2016;19(11):1463-1476. doi:10.1038/nn.4373
156. Cao X, Yeo G, Muotri AR, Kuwabara T, Gage FH. Noncoding RNAs in the mammalian central nervous system. Annu Rev Neurosci. 2006;29:77-103. doi:10.1146/annurev.neuro.29.051605.112839
157. Cao DD, Li L, Chan WY. MicroRNAs: Key Regulators in the Central Nervous System and Their Implication in Neurological Diseases. Int J Mol Sci. 2016;17(6):842. doi:10.3390/ijms17060842
158. Chen W, Qin C. General hallmarks of microRNAs in brain evolution and development. RNA Biol. 2015;12(7):701-708. doi:10.1080/15476286.2015.1048954
159. Devanna P, Vernes SC. A direct molecular link between the autism candidate gene RORa and the schizophrenia candidate MIR137. Sci Rep. 2015;4(1):3994. doi:10.1038/srep03994
160. Jin P, Alisch RS, Warren ST. RNA and microRNAs in fragile X mental retardation. Nat Cell Biol. 2004;6(11):1048-1053. doi:10.1038/ncb1104-1048
161. Bădescu GM, Fîlfan M, Sandu RE, Surugiu R, Ciobanu O, Popa-Wagner A. Molecular mechanisms underlying neurodevelopmental disorders, ADHD and autism. Romanian J Morphol Embryol Rev Roum Morphol Embryol. 2016;57(2):361-366.
162. Mor M, Nardone S, Sams DS, Elliott E. Hypomethylation of miR-142 promoter and upregulation of microRNAs that target the oxytocin receptor gene in the autism prefrontal cortex. Mol Autism. 2015;6(1):46. doi:10.1186/s13229-015-0040-1
163. Ander BP, Barger N, Stamova B, Sharp FR, Schumann CM. Atypical miRNA expression in temporal cortex associated with dysregulation of immune, cell cycle, and other pathways in autism spectrum disorders. Mol Autism. 2015;6(1):37. doi:10.1186/s13229-015-0029-9
164. Abu-Elneel K, Liu T, Gazzaniga FS, et al. Heterogeneous dysregulation of microRNAs across the autism spectrum. neurogenetics. 2008;9(3):153-161. doi:10.1007/s10048-008-0133-5
165. Huang F, Long Z, Chen Z, et al. Investigation of Gene Regulatory Networks Associated with Autism Spectrum Disorder Based on MiRNA Expression in China. PLOS ONE. 2015;10(6):e0129052. doi:10.1371/journal.pone.0129052
166. Kichukova TM, Popov NT, Ivanov IS, Vachev TI. Profiling of Circulating Serum MicroRNAs in Children with Autism Spectrum Disorder using Stem-loop qRT-PCR Assay. Folia Med (Plovdiv). 2017;59(1):43-52. doi:10.1515/folmed-2017-0009

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

167. Mundalil Vasu M, Anitha A, Thanseem I, et al. Serum microRNA profiles in children with autism. Mol Autism. 2014;5(1):40. doi:10.1186/2040-2392-5-40
168. Nakata M, Kimura R, Funabiki Y, Awaya T, Murai T, Hagiwara M. MicroRNA profiling in adults with high-functioning autism spectrum disorder. Mol Brain. 2019;12(1):82. doi:10.1186/s13041-019-0508-6
169. Ozkul Y, Taheri S, Bayram KK, et al. A heritable profile of six miRNAs in autistic patients and mouse models. Sci Rep. 2020;10(1):9011. doi:10.1038/s41598-020-65847-8
170. Popov NT, Madjirova NP, Minkov IN, Vachev TI. Micro RNA HSA-486-3P Gene Expression Profiling in the Whole Blood of Patients with Autism. Biotechnol Biotechnol Equip. 2012;26(6):3385-3388. doi:10.5504/BBEQ.2012.0093
171. Nguyen LS, Lepleux M, Makhlouf M, et al. Profiling olfactory stem cells from living patients identifies miRNAs relevant for autism pathophysiology. Mol Autism. 2016;7(1):1. doi:10.1186/s13229-015-0064-6
172. Talebizadeh Z, Butler MG, Theodoro MF. Feasibility and relevance of examining lymphoblastoid cell lines to study role of microRNAs in autism. Autism Res. 2008;1(4):240-250. doi:10.1002/aur.33
173. Ghahramani Seno MM, Hu P, Gwadry FG, et al. Gene and miRNA expression profiles in autism spectrum disorders. Brain Res. 2011;1380:85-97. doi:10.1016/j.brainres.2010.09.046
174. Vaishnavi V, Manikandan M, Tiwary BK, Munirajan AK. Insights on the Functional Impact of MicroRNAs Present in Autism-Associated Copy Number Variants. PLoS ONE. 2013;8(2):e56781. doi:10.1371/journal.pone.0056781
175. Brum CB, Paixão-Côrtes VR, Carvalho AM, et al. Genetic variants in miRNAs differentially expressed during brain development and their relevance to psychiatric disorders susceptibility. World J Biol Psychiatry Off J World Fed Soc Biol Psychiatry. 2021;22(6):456-467. doi:10.1080/15622975.2020.1834618
176. Toma C, Torrico B, Hervás A, et al. Common and rare variants of microRNA genes in autism spectrum disorders. World J Biol Psychiatry Off J World Fed Soc Biol Psychiatry. 2015;16(6):376-386. doi:10.3109/15622975.2015.1029518
177. Williams SM, An JY, Edson J, et al. An integrative analysis of non-coding regulatory DNA variations associated with autism spectrum disorder. Mol Psychiatry. 2019;24(11):1707-1719. doi:10.1038/s41380-018-0049-x
178. Schepici, Cavalli, Bramanti, Mazzon. Autism Spectrum Disorder and miRNA: An Overview of Experimental Models. Brain Sci. 2019;9(10):265. doi:10.3390/brainsci9100265
179. Siwach P. Tandem repeats in human disorders: mechanisms and evolution. Front Biosci. 2008;13:4467. doi:10.2741/3017
180. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 2009;10(10):691-703. doi:10.1038/nrg2640
181. Usdin K. The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases: Table 1. Genome Res. 2008;18(7):1011-1019. doi:10.1101/gr.070409.107
182. Mitra I, Huang B, Mousavi N, et al. Patterns of de novo tandem repeat mutations and their role in autism. Nature. 2021;589(7841):246-250. doi:10.1038/s41586-020-03078-7

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

183. Parikshak NN, Swarup V, Belgard TG, et al. Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. Nature. 2016;540(7633):423-427. doi:10.1038/nature20612
184. Smith RM, Sadee W. Synaptic Signaling and Aberrant RNA Splicing in Autism Spectrum Disorders. Front Synaptic Neurosci. 2011;3. doi:10.3389/fnsyn.2011.00001
185. Irimia M, Weatheritt RJ, Ellis JD, et al. A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains. Cell. 2014;159(7):1511-1523. doi:10.1016/j.cell.2014.11.035
186. Gonatopoulos-Pournatzis T, Wu M, Braunschweig U, et al. Genome-wide CRISPR-Cas9 Interrogation of Splicing Networks Reveals a Mechanism for Recognition of Autism-Misregulated Neuronal Microexons. Mol Cell. 2018;72(3):510-524.e12. doi:10.1016/j.molcel.2018.10.008
187. González-Castañeda RE, Sánchez-González VJ, Flores-Soto M, Vázquez-Camacho G, Macías-Islas MA, Ortiz GG. Neural restrictive silencer factor and choline acetyltransferase expression in cerebral tissue of Alzheimer's Disease patients: a pilot study. Genet Mol Biol. 2013;36(1):025-036. doi:10.1590/S1415-47572013000100005
188. Borges-Monroy R, Chu C, Dias C, et al. Whole-genome analysis reveals the contribution of non-coding de novo transposon insertions to autism spectrum disorder. Mob DNA. 2021;12(1):28. doi:10.1186/s13100-021-00256-w
189. Misiak B, Ricceri L, Sąsiadek MM. Transposable Elements and Their Epigenetic Regulation in Mental Disorders: Current Evidence in the Field. Front Genet. 2019;10:580. doi:10.3389/fgene.2019.00580
190. Douglas AGL, Wood MJA. RNA splicing: disease and therapy. Brief Funct Genomics. 2011;10(3):151-164. doi:10.1093/bfgp/elr020
191. Scotti MM, Swanson MS. RNA mis-splicing in disease. Nat Rev Genet. 2016;17(1):19-32. doi:10.1038/nrg.2015.3
192. Wang Y, Wang Z. Systematical identification of splicing regulatory cis-elements and cognate trans-factors. Methods San Diego Calif. 2014;65(3):350-358. doi:10.1016/j.ymeth.2013.08.019
193. Chen M, Manley JL. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nat Rev Mol Cell Biol. 2009;10(11):741-754. doi:10.1038/nrm2777
194. Raj B, Blencowe BJ. Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. Neuron. 2015;87(1):14-27. doi:10.1016/j.neuron.2015.05.004
195. Okay K, Varış PÜ, Miral S, et al. Alternative splicing and gene co-expression network-based analysis of dizygotic twins with autism-spectrum disorder and their parents. Genomics. 2021;113(4):2561-2571. doi:10.1016/j.ygeno.2021.05.038
196. Stamova BS, Tian Y, Nordahl CW, et al. Evidence for differential alternative splicing in blood of young boys with autism spectrum disorders. Mol Autism. 2013;4(1):30. doi:10.1186/2040-2392-4-30

usc UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

197. Gonatopoulos-Pournatzis T, Blencowe BJ. Microexons: at the nexus of nervous system development, behaviour and autism spectrum disorder. Curr Opin Genet Dev. 2020;65:22-33. doi:10.1016/j.gde.2020.03.007
198. Gandal MJ, Zhang P, Hadjimichael E, et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science. 2018;362(6420):eaat8127. doi:10.1126/science.aat8127
199. Quesnel-Vallières M, Dargaei Z, Irimia M, et al. Misregulation of an Activity-Dependent Splicing Network as a Common Mechanism Underlying Autism Spectrum Disorders. Mol Cell. 2016;64(6):1023-1034. doi:10.1016/j.molcel.2016.11.033
200. Quesnel-Vallières M, Irimia M, Cordes SP, Blencowe BJ. Essential roles for the splicing regulator nSR100/SRRM4 during nervous system development. Genes Dev. 2015;29(7):746-759. doi:10.1101/gad.256115.114
201. Zhou J, Park CY, Theesfeld CL, et al. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. Nat Genet. 2019;51(6):973-980. doi:10.1038/s41588-019-0420-0
202. Li YI, Sanchez-Pulido L, Haerty W, Ponting CP. RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. Genome Res. 2015;25(1):1-13. doi:10.1101/gr.181990.114
203. Lovci MT, Ghanem D, Marr H, et al. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nat Struct Mol Biol. 2013;20(12):1434-1442. doi:10.1038/nsmb.2699
204. Voineagu I, Wang X, Johnston P, et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature. 2011;474(7351):380-384. doi:10.1038/nature10110
205. Gueroussov S, Gonatopoulos-Pournatzis T, Irimia M, et al. An alternative splicing event amplifies evolutionary differences between vertebrates. Science. 2015;349(6250):868-873. doi:10.1126/science.aaa8381
206. Dweep H, Sticht C, Gretz N. In-Silico Algorithms for the Screening of Possible microRNA Binding Sites and Their Interactions. Curr Genomics. 2013;14(2):127-136. doi:10.2174/1389202911314020005
207. Riffo-Campos Á, Riquelme I, Brebi-Mieville P. Tools for Sequence-Based miRNA Target Prediction: What to Choose? Int J Mol Sci. 2016;17(12):1987. doi:10.3390/ijms17121987
208. Sample PJ, Wang B, Reid DW, et al. Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. Nat Biotechnol. 2019;37(7):803-809. doi:10.1038/s41587-019-0164-5
209. Wegrzyn JL, Drudge TM, Valafar F, Hook V. Bioinformatic analyses of mammalian 5'-UTR sequence properties of mRNAs predicts alternative translation initiation sites. BMC Bioinformatics. 2008;9(1):232. doi:10.1186/1471-2105-9-232
210. Rudie JD, Hernandez LM, Brown JA, et al. Autism-Associated Promoter Variant in MET Impacts Functional and Structural Brain Networks. Neuron. 2012;75(5):904-915. doi:10.1016/j.neuron.2012.07.010

