GENOTYPE AND PHENOTYPE RELATIONSHIPS IN NEURODEVELOPMENTAL DISORDERS



Katarina Vrcelj

Exeter College

Department of Physiology, Anatomy and Genetics

University of Oxford

A thesis submitted in partial fulfillment of the requirements

for the degree of *Doctor of Philosophy*

Trinity Term 2017

Genotype and Phenotype Relationships in Neurodevelopmental Disorders

Katarina Vrcelj, Exeter College, University of Oxford

DPhil in Physiology, Anatomy and Genetics

Trinity Term 2017

Abstract

NeuroDevelopmental Disorders (NDDs) are a group of heterogeneous neuropsychiatric conditions, encompassing Developmental Delay (DD), Intellectual Disability (ID), Autism Spectrum Disorder (ASD) and Gilles de la Tourette Syndrome (GTS). NDDs often involve a life-long need of supportive services and represent a major economic and societal burden in Europe and the United States. Hitherto, no safe and effective treatment strategies are available, underpinning the urgency of deciphering their elusive aetiology.

NDDs have a strong genetic component. Accordingly, the mechanistic understanding of disease involves (1) the description of mutational landscapes, (2) the molecular and cellular of characterisation of genetic variants and (3) their integration at circuit and system levels. In this thesis, we tackled these theoretical underpinnings to provide insights into the genetics of NDDs and the onset of associated clinical abnormalities.

Firstly, aiming to shape the genetic architecture of GTS, a condition characterised by the presence of motor and vocal tics, we undertook a whole-genome CNV study of a cohort of Danish GTS cases and healthy controls. Using statistical and functional genomics approaches, we proposed novel potential candidate genes and implicated the disruption of early neurodevelopmental and late synaptic processes in the aetiology of GTS.

Secondly, to elucidate the vast phenotypic heterogeneity of NDDs, we conducted a systematic investigation of genotype and phenotype relationships in DD, ID and ASD. Taking advantage of extensive phenotypic and genetic data available for DD/ID and ASD patients, we grouped individuals based on their functional rare CNV and gene disruptions but did not to identify distinguishing clinical archetypes. Instead, we showed converging molecular perturbations underlie the onset of globally more similar clinical presentations and investigated the role of common variants in modulating their expressivity.

Lastly, we established the relevance of mouse models in the study of human disease. By applying comprehensive genomics approaches to over 1,000 mouse neuroanatomical knockouts, we implicated early neurodevelopmental and adult synaptic processes in the aetiology of ID and brain malformations. Furthermore, we showed that functionally converging genetic disturbances translate at the phenotypic level and proposed novel candidate ID genes.

Declaration

I hereby declare that the work presented in this thesis is my own, and has been performed by myself, unless stated otherwise. The analyses were conducted in the MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, UK. This thesis has not been submitted for any other degree or professional qualification.

The following parts of this thesis were carried out in collaboration:

1. The Gilles de la Tourette syndrome project was performed in collaboration with Zeynep Tümer, Kennedy Center, Department of Clinical Genetics, The Juliane Marie Centre, Copenhagen University Hospital, Denmark. Shanmukha Sampath Padmanabhuni, Department of Molecuar Biology and Genetics Democritus, University of Thrace, Alexandroupoli, Greece, processed genotyping data.

2. The neuroanatomical mouse phenotyping project was performed in collaboration with Binnaz Yalcin, Institute of Genetics and Molecular and Cellular Biology (IGBMC), University of Strasbourg, France. Anna Mikhaleva, Centre for Integrative Genomics, University of Lausanne, Switzerland, collated the mouse neuroanatomical dataset and applied mixed models to identify genes inducing significant brain defects.

Katarina Vrcelj

Acknowledgements

I would like to thank my supervisor, Caleb Webber for taking me on to join his lab, providing me with invaluable guidance and teaching me the scientific approach to problem solving. I am also grateful to the other members from the Webber group for their support and feedback. In particular, I would like to express my gratitude to Cynthia Sandor for continuous discussions, advice and help with (and beyond) data analysis and scripting.

My thanks also go to Zeynep Tümer for seeking our collaboration for the Gilles de la Tourette CNV study and Shanmukha Sampath Padmanabhuni for processing the genotyping data. Similarly, I am obligated to Binnaz Yalcin for taking us onboard the high throughout mouse phenotyping project and Anna Mikhaleva for generating the dataset.

Finally, I am grateful to my family and friends for their support.

List of Abbreviations

ADHD	Attention deficit hyperactivity disorder
ASD	Autism spectrum disorders
BH	Benjamini-Hochberg correction for multiple testing
CA	Congenital abnormalities
CDS	Coding DNA sequence length
CNV	Copy number variation
CT	Control
CV	Common variation
DECIPHER	Database of genomic variation and phenotype in humans using Ensembl resources
DD	Developmental delay
dn	*de novo*
DNA	Deoxyribonucleic acide
FMRP	Fragile X Mental Retardation protein
FDR	False discovery rate
GBA	*Guilt by association* principle
GTEx	Genotype Tissue Expression project
GO	Gene Ontology
GTS	Gilles de la Tourette syndrome
GWAS	Genome-wide association study
HIS	Haploinsufficiency
HPO	Human Phenotype Ontology
ID	Intellectual disability
Inh	Inherited
IQ	Intellectual quotient
LGD	Likely gene-disrupting
LoF	Loss of function
Kb	Kilobase
KO	Knockout
Mb	Megabase

MGI	Mouse Genome Informatics Database
MIM	Phenotype from the Mendelian Inheritance in Man (accessed from the Online MIM data source; OMIM)
MP	Mouse phenotype
NDD	Neurodevelopmental disorder
NGS	Next generation sequencing
Nijmegen	NDD cohort of patients with DD, ID and CAs, collected at the Radboud University, Nijmegen Medical Centre, Netherlands
OCD	Obsessive compulsive disorder
OR	Odds ratio
p	P-value
PLN	Phenotypic linkage network
PPI	Protein-protein interaction
PSD	Post-synaptic density
RNA	Ribonucleic acid
QC	Quality control
r	Correlation coefficient
RV	Rare variation
SSC	Simons Simplex Collection
SLI	Specific language impairment
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variation
TF	Transcription factor
WES	Whole-exome sequencing
WT	Wildtype

This list includes the primary abbreviations used throughout the thesis, which are restated in each chapter. Chapter-specific abbreviations are also used.

Table of Contents

Abstract ii

Declaration iv

Acknowledgements v

List of Abbreviations vi

List of Figures xiv

List of Tables xvii

CHAPTER 1 | The genetic landscape of neurodevelopmental disorders: an introduction 1

1.1 Genetic models of NDDs 9
1.1.1 Genetic models of NDDs: monogenic versus polygenic models 9
1.1.2 Frequency of alleles: common versus rare variants 12
1.2 The genetic architecture of NDDs 19
1.2.1 GWA studies of NDDs 19
1.2.2 CNV studies of NNDs 24
1.2.3 Gene studies of NDDs 27
1.2.4 Integrative models of NDDs 29
1.2.5 Conclusions 32
1.3 Biological insights from gene, cellular and animal studies 33
1.3.1 Neurodevelopment and synaptic function 33
1.3.2 NDD genes affect neurodevelopmental and synaptic processes 36
1.3.3 Cellular and animal models 41
1.3.3 Conclusion 47
1.4 Comprehensive approaches to the study NDDs 47
1.4.1 Annotation datasets and functional enrichments 48
1.4.2 Transcriptomic analyses of NDDs 50
1.4.3 Protein-protein interaction 55
1.4.4 Integrative approaches 57
1.4.5 Conclusion 61
1.5 Genotype and phenotype relationships in NDDs 62

1.6 Conclusions and thesis aims 66

CHAPTER 2 | Common materials and methods 69

2.1 Gene lists 70
2.1.1 Gene haploinsufficiency scores 70
2.1.2 Human functional gene lists 70
2.1.3 High confidence disease gene lists 71
2.1.4 Genes and phenotypes from patients with NDDs 73
2.2 Functional gene annotation 74
2.2.1 GO annotations 74
2.2.2 MGI annotations 75
2.3 Expression datasets 75
2.3.1 Body-wide gene expression map 75
2.3.2 Spatiotemporal brain expression map 76
2.4 Gene networks 78
2.4.1 GTEx and Brainspan co-expression networks 79
2.4.2 Protein-protein interaction networks 79
2.4.3 Phenotypic linkage network 80
2.5 Network analysis 82
2.5.1 Gene matching 82
2.5.2 Gene set clustering 83
2.5.3 Clustering two gene sets 84
2.5.4 Gene module identification 85
2.6 Phenotypic similarity 87
2.7 Statistical tests 88
2.7.1 Gene set enrichments and comparisons 88
2.7.2 Gene expression 89
2.7.3 Phenotypic similarity comparisons 90

CHAPTER 3 | Gilles de la Tourette syndrome: a CNV study 91

3.1 Introduction 92
3.1.1 Chromosomal aberrations 93
3.1.2 Linkage studies 95
3.1.3 Candidate gene studies 96
3.1.4 Genome-wide association studies 97
3.1.5 Copy number variant studies 98
3.1.6 Whole-exome sequencing studies 99
3.1.7 Conclusion 100
3.2 Chapter aims 101
3.3 Specific Materials and Methods 102