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

211. Yang SY, Cho SC, Yoo HJ, et al. Association study between single nucleotide polymorphisms in promoter region of AVPR1A and Korean autism spectrum disorders. Neurosci Lett. 2010;479(3):197-200. doi:10.1016/j.neulet.2010.05.050
212. Koesterich J, An JY, Inoue F, et al. Characterization of De Novo Promoter Variants in Autism Spectrum Disorder with Massively Parallel Reporter Assays. Int J Mol Sci. 2023;24(4):3509. doi:10.3390/ijms24043509
213. Daoud H, Bonnet-Brilhault F, Védrine S, et al. Autism and Nonsyndromic Mental Retardation Associated with a De Novo Mutation in the NLGN4X Gene Promoter Causing an Increased Expression Level. Biol Psychiatry. 2009;66(10):906-910. doi:10.1016/j.biopsych.2009.05.008
214. Zhang L, Liu L, Wen Y, et al. Genome-wide association study and identification of chromosomal enhancer maps in multiple brain regions related to autism spectrum disorder. Autism Res. 2019;12(1):26-32. doi:10.1002/aur.2001
215. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat Genet. 2018;50(8):1171-1179. doi:10.1038/s41588-018-0160-6
216. Arrigoni A, Ranzani V, Rossetti G, et al. Analysis RNA-seq and Noncoding RNA. Methods Mol Biol Clifton NJ. 2016;1480:125-135. doi:10.1007/978-1-4939-6380-5_11
217. Ilott NE, Ponting CP. Predicting long non-coding RNAs using RNA sequencing. Methods. 2013;63(1):50-59. doi:10.1016/j.ymeth.2013.03.019
218. Khadirnaikar S, Chatterjee A, Kumar P, Shukla S. A Greedy Algorithm-Based Stem Cell LncRNA Signature Identifies a Novel Subgroup of Lung Adenocarcinoma Patients With Poor Prognosis. Front Oncol. 2020;10:1203. doi:10.3389/fonc.2020.01203
219. Sun Y, Zhao H, Zhou G, Guan T, Wang Y, Gao J. Random distributed logistic regression framework for predicting potential lncRNA–disease association. J Mol Cell Biol. 2021;13(5):386-388. doi:10.1093/jmcb/mjab005
220. Zhu R, Wang Y, Liu JX, Dai LY. IPCARF: improving lncRNA-disease association prediction using incremental principal component analysis feature selection and a random forest classifier. BMC Bioinformatics. 2021;22(1):175. doi:10.1186/s12859-021-04104-9
221. Wang J, Wang L. Prediction and prioritization of autism-associated long non-coding RNAs using gene expression and sequence features. BMC Bioinformatics. 2020;21(1):505. doi:10.1186/s12859-020-03843-5
222. Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. Profiling the genome-wide landscape of tandem repeat expansions. Nucleic Acids Res. 2019;47(15):e90-e90. doi:10.1093/nar/gkz501
223. Dolzhenko E, Bennett MF, Richmond PA, et al. ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. Genome Biol. 2020;21(1):102. doi:10.1186/s13059-020-02017-z
224. Riepe TV, Khan M, Roosing S, Cremers FPM, 't Hoen PAC. Benchmarking deep learning splice prediction tools using functional splice assays. Hum Mutat. 2021;42(7):799-810. doi:10.1002/humu.24212

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

225. Xiong HY, Alipanahi B, Lee LJ, et al. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015;347(6218):1254806. doi:10.1126/science.1254806
226. Cheng J, Çelik MH, Kundaje A, Gagneur J. MTSplice predicts effects of genetic variants on tissue-specific splicing. Genome Biol. 2021;22(1):94. doi:10.1186/s13059-021-02273-7
227. Ross PJ, Mok RSF, Smith BS, et al. Modeling neuronal consequences of autism-associated gene regulatory variants with human induced pluripotent stem cells. Mol Autism. 2020;11(1):33. doi:10.1186/s13229-020-00333-6
228. Lieberman-Aiden E, van Berkum NL, Williams L, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science. 2009;326(5950):289-293. doi:10.1126/science.1181369
229. Berkum N, Lieberman-Aiden E, Williams L, et al. Hi-C: A Method to Study the Three-dimensional Architecture of Genomes. J Vis Exp JoVE. 2010;39. doi:10.3791/1869
230. Dong S, Boyle AP. Prioritization of regulatory variants with tissue-specific function in the non-coding regions of human genome. Nucleic Acids Res. 2022;50(1):e6. doi:10.1093/nar/gkab924
231. Jiang S, Mortazavi A. Integrating ChIP-seq with other functional genomics data. Brief Funct Genomics. 2018;17(2):104-115. doi:10.1093/bfgp/ely002
232. Turner TN, Hormozdiari F, Duyzend MH, et al. Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. Am J Hum Genet. 2016;98(1):58-74. doi:10.1016/j.ajhg.2015.11.023
233. Packer A. Neocortical neurogenesis and the etiology of autism spectrum disorder. Neurosci Biobehav Rev. 2016;64:185-195. doi:10.1016/j.neubiorev.2016.03.002
234. Doan RN, Bae BI, Cubelos B, et al. Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. Cell. 2016;167(2):341-354.e12. doi:10.1016/j.cell.2016.08.071
235. Castelijns B, Baak ML, Timpanaro IS, et al. Hominin-specific regulatory elements selectively emerged in oligodendrocytes and are disrupted in autism patients. Nat Commun. 2020;11(1):301. doi:10.1038/s41467-019-14269-w
236. Rodin RE, Dou Y, Kwon M, et al. The landscape of somatic mutation in cerebral cortex of autistic and neurotypical individuals revealed by ultra-deep whole-genome sequencing. Nat Neurosci. 2021;24(2):176-185. doi:10.1038/s41593-020-00765-6
237. Trevino AE, Müller F, Andersen J, et al. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. Cell. 2021;184(19):5053-5069.e23. doi:10.1016/j.cell.2021.07.039
238. De Rubeis S, He X, Goldberg AP, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014;515(7526):209-215. doi:10.1038/nature13772
239. Jiang CC, Lin LS, Long S, et al. Signalling pathways in autism spectrum disorder: mechanisms and therapeutic implications. Signal Transduct Target Ther. 2022;7:229. doi:10.1038/s41392-022-01081-0
240. Kelleher RJ, Bear MF. The Autistic Neuron: Troubled Translation? Cell. 2008;135(3):401-406. doi:10.1016/j.cell.2008.10.017

241. Santini E, Huynh TN, MacAskill AF, et al. Exaggerated translation causes synaptic and behavioural aberrations associated with autism. Nature. 2013;493(7432):411-415. doi:10.1038/nature11782
242. Gkogkas CG, Khoutorsky A, Ran I, et al. Autism-related deficits via dysregulated eIF4E-dependent translational control. Nature. 2013;493(7432):371-377. doi:10.1038/nature11628
243. Pinto D, Delaby E, Merico D, et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am J Hum Genet. 2014;94(5):677-694. doi:10.1016/j.ajhg.2014.03.018
244. Xing L, Larsen RS, Bjorklund GR, et al. Layer specific and general requirements for ERK/MAPK signaling in the developing neocortex. eLife. 2016;5:e11123. doi:10.7554/eLife.11123
245. Caracci MO, Avila ME, Espinoza-Cavieres FA, López HR, Ugarte GD, De Ferrari GV. Wnt/β-Catenin-Dependent Transcription in Autism Spectrum Disorders. Front Mol Neurosci. 2021;14:764756. doi:10.3389/fnmol.2021.764756
246. Kwan V, Unda BK, Singh KK. Wnt signaling networks in autism spectrum disorder and intellectual disability. J Neurodev Disord. 2016;8. doi:10.1186/s11689-016-9176-3
247. Judson MC, Eagleson KL, Levitt P. A new synaptic player leading to autism risk: Met receptor tyrosine kinase. J Neurodev Disord. 2011;3(3):282-292. doi:10.1007/s11689-011-9081-8
248. de la Torre-Ubieta L, Won H, Stein JL, Geschwind DH. Advancing the understanding of autism disease mechanisms through genetics. Nat Med. 2016;22(4):345-361. doi:10.1038/nm.4071
249. Choquet D, Triller A. The dynamic synapse. Neuron. 2013;80(3):691-703. doi:10.1016/j.neuron.2013.10.013
250. Krumm N, O'Roak BJ, Shendure J, Eichler EE. A de novo convergence of autism genetics and molecular neuroscience. Trends Neurosci. 2014;37(2):95-105. doi:10.1016/j.tins.2013.11.005
251. Wen Z, Cheng TL, Li G zhi, et al. Identification of autism-related MECP2 mutations by whole-exome sequencing and functional validation. Mol Autism. 2017;8:43. doi:10.1186/s13229-017-0157-5
252. Zhubi A, Chen Y, Dong E, Cook EH, Guidotti A, Grayson DR. Increased binding of MeCP2 to the GAD1 and RELN promoters may be mediated by an enrichment of 5-hmC in autism spectrum disorder (ASD) cerebellum. Transl Psychiatry. 2014;4(1):e349. doi:10.1038/tp.2013.123
253. Sun W, Poschmann J, Rosario RCH del, et al. Histone Acetylome-wide Association Study of Autism Spectrum Disorder. Cell. 2016;167(5):1385-1397.e11. doi:10.1016/j.cell.2016.10.031
254. Stessman HAF, Xiong B, Coe BP, et al. Targeted sequencing identifies 91 neurodevelopmental-disorder risk genes with autism and developmental-disability biases. Nat Genet. 2017;49(4):515-526. doi:10.1038/ng.3792

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

255. Hoffmann A, Spengler D. Chromatin Remodeler CHD8 in Autism and Brain Development. J Clin Med. 2021;10(2):366. doi:10.3390/jcm10020366
256. Wang P, Mokhtari R, Pedrosa E, et al. CRISPR/Cas9-mediated heterozygous knockout of the autism gene CHD8 and characterization of its transcriptional networks in cerebral organoids derived from iPS cells. Mol Autism. 2017;8(1):11. doi:10.1186/s13229-017-0124-1
257. Yoon SH, Choi J, Lee WJ, Do JT. Genetic and Epigenetic Etiology Underlying Autism Spectrum Disorder. J Clin Med. 2020;9(4):966. doi:10.3390/jcm9040966
258. Wray NR, Ripke S, Mattheisen M, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat Genet. 2018;50(5):668-681. doi:10.1038/s41588-018-0090-3
259. Sánchez-Mora C, Ramos-Quiroga JA, Bosch R, et al. Case–Control Genome-Wide Association Study of Persistent Attention-Deficit Hyperactivity Disorder Identifies FBXO33 as a Novel Susceptibility Gene for the Disorder. Neuropsychopharmacology. 2015;40(4):915-926. doi:10.1038/npp.2014.267
260. Demontis D, Walters RK, Martin J, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nat Genet. 2019;51(1):63-75. doi:10.1038/s41588-018-0269-7
261. Pardiñas AF, Holmans P, Pocklington AJ, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nat Genet. 2018;50(3):381-389. doi:10.1038/s41588-018-0059-2
262. Sullivan PF, Agrawal A, Bulik CM, et al. Psychiatric Genomics: An Update and an Agenda. Am J Psychiatry. 2018;175(1):15-27. doi:10.1176/appi.ajp.2017.17030283
263. Waring SC, Rosenberg RN. Genome-Wide Association Studies in Alzheimer Disease. Arch Neurol. 2008;65(3):329-334. doi:10.1001/archneur.65.3.329
264. Neale BM, Medland SE, Ripke S, et al. Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. J Am Acad Child Adolesc Psychiatry. 2010;49(9):884-897. doi:10.1016/j.jaac.2010.06.008
265. Wray NR, Pergadia ML, Blackwood DHR, et al. Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. Mol Psychiatry. 2012;17(1):36-48. doi:10.1038/mp.2010.109
266. Sullivan PF. DON'T GIVE UP ON GWAS. Mol Psychiatry. 2012;17(1):2-3. doi:10.1038/mp.2011.94
267. Lasky-Su J, Neale BM, Franke B, et al. Genome-wide association scan of quantitative traits for attention deficit hyperactivity disorder identifies novel associations and confirms candidate gene associations. Am J Med Genet Part B Neuropsychiatr Genet Off Publ Int Soc Psychiatr Genet. 2008;147B(8):1345-1354. doi:10.1002/ajmg.b.30867
268. Demontis D, Walters GB, Athanasiadis G, et al. Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains. Nat Genet. 2023;55(2):198-208. doi:10.1038/s41588-022-01285-8
269. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511(7510):421-427. doi:10.1038/nature13595