3.3.1 Study Subjects 102
3.3.2 Genotyping 103
3.3.3 Genotype data pre-processing 103
3.3.4 Per individual quality control 104
3.3.5 Per SNP quality control 107
3.3.6 CNV calling and quality control 107
3.3.7 Rare CNVs and platform bias check 110
3.3.8 CNV burden and association analyses 112
3.2.9 GTS CNVs from Fernandez *et al.*, 2012 113
3.4 Results 114
3.4.1 Summary of data 114
3.4.2 Recurrence and disease gene analysis 114
3.4.3 Functional pathway and expression analyses 117
3.4.4 Subset of genes more likely to be implicated in GTS aetiology 119
3.5 Discussion 129
3.6 Conclusion 135

CHAPTER 4 | Genotype and phenotype convergence: a CNV study . 136

4.1 Introduction 137
4.2 Chapter aims 143
4.3 Specific materials and methods 145
4.3.1 Nijmegen cohort: patients with ID, DD and/or CAs 145
4.3.2 DECIPHER cohort: patients with chromosomal aberrations 146
4.3.3 SSC cohort of ASD patients: phenotypic data 147
4.3.4 SSC cohort of ASD patients: CNV data 150
4.3.5 Construction of patient modules 150
4.3.6 Global genotype and phenotype convergence 152
4.3.7 Specificity of phenotypic coherence 153
4.3.8 Identifying phenotypes driving phenotypic coherence 153
4.3.9 Investigating phenotypic relationships between individuals . 154
4.4 Results 157
4.4.1 Summary of data 157
4.4.2 Coherence of neurodevelopmental CNV carriers 160
4.4.3 Identification of modules of co-expressed genes 160
4.4.4 Coherence of patients with similar CNV disruptions 163
4.4.5 Replicating coherence using auxiliary genomics datasets ... 168
4.4.6 Replicating coherence using patients from DECIPHER 171
4.4.7 Coherence of patients with inherited CNVs 174
4.4.8 Phenotypic coherence of ASD patients 179
4.5 Discussion 186

4.6 Conclusion 193

CHAPTER 5 | Genotype and phenotype convergence: an exome study 194

5.1 Introduction 195
5.2 Chapter aims 200
5.3 Specific materials and methods 201
5.3.1. Gene Lists 201
5.3.2 SSC cohort: phenotypic data 201
5.3.3 SSC cohort: rare *de novo* exonic variants 201
5.3.4 SSC cohort: common variants 202
5.3.4.1 Samples and genotyping 202
5.3.4.2 Per individual quality control 203
5.3.4.3 Per SNP quality control 206
5.3.4.4 Genotype imputations 207
5.3.4.4 Genetic correlation 207
5.3.5 Expression quantitative trait loci 208
5.3.6 SNP and eQTL sets 210
5.3.7 Genotype and phenotype analysis pipeline 210
5.3.8 Genetic correlation comparisons 211
5.3.9 Centre of assessment bias 213
5.4 Results 215
5.4.1 Summary of data 215
5.4.2 Phenotypic coherence of single gene variants 216
5.4.3 Modules of co-expressed and functionally similar genes 217
5.4.4 Phenotypic coherence of biologically related perturbations 220
5.4.5 Autoimmunity and ASD 222
5.4.6 Phenotypically coherent modules underlie wide phenotypic spectra 227
5.4.7 Phenotypic coherence supports immune and non-immune ASD aetiologies 230
5.4.8 CVs as genetic modifiers of phenotypic variability 231
5.5 Discussion 241
5.6 Conclusions 249

CHAPTER 6 | Mouse neuroanatomical models: a brain functional genomics study 250

6.1 Introduction 251
6.2 Chapter aims 258
6.3 Specific materials and methods 259

6.3.1 Study samples and datasets 259
6.3.2 Experimental flow and statistical analysis 261
6.3.3 Data imputation and imputed datasets..... 263
6.3.4 Data availability 263
6.3.5 Mouse gene KO dataset 264
6.3.6 Genes causing strongest and weakest neuroanatomical defects..... 265
6.3.7 Phenotypic similarity 266
6.3.8 Human gene intolerance scores and mouse lethal genes ... 267
6.3.9 Mouse gene expression datasets 268
6.3.10 Gene expression levels for Bregma +0.98mm and -1.34 mm 269
6.3.11 Mouse gene networks 270
6.4 Results 271
6.4.1 Summary of data 271
6.4.2 Functional characterisation of mouse NAP genes 274
6.4.3 Functional characterisation of human NAP genes 279
6.4.4 Neuroanatomical functional gene modules 282
6.4.5 Functional translation between mouse and human..... 288
6.4.6 Neuroanatomical deficits and gene deleteriousness..... 290
6.4.7 NAP genes and human disease..... 291
6.5 Discussion..... 299
6.6 Conclusion 306