270. Ripke S, O'Dushlaine C, Chambert K, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet. 2013;45(10):1150-1159. doi:10.1038/ng.2742
271. Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium, Ripke S, Wray NR, et al. A mega-analysis of genome-wide association studies for major depressive disorder. Mol Psychiatry. 2013;18(4):497-511. doi:10.1038/mp.2012.21
272. Cai N, Bigdeli TB, Kretzschmar W, et al. Sparse whole-genome sequencing identifies two loci for major depressive disorder. Nature. 2015;523(7562):588-591. doi:10.1038/nature14659
273. Power RA, Tansey KE, Buttenschøn HN, et al. Genome-wide Association for Major Depression Through Age at Onset Stratification: Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. Biol Psychiatry. 2017;81(4):325-335. doi:10.1016/j.biopsych.2016.05.010
274. Hyde CL, Nagle MW, Tian C, et al. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. Nat Genet. 2016;48(9):1031-1036. doi:10.1038/ng.3623
275. Howard DM, Adams MJ, Clarke TK, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci. 2019;22(3):343-352. doi:10.1038/s41593-018-0326-7
276. Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. J Open Source Softw. 2018;3(25):731. doi:10.21105/joss.00731
277. Hormozdiari F, van de Bunt M, Segrè AV, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet. 2016;99(6):1245-1260. doi:10.1016/j.ajhg.2016.10.003
278. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. PLOS Genet. 2010;6(4):e1000888. doi:10.1371/journal.pgen.1000888
279. Wu Z, Qin J, You Y, et al. Genetic variants in the transcription regulatory region of MEGF10 are associated with autism in Chinese Han population. Sci Rep. 2017;7(1):2292. doi:10.1038/s41598-017-02348-1
280. Hickman AR, Selee B, Pauly R, Husain B, Hang Y, Feltus FA. Discovery of eQTL Alleles Associated with Autism Spectrum Disorder: A Case-Control Study. J Autism Dev Disord. 2023;53(9):3595-3612. doi:10.1007/s10803-022-05631-x
281. Rodriguez-Fontenla C, Carracedo A. UTMOST, a single and cross-tissue TWAS (Transcriptome Wide Association Study), reveals new ASD (Autism Spectrum Disorder) associated genes. Transl Psychiatry. 2021;11(256). doi:; https://doi.org/10.1101/2020.06.11.145524
282. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369(6509):1318-1330. doi:10.1126/science.aaz1776
283. Drivas TG, Lucas A, Ritchie MD. eQTpLot: a user-friendly R package for the visualization of colocalization between eQTL and GWAS signals. BioData Min. 2021;14(1):32. doi:10.1186/s13040-021-00267-6

284. Howard DM, Adams MJ, Clarke TK, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci. 2019;22(3):343-352. doi:10.1038/s41593-018-0326-7
285. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-575. doi:10.1086/519795
286. Cross-Disorder Group of the Psychiatric Genomics Consortium., Cross-Disorder Group of the Psychiatric Genomics Consortium. Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. Cell. 2019;179(7):1469-1482.e11. doi:10.1016/j.cell.2019.11.020
287. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics. 2016;32(10):1479-1485. doi:10.1093/bioinformatics/btv722
288. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci. 2003;100(16):9440-9445. doi:10.1073/pnas.1530509100
289. Võsa U, Claringbould A, Westra HJ, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021;53(9):1300-1310. doi:10.1038/s41588-021-00913-z
290. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017;8(1):1826. doi:10.1038/s41467-017-01261-5
291. Piñero J, Bravo À, Queralt-Rosinach N, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017;45(D1):D833-D839. doi:10.1093/nar/gkw943
292. Golovina E, Fadason T, Lints TJ, Walker C, Vickers MH, O'Sullivan JM. Understanding the impact of SNPs associated with autism spectrum disorder on biological pathways in the human fetal and adult cortex. Sci Rep. 2021;11(1):15867. doi:10.1038/s41598-021-95447-z
293. Piñero J, Bravo À, Queralt-Rosinach N, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017;45(Database issue):D833-D839. doi:10.1093/nar/gkw943
294. Zhang CC, Xing A, Tan MS, Tan L, Yu JT. The Role of MAPT in Neurodegenerative Diseases: Genetics, Mechanisms and Therapy. Mol Neurobiol. 2016;53(7):4893-4904. doi:10.1007/s12035-015-9415-8
295. Strang KH, Golde TE, Giasson BI. MAPT mutations, tauopathy, and mechanisms of neurodegeneration. Lab Invest. 2019;99(7):912-928. doi:10.1038/s41374-019-0197-x
296. Chen J, Lu Y, Cue JM, et al. Genetic Relationship between Alzheimer's Disease and Schizophrenia. Alzheimers Dement. 2022;18(S3):e065861. doi:10.1002/alz.065861
297. Huang K, Wu Y, Shin J, et al. Transcriptome-wide transmission disequilibrium analysis identifies novel risk genes for autism spectrum disorder. PLoS Genet. 2021;17(2):e1009309. doi:10.1371/journal.pgen.1009309

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

298. Hale AJ, den Hertog J. Studying Protein-Tyrosine Phosphatases in Zebrafish. Methods Mol Biol Clifton NJ. 2016;1447:351-372. doi:10.1007/978-1-4939-3746-2_19
299. Hendriks WJAJ, Elson A, Harroch S, Pulido R, Stoker A, den Hertog J. Protein tyrosine phosphatases in health and disease. FEBS J. 2013;280(2):708-730. doi:10.1111/febs.12000
300. Tomita H, Cornejo F, Aranda-Pino B, et al. The Protein Tyrosine Phosphatase Receptor Delta Regulates Developmental Neurogenesis. Cell Rep. 2020;30(1):215-228.e5. doi:10.1016/j.celrep.2019.11.033
301. O'Donnell-Luria AH, Pais LS, Faundes V, et al. Heterozygous Variants in KMT2E Cause a Spectrum of Neurodevelopmental Disorders and Epilepsy. Am J Hum Genet. 2019;104(6):1210-1222. doi:10.1016/j.ajhg.2019.03.021
302. Wang ZH, Liu P, Liu X, Yu SP, Wang JZ, Ye K. Delta-secretase (AEP) mediates tau-splicing imbalance and accelerates cognitive decline in tauopathies. J Exp Med. 2018;215(12):3038-3056. doi:10.1084/jem.20180539
303. Wei H, Alberts I, Li X. The apoptotic perspective of autism. Int J Dev Neurosci. 2014;36(1):13-18. doi:10.1016/j.ijdevneu.2014.04.004
304. Ramesh M, Gopinath P, Govindaraju T. Role of Post-translational Modifications in Alzheimer's Disease. ChemBioChem. 2020;21(8):1052-1079. doi:10.1002/cbic.201900573
305. Zhang Z, Tian Y, Ye K. δ-secretase in neurodegenerative diseases: mechanisms, regulators and therapeutic opportunities. Transl Neurodegener. 2020;9(1):1. doi:10.1186/s40035-019-0179-3
306. Tai C, Chang CW, Yu GQ, et al. Tau Reduction Prevents Key Features of Autism in Mouse Models. Neuron. 2020;106(3):421-437.e11. doi:10.1016/j.neuron.2020.01.038
307. Lettre G. One step closer to linking GWAS SNPs with the right genes. Nat Genet. 2022;54(6):748-749. doi:10.1038/s41588-022-01093-0
308. Liu Z, Liu W, Yao L, et al. Negative life events and corticotropin-releasing-hormone receptor1 gene in recurrent major depressive disorder. Sci Rep. 2013;3(1):1548. doi:10.1038/srep01548
309. Ribbe K, Ackermann V, Schwitulla J, et al. Prediction of the Risk of Comorbid Alcoholism in Schizophrenia by Interaction of Common Genetic Variants in the Corticotropin-Releasing Factor System. Arch Gen Psychiatry. 2011;68(12):1247-1256. doi:10.1001/archgenpsychiatry.2011.100
310. Refojo D, Schweizer M, Kuehne C, et al. Glutamatergic and Dopaminergic Neurons Mediate Anxiogenic and Anxiolytic Effects of CRHR1. Science. 2011;333(6051):1903-1907. doi:10.1126/science.1202107
311. Kumar S, Reynolds K, Ji Y, Gu R, Rai S, Zhou CJ. Impaired neurodevelopmental pathways in autism spectrum disorder: a review of signaling mechanisms and crosstalk. J Neurodev Disord. 2019;11(1):10. doi:10.1186/s11689-019-9268-y
312. Koolen DA, Morgan A, de Vries BB. Koolen-de Vries Syndrome. GeneReviews®. University of Washington, Seattle; 1993-2024. Accessed November 17, 2023. Available from: https://www.ncbi.nlm.nih.gov/books/NBK24676/
313. Chen X, Yao T, Cai J, et al. A novel genetic variant potentially altering the expression of MANBA in the cerebellum associated with attention deficit hyperactivity disorder in Han

Chinese children. World J Biol Psychiatry. 2022;23(7):548-559. doi:10.1080/15622975.2021.2014248
314. Beroun A, Mitra S, Michaluk P, Pijet B, Stefaniuk M, Kaczmarek L. MMPs in learning and memory and neuropsychiatric disorders. Cell Mol Life Sci. 2019;76(16):3207-3228. doi:10.1007/s00018-019-03180-8
315. Werling D, Brand H, An JY, et al. Limited contribution of rare, noncoding variation to Autism Spectrum Disorder from sequencing of 2,076 genomes in quartet families. Eur Neuropsychopharmacol. 2019;29:S784-S785. doi:10.1016/j.euroneuro.2017.08.010
316. Sun Y, Yao X, March ME, et al. Target Genes of Autism Risk Loci in Brain Frontal Cortex. Front Genet. 2019;10:707. doi:10.3389/fgene.2019.00707
317. Arpi MNT, Simpson TI. SFARI genes and where to find them; modelling Autism Spectrum Disorder specific gene expression dysregulation with RNA-seq data. Sci Rep. 2022;12(1):10158. doi:10.1038/s41598-022-14077-1
318. Quesnel-Vallières M, Weatheritt RJ, Cordes SP, Blencowe BJ. Autism spectrum disorder: insights into convergent mechanisms from transcriptomics. Nat Rev Genet. 2019;20(1):51-63. doi:10.1038/s41576-018-0066-2
319. Wang ET, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456(7221):470-476. doi:10.1038/nature07509
320. Chau KK, Zhang P, Urresti J, et al. Full-length isoform transcriptome of the developing human brain provides further insights into autism. Cell Rep. 2021;36(9). doi:10.1016/j.celrep.2021.109631
321. Shah S, Richter JD. Do Fragile X Syndrome and Other Intellectual Disorders Converge at Aberrant Pre-mRNA Splicing? Front Psychiatry. 2021;12:715346. doi:10.3389/fpsyt.2021.715346
322. Li R, Dong Q, Yuan X, et al. Misregulation of Alternative Splicing in a Mouse Model of Rett Syndrome. PLOS Genet. 2016;12(6):e1006129. doi:10.1371/journal.pgen.1006129
323. Urdinguio RG, Fernández AF, Lopez-Nieva P, et al. Disrupted microRNA expression caused by Mecp2 loss in a mouse model of Rett syndrome. Epigenetics. 2010;5(7):656-663. doi:10.4161/epi.5.7.13055
324. Shah S, Molinaro G, Liu B, Wang R, Huber KM, Richter JD. FMRP Control of Ribosome Translocation Promotes Chromatin Modifications and Alternative Splicing of Neuronal Genes Linked to Autism. Cell Rep. 2020;30(13):4459-4472.e6. doi:10.1016/j.celrep.2020.02.076
325. Dawes R, Bournazos AM, Bryen SJ, et al. SpliceVault predicts the precise nature of variant-associated mis-splicing. Nat Genet. 2023;55(2):324-332. doi:10.1038/s41588-022-01293-8
326. Wagner N, Çelik MH, Hölzlwimmer FR, et al. Aberrant splicing prediction across human tissues. Nat Genet. 2023;55(5):861-870. doi:10.1038/s41588-023-01373-3
327. Alonso-Gonzalez A, Calaza M, Amigo J, et al. Exploring the biological role of postzygotic and germinal de novo mutations in ASD. Sci Rep. 2021;11(1):319. doi:10.1038/s41598-020-79412-w

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

328. Lim ET, Uddin M, De Rubeis S, et al. Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. Nat Neurosci. 2017;20(9):1217-1224. doi:10.1038/nn.4598
329. Dawes R, Joshi H, Cooper ST. Empirical prediction of variant-activated cryptic splice donors using population-based RNA-Seq data. Nat Commun. 2022;13(1):1655. doi:10.1038/s41467-022-29271-y
330. Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation. 2021;2(3):100141. doi:10.1016/j.xinn.2021.100141
331. Lord J, Baralle D. Splicing in the Diagnosis of Rare Disease: Advances and Challenges. Front Genet. 2021;12:689892. doi:10.3389/fgene.2021.689892
332. Blakes AJM, Wai HA, Davies I, et al. A systematic analysis of splicing variants identifies new diagnoses in the 100,000 Genomes Project. Genome Med. 2022;14(1):79. doi:10.1186/s13073-022-01087-x
333. Ong CT, Adusumalli S. Increased intron retention is linked to Alzheimer's disease. Neural Regen Res. 2019;15(2):259-260. doi:10.4103/1673-5374.265549
334. Zhang R, He H, Yuan B, et al. An Intronic Variant of CHD7 Identified in Autism Patients Interferes with Neuronal Differentiation and Development. Neurosci Bull. 2021;37(8):1091-1106. doi:10.1007/s12264-021-00685-w
335. Niesler B, Rappold GA. Emerging evidence for gene mutations driving both brain and gut dysfunction in autism spectrum disorder. Mol Psychiatry. 2021;26(5):1442-1444. doi:10.1038/s41380-020-0778-5
336. Yi CX, Tschöp MH. Brain–gut–adipose-tissue communication pathways at a glance. Dis Model Mech. 2012;5(5):583-587. doi:10.1242/dmm.009902
337. Puente-Ruiz SC, Jais A. Reciprocal signaling between adipose tissue depots and the central nervous system. Front Cell Dev Biol. 2022;10:979251. doi:10.3389/fcell.2022.979251
338. Ge T, Fan J, Yang W, Cui R, Li B. Leptin in depression: a potential therapeutic target. Cell Death Dis. 2018;9(11):1-10. doi:10.1038/s41419-018-1129-1
339. Bouret SG. Neurodevelopmental actions of leptin. Brain Res. 2010;1350:2-9. doi:10.1016/j.brainres.2010.04.011
340. Beccano-Kelly D, Harvey J. Leptin: A Novel Therapeutic Target in Alzheimer's Disease? Int J Alzheimer's Dis. 2012;2012:e594137. doi:10.1155/2012/594137
341. McGregor G, Harvey J. Leptin Regulation of Synaptic Function at Hippocampal TA-CA1 and SC-CA1 Synapses: Implications for Health and Disease. Neurochem Res. 2019;44(3):650-660. doi:10.1007/s11064-017-2362-1
342. Naro C, Cesari E, Sette C. Splicing regulation in brain and testis: common themes for highly specialized organs. Cell Cycle. 2021;20(5-6):480-489. doi:10.1080/15384101.2021.1889187
343. Leung CS, Rosenzweig SJ, Yoon B, et al. Dysregulation of the chromatin environment leads to differential alternative splicing as a mechanism of disease in a human model of autism spectrum disorder. Hum Mol Genet. 2023;32(10):1634-1646. doi:10.1093/hmg/ddad002