CHAPTER 7 | From genomics to the clinic: Conclusions and perspectives 307

BIBLIOGRAPHY 318

SUPPLEMENTARY INFORMATION 368

S1 Overview of genes implicated in rare neurological syndromes and NDDs..... 369
S1.1 Genes involved in brain malformations and NDDs 369
S1.1.1 Microcephaly..... 369
S1.1.2 Lissencephaly 371
S1.1.3 Broader NDD categories..... 373
S1.2 Synaptic genes in NDDs..... 374
S1.2.1 Cell adhesion molecules..... 374
S1.2.2 Scaffolding proteins 375
S1.2.3 Ion channels and receptors 377

S1.2.4 Signal transduction pathways 378
S1.2.5 Transcription factors and chromatin regulators 379
S2 Construction of the phenotypic linkage network (PLN)..... 381
S.3 Phenotypic SSC dataset..... 384
S.3.1 Construction of the phenotypic dataset 384
S3.2 Imputation of the phenotypic dataset..... 392

SUPPLEMENTARY FIGURES 394

SUPPLEMENTARY TABLES 426

List of Figures

Fig 1.1 | Genetic models of NDDs 10
Fig 1.2 | Mechanisms of pathogenicity 12
Fig 1.3 | Variant frequency and penetrance in disease 14
Fig 1.4 | Genotype-first approach 65

Fig 2.1 | Phenotypic linkage network (PLN) construction 81
Fig 2.2 | Clustering gene sets 84
Fig 2.3 | Module detection 87

Fig 3.1 | Ancestry clustering 106
Fig 3.2 | CNV calls and platform bias 112
Fig 3.3 | GO enrichment analysis 118
Fig 3.4 | Case genes are co-expressed across bodily tissues 120
Fig 3.5 | GTS CNVR co-expression network 122
Fig 3.6 | GTS CNVR co-expressed genes are involved in transcriptional regulation 124
Fig 3.7 | Gene expression dynamics of case and GTS CNVR co-expressed genes 125
Fig 3.8 | Module characteristics 128

Fig 4.1 | Workflow to examine genotype and phenotype relationships in NDDs 144
Fig 4.2 | Choice of gene network thresholds 151
Fig 4.3 | Integrity of acyclic-like path trees 156
Fig 4.4 | CNV burden of the Nijmegen and SSC cohorts 158
Fig 4.5 | Phenotypic burden of the Nijmegen and SSC cohorts 159
Fig 4.6 | Genotype and phenotype convergence in Nijmegen patients with *de novo* CNVs 168
Fig 4.7 | Replication of phenotypic coherence using DECIPHER patients 173
Fig 4.8 | Genotype and phenotype convergence in Nijmegen patients with inherited CNVs 178
Fig 4.9 | Genotype and phenotype convergence in SSC patients with inherited CNVs 185

Fig 5.1 | Ancestry clustering 204
Fig 5.2 | Burden of *de novo* exonic events in the SSC. 216
Fig 5.3 | Genotype and phenotype convergence. 221
Fig 5.4 | Phenotypic coherence underlies phenotypic spectra. 229

Fig 5.5 | Correlation between phenotypic and genetic similarities across genome-wide SNPs, ASD GWAS SNPs and SNPs within NDD-associated genes 233
Fig 5.6 | Correlation between phenotypic and genetic similarities across SNPs near module genes 236

Fig 6.1 | Standardized workflow for morphological phenotyping of the adult mouse brain 260
Fig 6.2 | Integrative statistical pipeline 262
Fig 6.3 | Identification of mouse NAP genes 273
Fig 6.4 | GO enrichment analysis 275
Fig 6.5 | NAP genes are brain genes displaying synaptic functions 276
Fig 6.6 | NAP genes are functionally similar and co-expressed 277
Fig 6.7 | Expression patterns of mouse NAP and module genes 279
Fig 6.8 | Body-wide and brain expression patterns of human NAP orthologues 281
Fig 6.9 | Functional similarity network of human neuroanatomical orthologues 284
Fig 6.10 | Spatiotemporal brain expression patterns of human module genes 285
Fig 6.11 | Gene set enrichment across human modules 286
Fig 6.12 | Similarity of neuroanatomical defects caused by disruption of human module genes 287
Fig 6.13 | Overlap between mouse and human gene modules 289
Fig 6.14 | Predicted gene deleteriousness in humans is associated with stronger mouse neuroanatomical abnormalities 291
Fig 6.15 | NAP genes overlap ID genes 293