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

344. Yuen RKC, Thiruvahindrapuram B, Merico D, et al. Whole-genome sequencing of quartet families with autism spectrum disorder. Nat Med. 2015;21(2):185-191. doi:10.1038/nm.3792
345. Turner TN, Coe BP, Dickel DE, et al. Genomic Patterns of De Novo Mutation in Simplex Autism. Cell. 2017;171(3):710-722.e12. doi:10.1016/j.cell.2017.08.047
346. Ellingford JM, Ahn JW, Bagnall RD, et al. Recommendations for clinical interpretation of variants found in non-coding regions of the genome. Genome Med. 2022;14(1):73. doi:10.1186/s13073-022-01073-3
347. Snyder MP, Gingeras TR, Moore JE, et al. Perspectives on ENCODE. Nature. 2020;583(7818):693-698. doi:10.1038/s41586-020-2449-8
348. Chen KM, Wong AK, Troyanskaya OG, Zhou J. A sequence-based global map of regulatory activity for deciphering human genetics. Nat Genet. 2022;54(7):940-949. doi:10.1038/s41588-022-01102-2
349. Gates LA, Foulds CE, O'Malley BW. Histone marks in the 'drivers seat': functional roles in steering the transcription cycle. Trends Biochem Sci. 2017;42(12):977-989. doi:10.1016/j.tibs.2017.10.004
350. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma Oxf Engl. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324
351. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinforma. 2013;43(1110):11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43
352. Carson AR, Smith EN, Matsui H, et al. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. BMC Bioinformatics. 2014;15(1):125. doi:10.1186/1471-2105-15-125
353. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007;81(3):559-575. doi:10.1086/519795
354. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10(2):giab008. doi:10.1093/gigascience/giab008
355. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma Oxf Engl. 2010;26(6):841-842. doi:10.1093/bioinformatics/btq033
356. Helena Mangs A, Morris BJ. The Human Pseudoautosomal Region (PAR): Origin, Function and Future. Curr Genomics. 2007;8(2):129-136.
357. Edwards AWF. G. H. Hardy (1908) and Hardy–Weinberg Equilibrium. Genetics. 2008;179(3):1143-1150. doi:10.1534/genetics.104.92940
358. Sanders SJ, He X, Willsey AJ, et al. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron. 2015;87(6):1215-1233. doi:10.1016/j.neuron.2015.09.016
359. Girirajan S, Rosenfeld JA, Coe BP, et al. Phenotypic Heterogeneity of Genomic Disorders and Rare Copy-Number Variants. N Engl J Med. 2012;367(14):1321-1331. doi:10.1056/NEJMoa1200395

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

360. Jacquemont S, Coe BP, Hersch M, et al. A Higher Mutational Burden in Females Supports a "Female Protective Model" in Neurodevelopmental Disorders. Am J Hum Genet. 2014;94(3):415-425. doi:10.1016/j.ajhg.2014.02.001
361. Sanders SJ, Ercan-Sencicek AG, Hus V, et al. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron. 2011;70(5):863-885. doi:10.1016/j.neuron.2011.05.002
362. Krumm N, Turner TN, Baker C, et al. Excess of rare, inherited truncating mutations in autism. Nat Genet. 2015;47(6):582-588. doi:10.1038/ng.3303
363. Willsey AJ, Fernandez TV, Yu D, et al. De Novo Coding Variants Are Strongly Associated with Tourette Disorder. Neuron. 2017;94(3):486-499.e9. doi:10.1016/j.neuron.2017.04.024
364. Gregor A, Oti M, Kouwenhoven EN, et al. De novo mutations in the genome organizer CTCF cause intellectual disability. Am J Hum Genet. 2013;93(1):124-131. doi:10.1016/j.ajhg.2013.05.007
365. Nakamura T, Ueda J, Mizuno S, et al. Topologically associating domains define the impact of de novo promoter variants on autism spectrum disorder risk. Cell Genomics. 2024;(2). doi:10.1016/j.xgen.2024.100488
366. Oh S, Shao J, Mitra J, et al. Enhancer release and retargeting activates disease-susceptibility genes. Nature. 2021;595(7869):735-740. doi:10.1038/s41586-021-03577-1
367. Kubo N, Ishii H, Xiong X, et al. Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation. Nat Struct Mol Biol. 2021;28(2):152-161. doi:10.1038/s41594-020-00539-5
368. Deng L, Chen L, Zhao L, et al. Ubiquitination of Rheb governs growth factor-induced mTORC1 activation. Cell Res. 2019;29(2):136-150. doi:10.1038/s41422-018-0120-9
369. Wang B, Qin Y, Wu Q, et al. mTOR Signaling Pathway Regulates the Release of Proinflammatory Molecule CCL5 Implicated in the Pathogenesis of Autism Spectrum Disorder. Front Immunol. 2022;13:818518. doi:10.3389/fimmu.2022.818518
370. Obara Y, Sato H, Nakayama T, Kato T, Ishii K. Midnolin is a confirmed genetic risk factor for Parkinson's disease. Ann Clin Transl Neurol. 2019;6(11):2205-2211. doi:10.1002/acn3.50914
371. Butler MG, Rafi SK, Hossain W, Stephan DA, Manzardo AM. Whole Exome Sequencing in Females with Autism Implicates Novel and Candidate Genes. Int J Mol Sci. 2015;16(1):1312-1335. doi:10.3390/ijms16011312

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

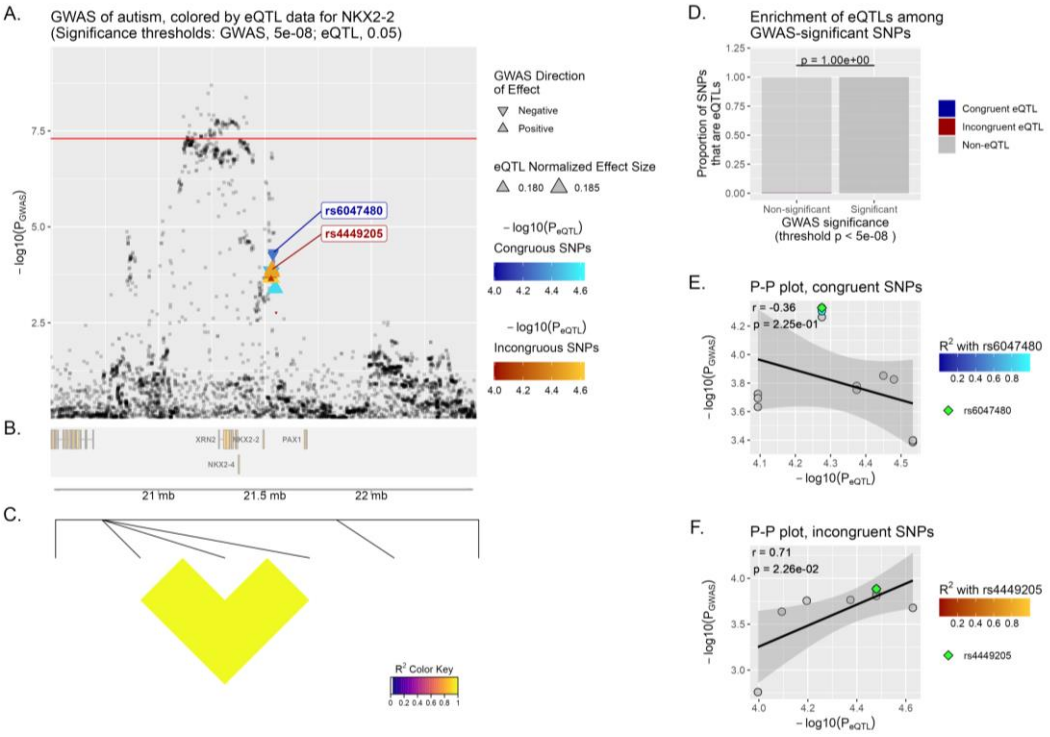
APPENDIX

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

10 APPENDIX

10.1 SUPPLEMENTARY FIGURES

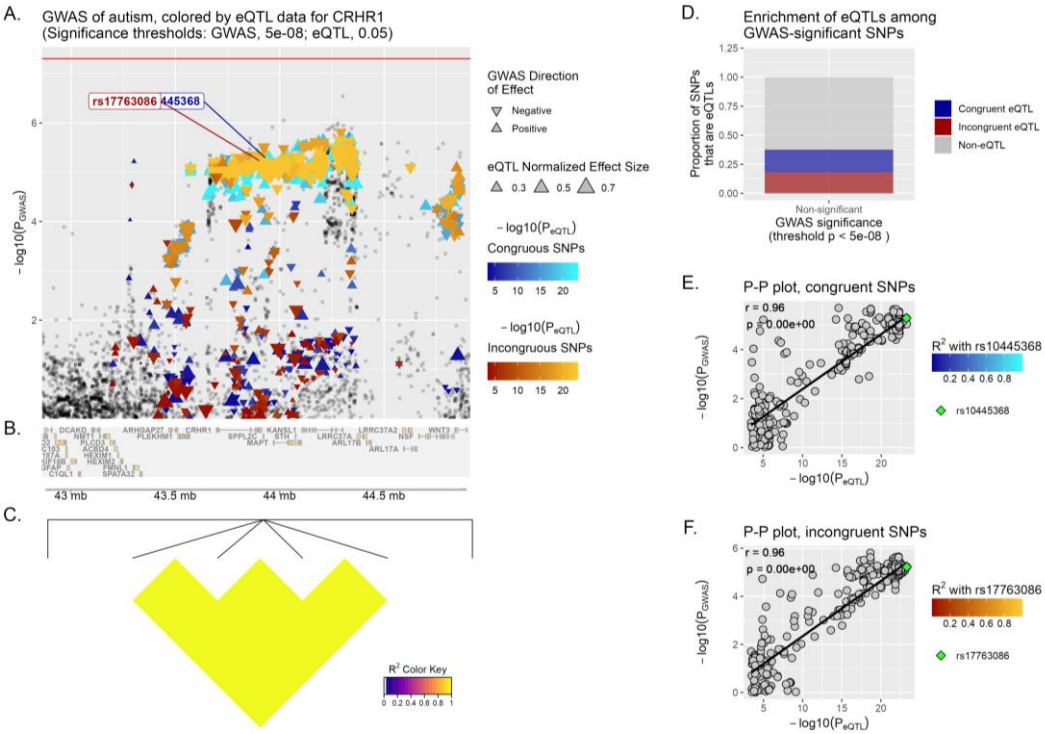
10.1.1 Supplementary Figure 1. eQTpLot for ASD and *NKX2-2* multi-tissue analysis.



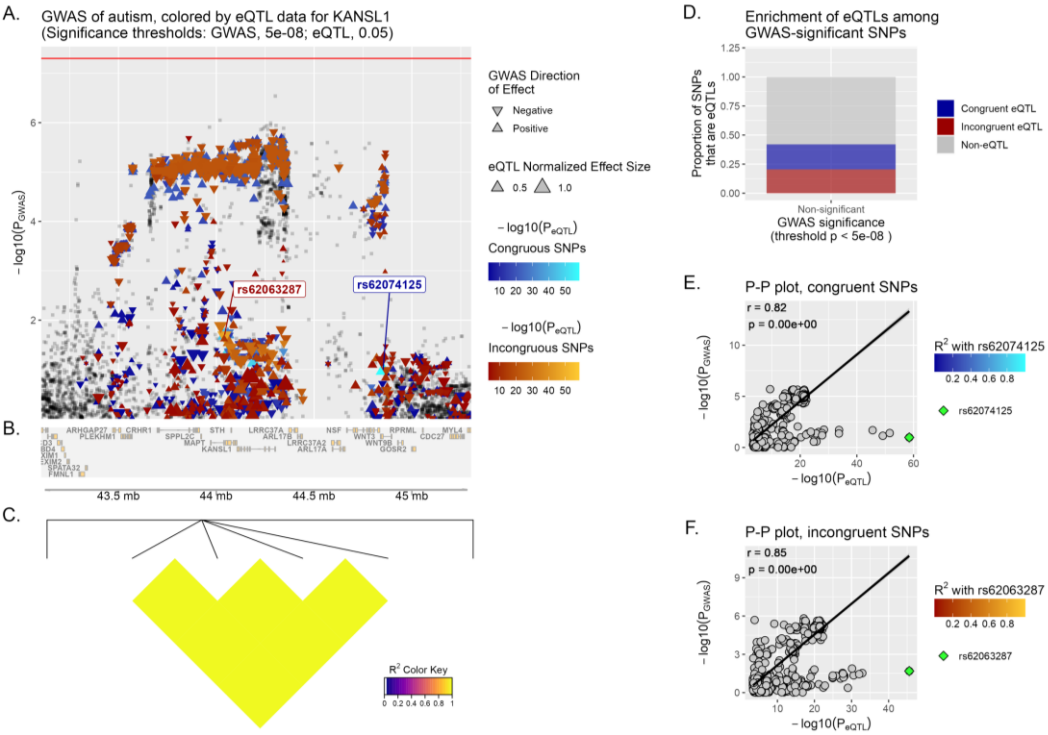
To view the legend for this figure (as well as for Supplementary Figures 2-6), please refer to the caption on Figure 14.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

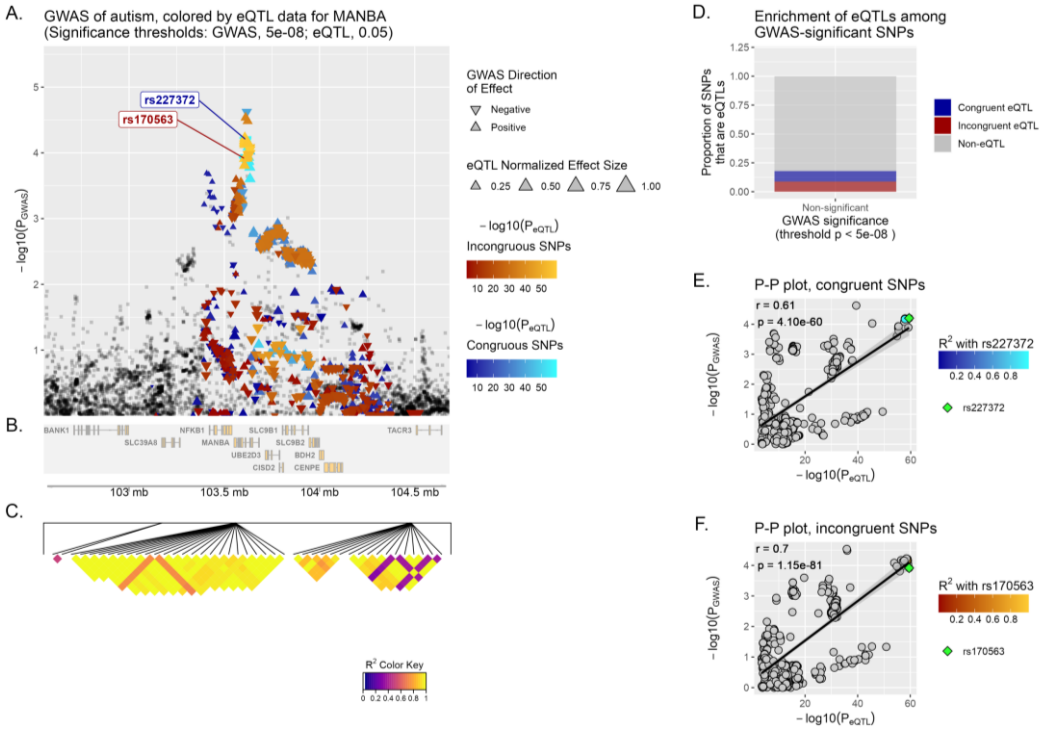
10.1.2 Supplementary Figure 2. eQTpLot eQTpLot for ASD and *CRHR1* pan-tissue analysis.



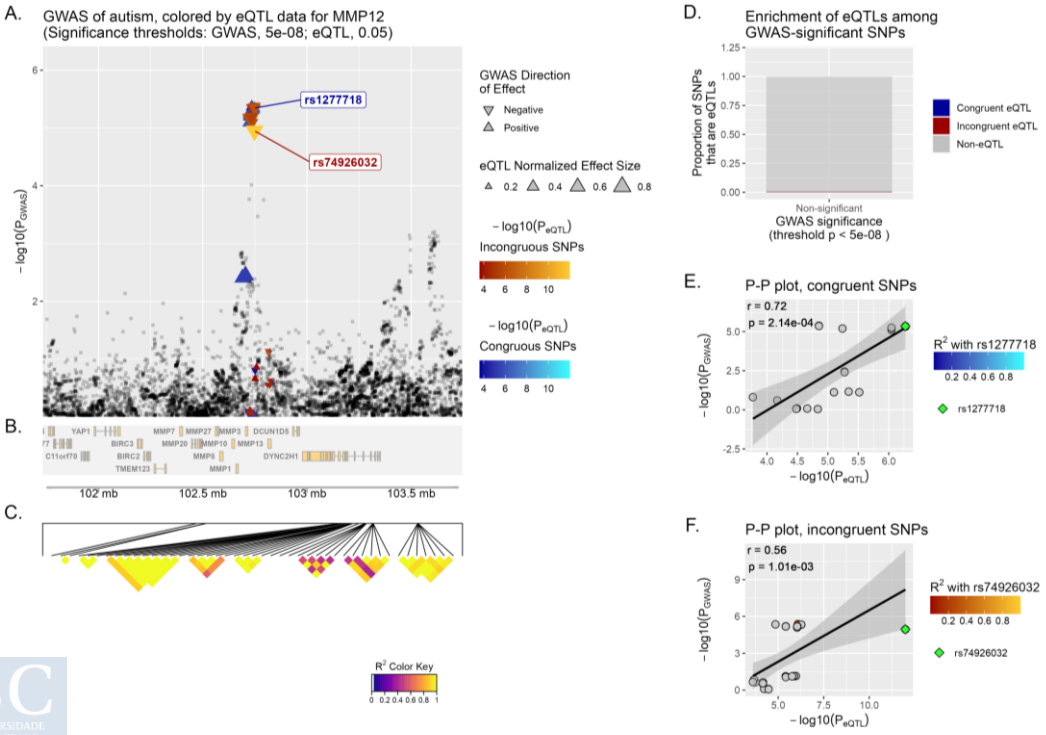
10.1.3 Supplementary Figure 3. eQTpLot for ASD and *KANSL1* pan-tissue analysis.



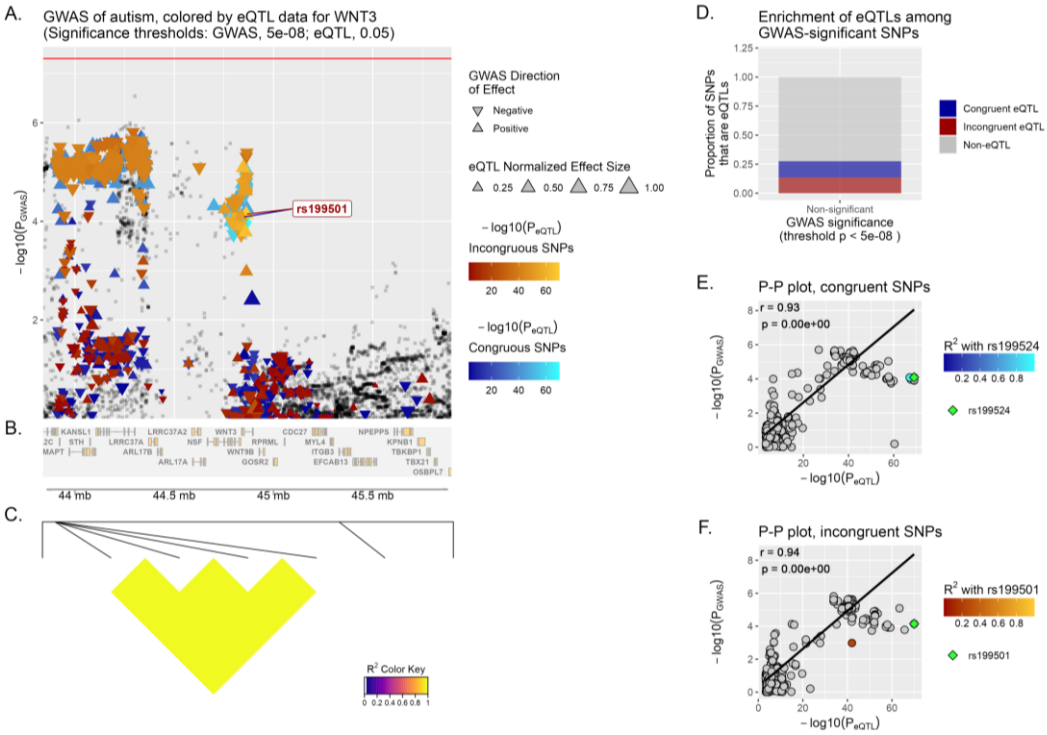
10.1.4 Supplementary Figure 4. eQTpLot for ASD and *MANBA* pan-tissue analysis.



10.1.5 Supplementary Figure 5. eQTpLot for ASD and *MMP12* pan-tissue analysis.

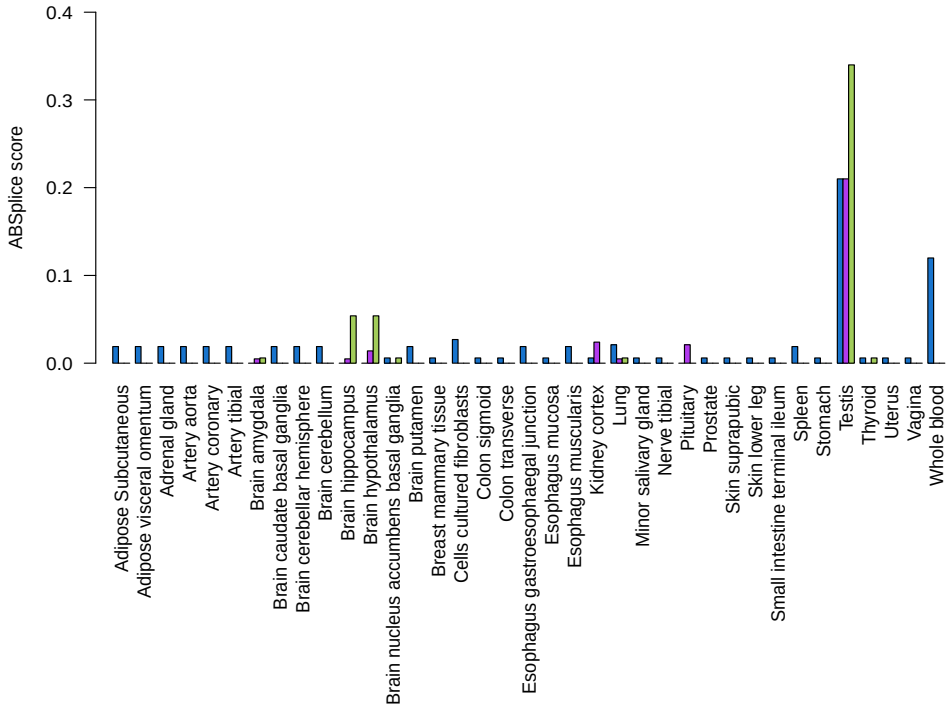


10.1.6 Supplementary Figure 6. eQTpLot for ASD and *WNT3* pan-tissue analysis.



10.1.7 Supplementary Figure 7. Validated variants with the highest impact restricted to testis.

The plot illustrates ABSplice scores for variants with their highest scores in the testis, distinguishing variants within the *CATSPER1* gene in blue, *VWA5B1* in purple, and *TECK4* in green.