Fig 7.1 | From genomics to the clinic 317

Fig S3.1 | Expression dynamics of module compared to genome genes. 395
Fig S3.2 | Expression dynamics of module compared to CT genes 396
Fig S3.3 | Expression dynamics of module compared to co-expressed CT genes 397
Fig S3.4 | Co-expression patterns of module genes and various gene sets 398

Fig S4.1 | Clinical correlates in ASD 399
Fig S4.2 | Frequency of phenotypes in the Nijmegen and SSC cohorts .. 400
Fig S4.3 | GO enrichments of GTEx modules identified from genes mapping to *dn* CNVs from Nijmegen patients 401

Fig S4.4 | MGI enrichments of GTEx modules identified from genes mapping to *de novo* CNVs from Nijmegen patients 402
Fig S4.5 | Phenotype enrichments of patient modules 403
Fig S4.6 | Distribution of patient phenotypic similarity scores 404
Fig S4.7 | Odds ratios of essential phenotypes 405
Fig S4.8 | Module overlaps across networks and cohorts 406
Fig S4.9 | Overlap of essential HPO term between Nijmegen and DECIPHER modules 407
Fig S4.10 | Power and essential phenotypes for module of Nijmegen and SSC patients with inherited CNVs 408
Fig S4.11 | Temporal brain expression patterns of genes from modules identified using *de novo* CNVs from Nijmegen patients 409

Fig S5.1 | Choice of gene network thresholds 410
Fig S5.2 | Frequency of the 30 most common clinical phenotypes among sequenced SSC patients (n = 2,498) 411
Fig S5.3 | Module overlaps across networks and studies 412
Fig S5.4 | Temporal brain expression patterns module genes 413
Fig S5.5 | Distribution of patient phenotypic similarity scores 414
Fig S5.6 | Genes affected by LoF and delMIS variants affect common processes 415
Fig S5.7 | Phenotype enrichments across patient modules 416
Fig S5.8 | Statistical power and essential phenotypes 417
Fig S5.9 | Odds ratios of essential phenotypes 418
Fig S5.10 | Correlation between phenotypic and genetic similarities across eQTLs affecting NDD-associated genes 419

Fig S6.1 | Strategy used by the IMPC to generate knockout mice 420
Fig S6.2 | GO enrichment analysis of NAP genes determined at 5% FDR 421
Fig S6.3 | GO enrichment analysis of NAP genes determined at 1% FDR 422
Fig S6.4 | Developing CNS and adult brain gene co-expression network of NAP genes 423
Fig S6.5 | Functional similarity network of human NAP orthologues 424
Fig S6.6 | Functional similarity of human orthologues of NAP genes or human module genes and disease genes 425

List of Tables

Table 1.1 | High confidence ASD genes and their pleiotropic effects..... 19
Table 1.2 | Neurodevelopmental CNVs recurrently associated with ASD and their pleiotropic effects..... 27
Table 1.3 | Functions and genes implicated in NDDs..... 38
Table 1.4 | Large-scale mouse phenotyping initiatives..... 46
Table 1.5 | Properties and considerations of co-expression and PPI gene networks. 56

Table 2.1 | Description of genes lists formed. 72
Table 2.2 | Description of bodily structures constructed using the GTEx expression dataset..... 78
Table 2.3 | Description of gene networks. 80
Table 2.4 | Genomics datasets used to build the PLN. 82

Table 3.1 | Per sample quality control. 104
Table 3.2 | Per SNP quality control. 110
Table 3.3 | Number of common CNV calls stratified by size and platform biases. 110
Table 3.4 | Burden of rare CNVs ≥ 1 kb and ≥ 20 kb..... 113
Table 3.5 | Recurrently disrupted genes in GTS patients..... 115
Table 3.6 | OMIM genes observed in GTS patients. 131

Table 4.1 | Prevalence of phenotypes in the SSC cohort..... 149

Table 5.1 | Per sample quality control. 203
Table 5.2 | Per SNP quality control. 206
Table 5.3 | SNP and eQTL sets..... 211

Table 6.1 | ID associated genes among human NAP orthologues..... 295
Table 6.2 | List of genes mapping to ID, ASD and/or DD patients with cranial abnormalities. 298

Table S3.1 | Transcriptional gene regulation is affected in GTS patients. 427
Table S3.2 | GO enrichment analysis stratified by CNV size and direction. 428
Table S3.3 | Case gene clustering in GTEx, Brainspan and PPI networks. 429

Table S3.4 | Likelihood of haploinsufficiency of case genes and module gene. 430
Table S3.5 | Gene set enrichments of GTS module genes. 431
Table S3.6 | GO enrichment analysis of GTS modules. 432

Table S5.1 | Summary of GO terms over-represented among GTEx modules. 433
Table S5.2 | Summary of GO terms over-represented among Brainspan modules. 434
Table S5.3 | Summary of GO terms over-represented among PLN modules. 435
Table S5.4 | Summary of MGI phenotypes over-represented among GTEx modules. 436
Table S5.5 | Summary of MGI phenotypes over-represented among Brainspan modules. 437
Table S5.6 | Summary of MGI phenotypes over-represented among PLN modules. 438

Table S6.1 | Clustering of NAP genes and their human orthologues. 439
Table S6.2 | GO enrichment analysis of mouse modules. 440
Table S6.3 | GO enrichment analysis of human modules. 441

CHAPTER 1

The genetic landscape of neurodevelopmental disorders: an introduction

The development of the human brain is remarkably complex and relies on the synchronisation of proliferation, differentiation, migration and adhesion events to form the unified circuitry of approximately 100 billion neurons which governs human motion, cognition, emotion and language (Herculano-Houzel, 2009). It is, hence, unsurprising that abnormalities during brain development and in brain functioning lead to a wide range of neurological and psychiatric conditions, collectively termed NeuroDevelopmental Disorders (NDDs). NDDs range from rare but distinctive neurological syndromes to more common yet clinically broader conditions, including Developmental Delay (DD), Intellectual Disability (ID), Autism Spectrum Disorders (ASD), Attention Deficit Hyperactivity Disorder (ADHD), Gilles de la Tourette Syndrome (GTS), and language impairments (Specific Language Impairment; SLI). Of note, although schizophrenia is often considered an NDD, we mainly exclude it from this thesis due to the delayed period of onset compared to the remaining conditions.

NDDs are collectively characterised by deficits in personal, academic or occupational functioning, arising before puberty. Individually, each disorder underlies a distinctive set of core diagnostic symptoms, as defined by the American Psychiatric Association in the Diagnostic and Statistical Manual of mental disorders, fifth edition (DSM-5) (American Psychiatric Association, 2013). For example, DD is broadly linked to delays in functional aspects of development while ID is specifically characterised by substantial limitations in intellectual functioning and adaptive behaviours, and defined by Intellectual Quotients (IQ) below 70; and ASD is associated with deficits in social

interaction and communication, as well as restricted and repetitive behaviours. Other NDDs, such GTS, ADHD and SLI are characterised by the presence of (i) motor and vocal tics, (ii) inattentiveness, hyperactivity and impulsiveness and (iii) difficulties in speech and language understanding, respectively. The qualitative impairments of the diagnostic features of each disorder are highly variable and often accompanied by additional comorbidities, malformations and medical problems resulting in a compelling amount of clinical heterogeneity. Accordingly, epidemiological studies revealed that between 30% and 70% of ID patients display additional comorbidities (Cooper *et al.*, 2007; Einfeld *et al.*, 2011; Cooper *et al.*, 2015), such as ASD (i.e. 4-30%) (Cooper *et al.*, 2007; Van Naarden Braun *et al.*, 2015), epilepsy (i.e. 18-30%) (Cooper *et al.*, 2015), anxiety disorders (i.e. 3-8%) (Cooper *et al.*, 2007; Cooper *et al.*, 2015), ADHD (i.e. 1.5%) (Cooper *et al.*, 2007) or cerebral palsy (i.e. 13%) (Van Naarden Braun *et al.*, 2015). Conversely, among ASD patients, 30% to 61% display comorbid ID while 30%, 14-78% and 15-60% have epilepsy, ADHD and anxiety disorders, respectively and up to 85% and 90% display sleep and gastrointestinal disturbances, respectively (Jeste, 2011; Mannion and Leader, 2013; Mannion *et al.*, 2014; Van Naarden Braun *et al.*, 2015; Christensen *et al.*, 2016). Comorbidities are also widespread among less severe NDDs, such as GTS, where 11-80% and 15-90% of affected individuals display comorbid Obsessive Compulsive Disorder (OCD) and ADHD, respectively (Sheppard *et al.*, 1999; Mol Debes, 2013). Conversely, in ADHD, 44%, 1-8% and 1% of patients have an anxiety disorder, GTS or tic disorder and/or epilepsy, respectively (Steinhausen *et al.*, 2006).