10.2 SUPPLEMENTARY TABLES

All supplementary tables not included in this Appendix have been made available at: https://drive.google.com/drive/folders/1bIkf9fkSwQK0j2eE82O1r304u5Qq22_Q?usp=sharing

10.2.1 Supplementary Table 1. *MAPT, NKX2-2* and *PTPRE* functionally related genes in brain tissue.

Symbol	Name
SH2B3	SH2B adaptor protein 3
RAB31	RAB31
ABR	ABR
ETS2	ETS proto-oncogene 2
DUSP6	dual specificity phosphatase 6
HBEGF	heparin binding EGF like growth factor
CSGALNACT2	chondroitin sulfate N-acetylgalactosaminyltransferase 2
NUMB	NUMB
CCR1	C-C motif chemokine receptor 1
ATP2B1	ATPase plasma membrane Ca2+ transporting 1
LTBP2	latent transforming growth factor beta binding protein 2
ARL4C	ADP ribosylation factor like GTPase 4C
IDS	iduronate 2-sulfatase
TRIM8	tripartite motif containing 8
GLIPR1	GLI pathogenesis related 1
HIVEP2	human immunodeficiency virus type I enhancer binding protein 2
ARNTL	aryl hydrocarbon receptor nuclear translocator like
TNFAIP2	TNF alpha induced protein 2
DOCK4	dedicator of cytokinesis 4
VCAN	versican
IRS2	insulin receptor substrate 2
THEMIS2	thymocyte selection associated family member 2
EPHB2	EPH receptor B2
MTMR6	myotubularin related protein 6
SNN	stannin
ITGAM	integrin subunit alpha M
LAPTM5	lysosomal protein transmembrane 5
LRRFIP1	LRR binding FLII interacting protein 1
ATP6V1A	ATPase H+ transporting V1 subunit A
TACC1	transforming acidic coiled-coil containing protein 1
KLF4	Kruppel like factor 4
CXCL1	C-X-C motif chemokine ligand 1
LCP2	lymphocyte cytosolic protein 2
FAM49A	family with sequence similarity 49 member A
TRIM33	tripartite motif containing 33
CREM	cAMP responsive element modulator
IGF2R	insulin like growth factor 2 receptor
ITPR1	inositol 1
IL1R1	interleukin 1 receptor type 1

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ITSN2	intersectin 2
PCDH7	protocadherin 7
PELI1	pellino E3 ubiquitin protein ligase 1
ABCA1	ATP binding cassette subfamily A member 1
PTEN	phosphatase and tensin homolog
KIF1B	kinesin family member 1B
IL1B	interleukin 1 beta
IL13RA1	interleukin 13 receptor subunit alpha 1
SPRED1	sprouty related EVH1 domain containing 1
SIRPA	signal regulatory protein alpha
MICAL2	microtubule associated monooxygenase

10.2.2 Supplementary Table 2. *De novo* variants with high confidence predicted splice altering scores (SpliceAI ≥ 0.8) and *in silico* validation results

Genomic positions are in GRCh37. The last three columns show validation results in SpliceVault, ABSplice and total validation ("Y" stands for validated variants in at least one of the datasets).

10.2.3 Supplementary Table 3. *CACNA1I*, *CBLB*, *CLTB*, *DLGAP1*, *DVL3*, *KIAA0513*, *OFD1*, *PKD1*, *SLC13A3* and *SCN2A* functionally related genes in brain tissue (splicing gene list).

Symbol	Name
SHANK2	SH3 and multiple ankyrin repeat domains 2
CAMK2B	calcium/calmodulin dependent protein kinase II beta
DLGAP4	DLG associated protein 4
EGFR	epidermal growth factor receptor
THRA	thyroid hormone receptor alpha
DDR1	discoidin domain receptor tyrosine kinase 1
NCOR2	nuclear receptor corepressor 2
CSNK1D	casein kinase 1 delta
PDLIM7	PDZ and LIM domain 7
FKBP8	FK506 binding protein 8
ATN1	atrophin 1
MAPK8IP3	mitogen-activated protein kinase 8 interacting protein 3
DNAJC4	DnaJ heat shock protein family (Hsp40) member C4
GNA11	G protein subunit alpha 11
SYN2	synapsin II
NTRK3	neurotrophic receptor tyrosine kinase 3
BCAN	brevican
MINK1	misshapen like kinase 1
ADORA1	adenosine A1 receptor
NECTIN1	nectin cell adhesion molecule 1
DMWD	DM1 locus WD repeat containing
TAOK2	TAO kinase 2
ALPG	alkaline phosphatase germ cell
ZER1	zyg-11 related cell cycle regulator
MAZ	MYC associated zinc finger protein

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ATP2B2	ATPase plasma membrane Ca2+ transporting 2
MPP2	membrane palmitoylated protein 2
EFNB3	ephrin B3
ATP6V0C	ATPase H+ transporting V0 subunit c
USP34	ubiquitin specific peptidase 34
SPTAN1	spectrin alpha non-erythrocytic 1
NEBL	nebulette
IDS	iduronate 2-sulfatase
MAP1A	microtubule associated protein 1A
HGS	hepatocyte growth factor-regulated tyrosine kinase substrate
ACTL6B	actin like 6B
ADCY1	adenylate cyclase 1
RARB	retinoic acid receptor beta
MAPK8IP2	mitogen-activated protein kinase 8 interacting protein 2
NUMA1	nuclear mitotic apparatus protein 1
ACVR1B	activin A receptor type 1B
KLC1	kinesin light chain 1
AP2A2	adaptor related protein complex 2 subunit alpha 2
TRPM3	transient receptor potential cation channel subfamily M member 3
AGRN	agrin
VAMP2	vesicle associated membrane protein 2
GRIN1	glutamate ionotropic receptor NMDA type subunit 1
MZF1	myeloid zinc finger 1
TNK2	tyrosine kinase non receptor 2
ADCY2	adenylate cyclase 2
DVL3	dishevelled segment polarity protein 3
PKD1	polycystin 1 transient receptor potential channel interacting
DLGAP1	DLG associated protein 1
CACNA1I	calcium voltage-gated channel subunit alpha1 I
SLC13A3	solute carrier family 13 member 3
KIAA0513	KIAA0513
SCN2A	sodium voltage-gated channel alpha subunit 2
CBLB	Cbl proto-oncogene B
CLTB	clathrin light chain B
OFD1	OFD1 centriole and centriolar satellite protein

10.2.4 Supplementary Table 4. Final list of *de novo* high regulatory impact variants.

10.2.5 Supplementary Table 5. Enrichments in the number of sequence classes assigned to ultra-rare inherited variants in each of ENCODE's cCREs categories.

10.2.6 Supplementary Table 6. Final list of inherited high regulatory impact variants

10.2.7 Supplementary Table 7. Enrichment of TFs among high regulatory impact *de novo* variants

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

10.2.8 Supplementary Table 8. Enrichment of TFs among high regulatory impact inherited variants

10.2.9 Supplementary Table 9. ASD risk genes custom list

Gene-symbol	Gene-name
ABCE1	ATP binding cassette subfamily E member 1
ACHE	acetylcholinesterase (Cartwright blood group)
ACTB	actin beta
ADNP	Activity-dependent neuroprotector homeobox
ADSL	adenylosuccinate lyase
AFF2	AF4/FMR2 family, member 2
AHDC1	AT-hook DNA binding motif containing 1
AKAP9	A-kinase anchoring protein 9
ALDH5A1	aldehyde dehydrogenase 5 family, member A1 (succinate-semialdehyde dehydrogenase)
ANK2	Ankyrin 2, neuronal
ANK3	ankyrin 3
ANKRD11	ankyrin repeat domain 11
ANP32A	acidic nuclear phosphoprotein 32 family member A
AP2S1	adaptor related protein complex 2 subunit sigma 1
APH1A	aph-1 homolog A, gamma-secretase subunit
ARF3	ADP ribosylation factor 3
ARHGEF9	Cdc42 guanine nucleotide exchange factor (GEF) 9
ARID1B	AT-rich interaction domain 1B
ARX	aristaless related homeobox
ASH1L	Ash1 (absent, small, or homeotic)-like (Drosophila)
ASXL3	Additional sex combs like 3 (Drosophila)
ATRX	alpha thalassemia/mental retardation syndrome X-linked
AUTS2	activator of transcription and developmental regulatorAUTS2
BAZ2B	bromodomain adjacent to zinc finger domain 2B
BCKDK	Branched chain ketoacid dehydrogenase kinase
BCL11A	B-cell CLL/lymphoma 11A (zinc finger protein)
BRAF	v-raf murine sarcoma viral oncogene homolog B
BRSK2	BR serine/threonine kinase 2
BTRC	beta-transducin repeat containing E3 ubiquitin protein ligase
CACNA1A	Calcium channel, voltage-dependent, P/Q type, alpha 1A subunit
CACNA1C	calcium channel, voltage-dependent, L type, alpha 1C subunit
CACNA1E	calcium voltage-gated channel subunit alpha1 E
CACNA2D3	Calcium channel, voltage-dependent, alpha 2/delta subunit 3
CAMTA2	calmodulin binding transcription activator 2
CAPN12	calpain 12
CAPRIN1	Cell cycle associated protein 1
CASK	calcium/calmodulin dependent serine protein kinase
CASZ1	castor zinc finger 1
CBL	Cbl proto-oncogene
CCSER1	coiled-coil serine rich protein 1

CDKL5	cyclin-dependent kinase-like 5
CELF4	CUGBP, Elav-like family member 4
CHAMP1	chromosome alignment maintaining phosphoprotein 1
CHD2	Chromodomain helicase DNA binding protein 2
CHD3	chromodomain helicase DNA binding protein 3
CHD7	chromodomain helicase DNA binding protein 7
CHD8	chromodomain helicase DNA binding protein 8
CIC	capicua transcriptional repressor
CMPK2	cytidine/uridine monophosphate kinase 2
CNOT3	CCR4-NOT transcription complex subunit 3
CORO1A	coronin 1A
CPSF7	cleavage and polyadenylation specific factor 7
CREBBP	CREB binding protein
CSDE1	cold shock domain containing E1
CSNK2A1	casein kinase 2 alpha 1
CTCF	CCCTC-binding factor
CTNNB1	catenin beta 1
CTTNBP2	cortactin binding protein 2
CUL3	Cullin 3
DDX3X	DEAD (Asp-Glu-Ala-Asp) box helicase 3, X-linked
DEAF1	DEAF1 transcription factor
DHCR7	7-dehydrocholesterol reductase
DIP2A	DIP2 disco-interacting protein 2 homolog A (Drosophila)
DLG4	discs large MAGUK scaffold protein 4
DMPK	dystrophia myotonica-protein kinase
DNMT3A	DNA (cytosine-5-)-methyltransferase 3 alpha
DPYSL2	dihydropyrimidinase like 2
DSCAM	Down syndrome cell adhesion molecule
DYNC1H1	dynein cytoplasmic 1 heavy chain 1
DYRK1A	Dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 1A
EBF3	early B-cell factor 3
EHMT1	Euchromatic histone-lysine N-methyltransferase 1
EIF3F	eukaryotic translation initiation factor 3 subunit F
EIF3G	eukaryotic translation initiation factor 3 subunit G
ELAVL3	ELAV like neuron-specific RNA binding protein 3
EP300	E1A binding protein p300
ERBIN	erbb2 interacting protein
ETFB	electron transfer flavoprotein subunit beta
FAM98C	family with sequence similarity 98 member C
FBXL13	F-box and leucine rich repeat protein 13
FMR1	fragile X mental retardation 1
FOXG1	Forkhead box G1
FOXP1	forkhead box P1
FOXP2	forkhead box P2
GABRB2	gamma-aminobutyric acid type A receptor subunit beta2

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GABRB3	gamma-aminobutyric acid (GABA) A receptor, beta 3
GFAP	glial fibrillary acidic protein
GIGYF1	GRB10 interacting GYF protein 1
GIGYF2	GRB10 interacting GYF protein 2
GNAI1	G protein subunit alpha i1
GRIA1	glutamate ionotropic receptor AMPA type subunit 1
GRIA2	glutamate ionotropic receptor AMPA type subunit 2
GRIN1	Glutamate receptor, ionotropic, N-methyl D-aspartate 1
GRIN2A	glutamate receptor, ionotropic, N-methyl D-aspartate 2A
GRIN2B	glutamate receptor, inotropic, N-methyl D-apartate 2B
HDLBP	high density lipoprotein binding protein
HECTD4	HECT domain E3 ubiquitin protein ligase 4
HIVEP2	HIVEP zinc finger 2
HNRNPH2	heterogeneous nuclear ribonucleoprotein H2
HNRNPU	heterogeneous nuclear ribonucleoprotein U
HNRNPUL2	Heterogeneous Nuclear Ribonucleoprotein U Like 2
HRAS	v-Ha-ras Harvey rat sarcoma viral oncogene homolog
ILF2	interleukin enhancer binding factor 2
INTS6	integrator complex subunit 6
IQSEC2	IQ motif and Sec7 domain 2
IRF2BPL	Interferon regulatory factor 2 binding protein-like
ITSN1	Intersectin 1
KANSL1	KAT8 regulatory NSL complex subunit 1
KAT2B	lysine acetyltransferase 2B
KATNAL2	Katanin p60 subunit A-like 2
KCNB1	potassium voltage-gated channel subfamily B member 1
KCNMA1	potassium calcium-activated channel subfamily M alpha 1
KCNQ3	potassium voltage-gated channel subfamily Q member 3
KDM2B	lysine demethylase 2B
KDM3B	lysine demethylase 3B
KDM5B	Lysine (K)-specific demethylase 5B
KDM5C	lysine demethylase 5C
KDM6B	Lysine (K)-specific demethylase 6B
KIAA0232	KIAA0232
KLHL20	kelch like family member 20
KMT2A	Lysine (K)-specific methyltransferase 2A
KMT2C	Lysine (K)-specific methyltransferase 2C
KMT2E	Lysine (K)-specific methyltransferase 2E
KMT5B	lysine methyltransferase 5B
KRAS	KRAS proto-oncogene, GTPase
LDB1	LIM domain binding 1
LRRC4C	leucine rich repeat containing 4C
LZTR1	Leucine-zipper-like transcription regulator 1
MAGEL2	MAGE-like 2
MAP1A	microtubule associated protein 1A