Collectively, NDDs have an estimated prevalence of 15% and individually DD, ID, ASD and GTS occur in 1% of the population (Shevell *et al.*, 2003; Robertson *et al.*, 2009; Maulik *et al*, 2011; Lai *et al.*, 2014) while ADHD and language impairments are more common and affect 5-7.2% (Willcut, 2012; Thomas *et al.*, 2015) and 7-10% (Tomblin *et al.*, 1997; Norbury *et al.*, 2016) of children, respectively. There are substantial gender differences in prevalence across disorders, with males being two to four times more commonly affected than females (Geller, 2006; Newschaffer *et al.*, 2007; Ramtekkar *et al.*, 2010; Boyle *et al.*, 2011; Lichter and Finnegan, 2015). NDDs represent a major public health problem with an annual economic burden of £32 billion for autism alone in the UK (Buescher *et al.*, 2014), making the development of effective therapies a medical priority. Current treatment strategies include the use of educational interventions to alleviate the behavioural or cognitive difficulties encountered in ASD or ID and the use of psychopharmacology to target psychiatric symptoms, such as depression, anxiety or epilepsy using anti-depressants, anxiolytics and anticonvulsants, respectively. These treatments, however, display only limited effectiveness and drugs can have dangerous side effects. As predicting the outcome of various therapies is difficult based on clinical profiles alone, a better understanding of the underlying pathogenic mechanisms of NDDs is necessary to specifically tailor safe and effective treatment strategies to affected individuals.

NDDs have a strong genetic component (van Loo and Martens, 2007; Stessman *et al.*, 2016). Indeed, family studies identified an increased risk in siblings of NDD patients compared to the general population (ASD: Ozonoff *et*

al., 2011; Constantino *et al.*, 2010; ADHD: Biederman *et al.*, 1990; Faraone *et al.*, 1991; GTS: Pauls *et al.*, 1991; O'Rourke *et al.*, 2009). Moreover, twin studies revealed stronger concordances among monozygotic compared to dizygotic twins (ASD: Bailey *et al.*, 1995; Rosenberg *et al.*, 2009; ADHD: Faraone *et al.*, 2005; GTS: Price *et al.*, 1985; O'Rourke *et al.*, 2009; SLI: Lewis and Thompson, 1992; Bishop *et al.*, 1995; Tomblin and Buckwalter, 1998), suggesting the genetic origin of NDDs is not confounded by the presence of shared environments.

Over the past decade, the advent of using high throughput technologies such as chromosomal microarrays and Next Generation Sequencing (NGS) enabled the genetic screening of thousands of patients (**Boxes 1 and 2**). This led to the discovery of genetic variants (i.e. whole chromosomes, DNA segments or single genes) specifically associated with neurological syndromes, as well as hundreds of other genes and loci conferring risk to individual or multiple NDDs. The simultaneous examination of hundreds of such genetic variants was further enabled by the development of holistic genomics approaches, such as pathway analysis, which uncovered common neurodevelopmental processes defective in NDDs. The most prominent classes include (i) cell proliferation and patterning events during early brain development, (ii) cell migration events and axon guidance required to form the cohesive neuronal circuitry and (iii) synapse formation, pruning and plasticity. The identification of causal genes linked to these categories across NDDs suggests an overlapping genetic aetiology and propose that the same disorders might result from the disruption of biologically coherent gene

groups, commonly referred to as a *biological pathway*. Beyond the identification of such pathways, cellular and animal models are used to study the underlying pathogenic mechanisms and uncover the molecular and systems abnormalities causing the observed clinical characteristics (i.e. phenotype). Consequentially, disentangling the symptomatic, genetic and biological heterogeneity of NDDs should facilitate the development of effective personalised therapies, based on patients' full clinical and aetiological profiles.

In this thesis, I investigated the genetic architecture of GTS (**chapter 3**), examined the relationship between genotype and phenotype in DD, ID and ASD (**chapters 4 and 5**) and evaluated the relevance of mouse models to study ID and identify novel candidate genes (**chapter 6**). This introductory chapter outlines the current genetic and aetiological understanding of NDDs.