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

MAP2K1	mitogen-activated protein kinase kinase 1
MAP2K2	mitogen-activated protein kinase kinase 2
MARK2	Microtubule Affinity Regulating Kinase 2
MBD5	Methyl-CpG binding domain protein 5
MBOAT7	membrane bound O-acyltransferase domain containing 7
MECP2	Methyl CpG binding protein 2
MED13	mediator complex subunit 13
MED13L	Mediator complex subunit 13-like
MEF2C	myocyte enhancer factor 2C
MEIS2	Meis homeobox 2
METTL26	methyltransferase like 26
MFRP	membrane frizzled-related protein
MIB1	MIB E3 ubiquitin protein ligase 1
MKX	mohawk homeobox
MLANA	melan-A
MSL3	MSL complex subunit 3
MTOR	mechanistic target of rapamycin kinase
MYCBP2	MYC binding protein 2
MYO5A	myosin VA
MYT1L	Myelin transcription factor 1-like
NAA15	N(alpha)-acetyltransferase 15, NatA auxiliary subunit
NACC1	nucleus accumbens associated 1
NAV3	Neuron Navigator 3
NBEA	neurobeachin
NCKAP1	NCK-associated protein 1
NCOA1	nuclear receptor coactivator 1
NEXMIF	neurite extension and migration factor
NF1	neurofibromin 1 (neurofibromatosis, von Recklinghausen disease, Watson disease)
NINL	ninein like
NIPBL	Nipped-B homolog (Drosophila)
NLGN2	Neuroligin 2
NLGN3	neuroligin 3
NLGN4X	neuroligin 4, X-linked
NR3C2	Nuclear receptor subfamily 3, group C, member 2
NR4A2	nuclear receptor subfamily 4 group A member 2
NRAS	NRAS proto-oncogene, GTPase
NRXN1	neurexin 1
NRXN2	neurexin 2
NRXN3	neurexin 3
NSD1	nuclear receptor binding SET domain protein 1
NUP155	nucleoporin 155
OR52M1	olfactory receptor family 52 subfamily M member 1
P2RX5	purinergic receptor P2X 5
PACS1	phosphofurin acidic cluster sorting protein 1
PAH	Phenylalanine hydroxylase

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

PAX5	Paired box 5
PCCB	propionyl-CoA carboxylase beta subunit
PCDH19	protocadherin 19
PCM1	pericentriolar material 1
PHF12	PHD finger protein 12
PHF2	PHD finger protein 2
PHF21A	PHD finger protein 21A
PHF3	PHD finger protein 3
PHIP	pleckstrin homology domain interacting protein
PLEKHA8	pleckstrin homology domain containing A8
POGZ	Pogo transposable element with ZNF domain
POMGNT1	protein O-linked mannose N-acetylglucosaminyltransferase 1 (beta 1,2-)
PPP1CB	protein phosphatase 1 catalytic subunit beta
PPP1R9B	protein phosphatase 1 regulatory subunit 9B
PPP2R5D	Protein phosphatase 2, regulatory subunit B', delta
PPP5C	protein phosphatase 5 catalytic subunit
PRKAR1B	protein kinase cAMP-dependent type I regulatory subunit beta
PRR12	proline rich 12
PRR14L	proline rich 14 like
PSMD11	proteasome 26S subunit, non-ATPase 11
PSMD12	proteasome 26S subunit, non-ATPase 12
PSMD6	proteasome 26S subunit, non-ATPase 6
PTCHD1	patched domain containing 1
PTEN	phosphatase and tensin homolog (mutated in multiple advanced cancers 1)
PTK7	Protein tyrosine kinase 7 (inactive)
PTPN11	protein tyrosine phosphatase, non-receptor type 11
RAF1	Raf-1 proto-oncogene, serine/threonine kinase
RAI1	retinoic acid induced 1
RALGAPB	Ral GTPase activating protein non-catalytic beta subunit
RANBP17	proline rich 25
RAPGEF4	RAN binding protein 17
RELN	Reelin
RERE	Arginine-glutamic acid dipeptide (RE) repeats
RFX3	regulatory factor X3
RIMS1	Regulating synaptic membrane exocytosis 1
RIT1	Ras like without CAAX 1
RORB	RAR related orphan receptor B
RUNX1T1	RUNX1 partner transcriptional co-repressor 1
SATB1	SATB homeobox 1
SCAF1	SR-Related CTD Associated Factor 1
SCN1A	sodium channel, voltage-gated, type I, alpha subunit
SCN2A	sodium channel, voltage-gated, type II, alpha subunit
SCN8A	sodium channel, voltage gated, type VIII, alpha subunit
SETBP1	SET binding protein 1
SETD1A	SET domain containing 1A, histone lysine methyltransferase

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

SETD2	SET domain containing 2
SETD5	SET domain containing 5
SHANK2	SH3 and multiple ankyrin repeat domains 2
SHANK3	SH3 and multiple ankyrin repeat domains 3
SHOC2	SHOC2 leucine rich repeat scaffold protein
SIN3A	SIN3 transcription regulator family member A
SKI	SKIproto-oncogene
SLC6A1	Solute carrier family 6 (neurotransmitter transporter), member 1
SLC9A6	solute carrier family 9 (sodium/hydrogen exchanger), member 6
SLFN5	Rap guanine nucleotide exchange factor 4
SMARCA2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2
SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4
SMARCC2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 2
SMURF1	schlafen family member 5
SMURF1	SMAD specific E3 ubiquitin protein ligase 1
SNCAIP	synuclein alpha interacting protein
SON	SONDNA binding protein
SOS1	SOS Ras/Rac guanine nucleotide exchange factor 1
SOS2	SOS Ras/Rho guanine nucleotide exchange factor 2
SOX5	SRY-box 5
SPAST	Spastin
SPTBN1	Spectrin Beta, Non-Erythrocytic 1
SRCAP	Snf2 related CREBBP activator protein
SRPRA	SRP receptor subunit alpha
STXBP1	Syntaxin binding protein 1
SYN1	Synapsin 1
SYNGAP1	synaptic Ras GTPase activating protein 1
TANC2	etratricopeptide repeat, ankyrin repeat and coiled-coil containing 2
TAOK1	TAO kinase 1
TBCEL	tubulin folding cofactor E like
TBCK	TBC1 domain containing kinase
TBL1XR1	transducin beta like 1 X-linked receptor 1
TBR1	T-box, brain, 1
TCF20	Transcription factor 20 (AR1)
TCF4	Transcription factor 4
TCF7L2	Transcription factor 7-like 2 (T-cell specific, HMG-box)
TEK	TEKreceptortyrosine kinase
TGM1	transglutaminase 1
TLE3	TLE family member 3, transcriptional corepressor
TLK2	tousled-like kinase 2
TM9SF4	transmembrane 9 superfamily member 4
TMEM39B	transmembrane protein 39B
TNRC6B	trinucleotide repeat containing adaptor 6B

TRAF7	TNF receptor associated factor 7
TRIM23	tripartite motif containing 23
TRIO	Trio Rho guanine nucleotide exchange factor
TRIP12	Thyroid hormone receptor interactor 12
TSC1	tuberous sclerosis 1
TSC2	tuberous sclerosis 2
TSHZ1	teashirt zinc finger homeobox 1
TSHZ3	teashirt zinc finger homeobox 3
TSPAN4	tetraspanin 4
UBAP2L	ubiquitin associated protein 2 like
UBE3A	ubiquitin protein ligase E3A
UBR1	ubiquitin protein ligase E3 component n-recognin 1
UIMC1	ubiquitin interaction motif containing 1
UPF3B	UPF3B, regulator of nonsense mediated mRNA decay
USP45	ubiquitin specific peptidase 45
USP9X	ubiquitin specific peptidase 9 X-linked
VEZF1	vascular endothelial zinc finger 1
VPS13B	vacuolar protein sorting 13 homolog B (yeast)
VPS54	VPS54 subunit of GARP complex
WAC	WW domain containing adaptor with coiled-coil
WDFY3	WD repeat and FYVE domain containing 3
YY1	YY1transcription factor
ZBTB20	Zinc finger and BTB domain containing 20
ZBTB21	zinc finger and BTB domain containing 21
ZMYND8	zinc finger MYND-type containing 8
ZNF292	zinc finger protein 292
ZNF462	Zinc finger protein 462
ZNF559	zinc finger protein 559

10.2.10 Supplementary Table 10. Inherited variants prioritized by the presence of the assigned PLARV gene into our custom list of ASD risk genes

10.3 VQSR IN DETAIL

10.3.1 Resources to build the recalibration model

The VQSR model metrics were created from a training set of validated resources: dbSNP build 138, HapMap 3.3, 1000 Genomes OMNI 2.5, and 1000 Genomes Phase 1.

As reliable reference data (referred to as truth resources), consisting of call sets validated to a very high degree of confidence, HapMap 3.3 and 1000 Genomes OMNI 2.5 were used for SNPs, while 1000 Genomes Phase 1 was utilized for indels, all with the parameter truth = true (*i.e*. the model will consider that the variants in these resources are representative of true sites). These were employed for training the model and establishing a threshold to filter variants based on their sensitivity to truth sites. Priors, denoting phred-scaled estimates of data accuracy, were set at Q15 (equivalent to a 96.84% prior likelihood assigned to these variants) for HapMap SNPs, and Q12 (93.69% prior likelihood) for SNPs in 1000 Genomes OMNI 2.5, as well as for indels in 1000 Genomes Phase 1.

As for the training resources (*i.e.* validated to a certain degree of confidence and potentially containing false positives along with true variants), these were utilized for training the recalibration model (training = true). The same resources and priors were applied for training, and 1000 Genomes Phase 1 was specifically employed as an additional training resource for SNPs, with a prior of Q10 (90% prior likelihood).

For the known sites resources, corresponding to call sets not validated to a high degree (truth = false) of confidence and excluded from the recalibration model training (training = false), these were used to stratify output metrics such as the Ti/Tv ratio based on the presence or absence of variants in this resource (*i.e.* known resources are not used for any calculations, they are only valuable for reporting purposes). dbSNP build 138 was employed for SNPs, whereas the set of known indels in the Homo sapiens assembly 38 was used for indels (priors were set to Q2, corresponding to a 36.90% prior likelihood). Of note, this recommended low prior value reflects the GATK callers' rationale, which consists of creating an extensive and highly sensitive call set (*i.e.* not discarding any true positive but including false ones) that necessitates substantial refinement through supplementary filtering processes, as we will be discussing next.

10.3.2 Covariates employed for variant recalibration

Next, we specified which annotations should be included in the model in order to evaluate the likelihood of SNPs and indels being real.

For SNPs, we employed the following annotations:

i. QualByDepth (QD): This assesses variant confidence by normalizing the QUAL score for each variant with respect to allele depth.
ii. FisherStrand (FS): This measures strand bias, indicating whether there is a bias between forward and reverse strands for the reference or alternate allele. Increased bias suggests potential false positive calls.
iii. StrandOddsRatio (SOR): Complementary to FS, SOR is estimated using the Symmetric Odds Ratio test instead of Fisher's Exact Test. This method is better suited for handling large amounts of data in high-coverage situations.
iv. MappingQualityRankSumTest (MQRankSum): This test compares the mapping qualities of reads supporting the reference/alternate allele. Values close to zero suggest little to no difference, while a statistically significant difference may indicate bias or artifacts in the sequencing and/or mapping process.
v. ReadPosRankSumTest (ReadPosRankSum): This test examines whether there is evidence of bias in the position of reference/alternate alleles within supporting reads. Similar to MQRankSum, values close to zero indicate no differences in allele positions relative to the ends of reads. Negative values suggest that the alternate allele is more frequently found at the ends of reads, which is indicative of error.
vi. RMSMappingQuality (MQ): This provides an estimation of the overall mapping quality of reads supporting a variant call, averaged over all samples in a cohort.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Annotations identical to those used for SNPs were applied to indels, excluding MQ, which is generally not recommended for indel calibration. The reason for this is related to the nature of the MQ annotation and its suitability for accurately assessing the quality of indel variants. MQ is a measure of the likelihood that a read is mapped to its reported genomic location correctly. For SNPs, MQ can be a useful indicator of the reliability of the variant call. However, for indels, the situation is more complex. Indels, especially longer ones, may involve misalignments or shifts in the alignment of reads. Thus, the standard MQ may not effectively capture these complexities for indels and using it could lead to inaccurate recalibration and filtering of indel variants. Instead of MQ, VQSR typically relies on other annotations that are more informative for indels, such as ReadPosRankSum, FS, and SOR.

Moreover, an additional parameter, the maximum number of Gaussians, was set to 4, as recommended for indel recalibration. The VQSR method uses Gaussian mixture models to infer variant quality scores for true variants/false positive artifacts. The maximum number of Gaussians determines how many individual Gaussian distributions (*i.e.* symmetrical bell-shaped curves that represent the distribution of quality scores for clusters of variants that have similar properties) will be used to represent each of the components of the overall distribution of variant quality scores, aiming to capture the complexity of the data.

Of note, the unfiltered depth of coverage (DP), which may be tremendously informative in sequencing analysis such as WGS, was not used in our analysis for indels nor for SNPs. Because of the nature of our analysis, which focuses on specific regions of the genome through targeted sequencing, the DP is expected to vary across the interrogated regions (some regions may have high coverage while others do not, due to capture efficiency issues), and thus, will not be a reliable indicator of the quality of variant calls.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

10.4 FAVORABLE REPORTS FROM THE ETHICS COMMITTEE

The subsequent pages encompass reproductions of the favorable reports issued by the Ethics Committee pertaining to the investigations that led to the research developed in chapters 2 and 3

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

XUNTA DE GALICIA
CONSELLERÍA DE SANIDADE
Secretaría Xeral

Comité Ético de Investigación Clínica de Galicia
Edificio Administrativo de San Lázaro
15781 SANTIAGO DE COMPOSTELA
Tlf: 881 545425 Fax: 881 541804
ceic@sergas.es

galicia

DICTAMEN DEL COMITÉ ÉTICO DE INVESTIGACIÓN CLÍNICA DE GALICIA

Paula M. López Vázquez, Secretaria del Comité Ético de Investigación Clínica de Galicia

CERTIFICA:

Que este Comité evaluó en su reunión del día 28/06/2012 el estudio:

Título: Contribución a la búsqueda de las causas genéticas de los Trastornos del espectro autista con retraso mental
Promotor: Ángel Carracedo Álvarez
Código de Registro CEIC de Galicia: 2012/098

Y que este Comité de conformidad con sus Procedimientos Normalizados de Trabajo y tomando en cuenta los requisitos éticos, metodológicos y legales exigibles a los estudios de investigación con seres humanos, sus muestras o registros, emite un **DICTAMEN FAVORABLE** al estudio propuesto y que se llevará a cabo en:

Centros	Investigadores principales
Fundación Pública Galega de Medicina Xenómica	Ángel Carracedo Álvarez

En Santiago de Compostela a 11 de julio de 2012
La Secretaria

Paula M. López Vázquez

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

XUNTA DE GALICIA
CONSELLERÍA DE SANIDADE
Secretaría Xeral Técnica

SERVIZO GALEGO de SAÚDE | Xerencia do Servizo Galego de Saúde

galicia

DICTAMEN DEL COMITÉ DE ÉTICA DE LA INVESTIGACIÓN DE SANTIAGO-LUGO

Ana Estany Gestal, Secretaria del Comité de Ética de la Investigación de Santiago-Lugo,

CERTIFICA:

Que este Comité evaluó en su reunión del día 19 de enero de 2021 el estudio**:**

> **Título:** Enfoque integral y traslacional del diagnóstico y tratamiento de trastornos del neurodesarrollo (TEA,TDAH, DI) y del TOC: Una aproximación genética y funcional.
> **Versión:** 3.0
> **Promotor/a**: Ángel Carracedo Álvarez
> **Código de Registro**: 2020/400

Y que este Comité, tomando en consideración la pertinencia del estudio, el conocimiento disponible, los requisitos legales aplicables y los Procedimientos Normalizados de Trabajo del Comité, emite un dictamen **FAVORABLE** para la realización del citado estudio.

Documento asinado dixitalmente por:
Ana Estany Gestal (27/01/2021 14:06)
https://sede.xunta.gal/cve?idcve=5A0C-C4C4-B0AH-0F4O-TG3H-L7YL-1SYM-W161-1752-8080-09

NOTA: Se le recuerda que en el caso de que en este estudio se recluten pacientes, el equipo investigador debe tener disponible el Documento de Consentimiento Informado (Hojas de Información y Hojas de Firma) tanto en *galego* como en castellano en el momento de comenzar el reclutamiento.

Xacobeo 2021

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

LIST OF TABLES

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

11 LIST OF TABLES

Table 1. Diagnostic criteria for ASD according to the DSM-5. 14
Table 2. Severity levels for ASD according to the DSM-5. 14
Table 3. Most common comorbidities in ASD. 16
Table 4. Main ASD consortia. 24
Table 5. Key supporting evidence for the contribution of *de novo*/inherited variants (SNVs and indels) to the genetic etiology of ASD. 27
Table 6. Known CNVs in ASD curated from SFARI Gene. 31
Table 7. Genome-wide significant loci from the largest ASD GWAS. 33
Table 8. Key supporting evidence for the contribution of non-coding variation to the genetic etiology of ASD. 52
Table 9. Characterization of the cohorts included in the colocalization analysis. 66
Table 10. ASD associated genes collected from bibliography and selected for eQTL colocalization study 67
Table 11. eQTL colocalization analysis results for the 27 LOI in ASD. 72
Table 12. Colocalization results for potentially causal genes selected from the ASD eQTL colocalization analysis in correlated NDDs. 73
Table 13. Functional enrichment for *MAPT, NKX2-2, PTPRE* and their interactors in brain tissue 78
Table 14. Complementary datasets included in gene validation and GO analysis. 89
Table 15. Variants with an AL/DL $\Delta \geq 0.8$ and AG/DG $\Delta \geq 0.5$ 94
Table 16. Variants with AG/AL/DG/DL $\Delta \geq 0.8$. 94
Table 17. ABSplice prediction. 96
Table 18. Common/unique significantly enriched BP in splicing-related gene lists and the non-splicing gene list. 101
Table 19. Sequence classes defined in Sei. 117
Table 20. Enrichment of cCREs in *de novo* and inherited variants. 123
Table 21. Enrichment of variant scores within each of the regulatory classes, divided in positive and negative effects. 125

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

LIST OF FIGURES

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

12 LIST OF FIGURES

Figure 1. Global ASD prevalence in 2019. ...15
Figure 2. Diagram of the interaction between genetic and environmental risk factors in the development of ASD..18
Figure 3. AE model for ASD variance. 23
Figure 4. Frequency of genetic variants and disease susceptibility..32
Figure 5. Non-coding regulatory elements affecting gene expression. ...37
Figure 6. Examples of non-coding ASD-associated variants affecting proper functioning of UTRs, promoters, and enhancers. 39
Figure 7. Examples of the impact of miRNAs and lncRNAs on ASD.-42
Figure 8. Examples of non-coding ASD-associated variants interfering tandem repeats and variants affecting proper functioning of the splicing process45
Figure 9. Strategy for the characterization of non-coding variants47
Figure 10. Biological pathways frequently associated with ASD..57
Figure 11. Comparison of GWAS sample sizes and discovery rates..63
Figure 12. Manhattan plot for the latest ASD GWAS meta-analysis63
Figure 13. eQTL colocalization analysis...65
Figure 14. eQTpLot for ASD and MAPT multi-tissue analysis....74
Figure 15. eQTpLot for ASD and PTPRE multi-tissue analysis...75
Figure 16. Expression and disease enrichment analysis77
Figure 17. SpliceAI interpretation.87
Figure 18. Workflow for *in silico* variant prediction, variant validation, and gene validation..92
Figure 19. Tissue specific expression of genes harboring 31 *in silico* validated variants. ...98
Figure 20. Top 10 enriched biological processes in the splicing gene-list100
Figure 21. Cluster enrichment analysis for genes harboring splicing *vs.* coding variants (splicing gene list *vs.* non-splicing)102
Figure 22. cCREs' classification in the ENCODE Encyclopedia. 110
Figure 23. Tranches plot generated by the VQSR. 112
Figure 24. Workflow for variant quality control (QC) and filtering, and variant/gene prioritization...113
Figure 25. Sei, a deep convolutional network sequence model to globally assess the regulatory landscape of the genome. 118
Figure 26. Long range chromatin interactions. 120
Figure 27. Pie chart depicting the classification of cCREs. 122
Figure 28. Pie chart representation of *de novo* and inherited variants based on cCREs' classification. 123
Figure 29. Counts of *de novo* variants assigned to each regulatory class...124
Figure 30. Counts of inherited variants assigned to each regulatory class. ...125
Figure 31. Venn diagram for *de novo* PARV and PLARV genes. ...127
Figure 32. UCSC browser of ATAC-seq, H3K27ac and H3K4me3 ChIP-seq and PLAC-seq loops at the ETFBKMT and AMN1 loci...128
Figure 33. Venn diagram for inherited PARV and PLARV genes..129
Figure 34. Biological processes significantly enriched within PLARV genes from the inherited variant list across the three cell types...130

PUBLICATIONS AND CONTRIBUTIONS

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

13 PUBLICATIONS AND CONTRIBUTIONS

As indicated in the Introduction, a relatively significant amount of data was already published in the *European* Journal *of Medical Genetics*, published in June 2023, where I am first author.

Dominguez-Alonso S[1], Carracedo A[2], Rodriguez-Fontenla C[3]. The non-coding genome in Autism Spectrum Disorders. Eur J Med Genet. 2023;66(6):104752. doi: 10.1016/j.ejmg.2023.104752. Epub 2023 Apr 5. PMID: 37023975.

1. Grupo de Medicina Xenómica, Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidad de Santiago de Compostela, Santiago de Compostela, Spain.

2. Grupo de Medicina Xenómica, Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidad de Santiago de Compostela, Santiago de Compostela, Spain; Grupo de Medicina Xenómica, Fundación Instituto de Investigación Sanitaria de Santiago de Compostela (FIDIS), Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidad de Santiago de Compostela, Santiago de Compostela, Spain.

3. Grupo de Medicina Xenómica, Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidad de Santiago de Compostela, Santiago de Compostela, Spain.

However, contents along the different sections were partially rearranged, rewritten and extended to include additional information and details that were not incorporate into the original manuscript.

The introduction figures are original and created with Adobe Illustrator, BioRender, Rstudio or Microsoft PowerPoint softwares, and they are not subjected to copyright.

The official statement that has been provided by this journal to preserve its copyrights is depicted in the “Declaración de autoría de los datos presentados”.

Additional information includes:

i. Funding: Fundación María José Jove supported this work. SDA was supported by a Xunta de Galicia predoctoral fellowship. CRF was supported by a contract from the REACT-EU.

ii. Credit authorship contribution statement: S. Dominguez-Alonso: wrote the manuscript and did the review work, participated in the design of this review. A. Carracedo: participated in the design of this review, critically revised the work and approved the final content. C. Rodriguez-Fontenla: participated in the design of this review, critically revised the work and approved the final content.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

iii. Declaration of competing interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
iv. Acknowledgements: We thank the project: Instituto de Salud Carlos III (ISCIII)/PI1900809/Cofinanciado FEDER.
v. Publisher: Elsevier
vi. ISSN: 18780849, 17697212
vii. Category: Biochemistry, Genetics and Molecular Biology. Genetics (quartile Q2).
viii. H-index 2022: 60

Moreover, as indicated in the Chapter 1, a major amount of data was already published in the *European* Journal *of Medical Genetics,* published in October 2023, where I am first author.

Part of the content of this chapter has been published as Dominguez-Alonso S[1], Carracedo A[2], Rodriguez-Fontenla C[3]. eQTL colocalization analysis highlights novel susceptibility genes in Autism Spectrum Disorders (ASD). Transl Psychiatry. 2023;13(1):1-9. doi:10.1038/s41398-023-02621-0

1. Grupo de Medicina Xenómica, Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidad de Santiago de Compostela, Santiago de Compostela, Spain.
2. Grupo de Medicina Xenómica, Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidad de Santiago de Compostela, Santiago de Compostela, Spain; Grupo de Medicina Xenómica, Fundación Instituto de Investigación Sanitaria de Santiago de Compostela (FIDIS), Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidad de Santiago de Compostela, Santiago de Compostela, Spain.
3. Grupo de Medicina Xenómica, Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidad de Santiago de Compostela, Santiago de Compostela, Spain.

However, contents along the different sections were partially rewritten and extended to include additional information and details that were not incorporated into the original manuscript.

The figures are original and created with Adobe Illustrator, Rstudio, GTEx or Microsoft PowerPoint softwares, and they are not subjected to copyright.

The official statement that has been provided by this journal to preserve its copyrights is depicted in the “Declaración de autoría de los datos presentados”.

Additional information includes:

i. Funding: Fundación María José Jove supported this work. SDA was supported by a Xunta de Galicia predoctoral fellowship. CRF was supported by a contract from the REACT-EU

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ii. Credit authorship contribution statement: S. Dominguez-Alonso: wrote the manuscript. A. Carracedo and C. Rodriguez-Fontenla critically revised the work and approved the final content. All authors participated in the design and coordination of this study.
iii. Publisher: Nature Publishing Group
iv. ISSN: 21583188
v. Category: Neuroscience, Biological Psychiatry (quartile Q1).
vi. H-index 2022: 113

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Through a combination of bioinformatics methods and diverse genetic data collections, we have uncovered insights into a broad spectrum of genetic variation. In summary, the findings from this in silico characterization of common, splicing, and regulatory variants include the identification of specific genes most likely to be affected by non-coding common variation, a deeper understanding of how splicing variation may mediate disease risk, and the establishment of a crucial role for chromatin conformation in ASD.