Box 1 | Types of human genetic variation

The chromosomal structure and DNA sequence differs between any two genomes, such that the examination of these differences provides a powerful tool to detect the genetic variation underlying a trait or disease of interest. Multiple types of variation from the reference human genome (The Genome Sequencing Consortium, 2001) have been described:

- Common variation (CV): variation present in more than 1% of the population.
- Rare variation (RV): variation present in less than 1% of the population.
- Inherited variation (inh): variation present in all cells from at least one parent, which was transmitted to the child.
- *De novo* variation: spontaneous variation, absent from parents, arising in a germ cell (egg or sperm) and affecting all cells from the individual.
- Coding variation: variation affecting protein-coding regions of the genome.
- Non-coding variantion: variation affecting non-coding regions of the genome, including introns, intergenic segments and regulatory elements.
- Chromosomal aneuploidy: abnormalities in chromosome number - only one or more than two chromosome copies are present in a cell.
- Chromosomal rearrangement: structural chromosomal abnormality, whereby large DNA segments are deleted, duplicated, inverted or translocated by means of anomalous repair of double-strand DNA breaks.
- Copy number variation (CNV): submicroscopic structural DNA variation involving the deletion or duplication of more than 1000 bases (or 1 kb).
- Single nucleotide polymorphism (SNP): single nucleotide variation common in the population.
- Single nucleotide variation (SNV): single nucleotide variation rare in the population.
- Insertion-deletion (indel): DNA variation involving the deletion or insertion of a small number of nucleotides. Throughout this thesis, indels are included in the definition of SNVs.
- Synonymous mutation: nucleotide substitution which does not alter the amino acid encoded by the impacted codon.
- Missense mutation: non-synonymous nucleotide substitution affecting the amino acid encoded by the impacted codon.
- Nonsense mutation: nucleotide substitution switching an amino acid for a chain-terminating codon.
- Frameshift mutation: indel of a number of nucleotides not divisible by three leading to a change in codon grouping (reading frame), such that the amino acid sequence is altered.
- Splice-site mutation: indel affecting the specific sequence site at which splicing occurs during the processing of precursor messenger RNA.
- Loss of function variation (LoF): nonsense, frameshift or splice-site mutations.
- Likely gene-disrupting variation: LoF or missense mutations.
- Penetrance: the proportion of mutation carriers displaying a particular phenotype. More generally, we use the terms penetrant and high-confidence genes to describe those recurrently observed in patients and confidently associated with the disease. Conversely, disease risk, susceptibility and associated genes refer those underpinning variable lines of evidence.
- Expressivity: proportion of mutation carriers displaying a phenotype of interest relative to all observed phenotypes.

* N.B. throughout this thesis, the terms variation and variant are interchangeable.

Box 2 | Types of genetic studies used to examine genomic variation

The tremendous amount of methodological progress over these past decades has led to the development of high throughput methodologies capable of detecting various types of DNA variation. Described below are the different approaches used to investigate genetic variation in disease:

- Cytogenetic analysis: study of the number and structure of chromosomes in order to identify microscopically-visible chromosomal abnormalities of 4 to 6 megabases (mb).
- Linkage analysis: Study of large multi-generational families, in which a subset of members display the disease (or trait) of interest. The aim is to map disease-associated genetic loci by identifying those inherited by family members with the disease but not by those without the disease.
- Genome-wide association studies (GWAS): genome-wide scan of a large set of genetic variants (i.e. SNP markers detected using oligonucleotide arrays) to identify those occurring more commonly in individuals with a disease (or trait) of interest compared to those without. GWAS approaches are designed to examine common DNA variation, and particularly SNPs although CNVs can also be detected using the associated methodology.
- Whole-exome sequencing (WES): Next generation sequencing technology aiming to elucidate the whole coding DNA sequence, by firstly (1) selecting the DNA sequences encoding proteins (exons) using a target-enrichment strategy (e.g. array-based capture) and subsequently (2) sequencing that DNA using a high-throughput technology. WES approaches are powered to detect any types of DNA variation, although it is primarily used in the study of rare variation (SNVs and CNVs).
- Whole-genome sequencing (WGS): Next generation sequencing technology to elucidate the whole DNA sequence of an individual. It is powered to detect any type of variation in a genome-wide fashion and will specifically be used to investigate the roles of non-coding variation. Large-scale applications of WGS are uncommon due to their high cost and our inability to interpret non-coding variation.

1.1 Genetic models of NDDs

This section reviews the diverse genetic models of NDDs and explores their theoretical and empirical foundations (**Fig 1.1**).

1.1.1 Genetic models of NDDs: monogenic versus polygenic models

The Mendelian genetic model, which proposes that a phenotype or trait is determined by a single locus (**Fig 1.1**), has been used to describe over 4,100 single-gene disorders (Chong *et al.*, 2015). Linkage analyses aim to uncover disease-associated loci by examining genomic DNA markers in families with several affected individuals (termed *multiplex* families) (**Box 2**). Such studies have successfully identified causal genes for many rare clinical syndromes, such as Huntington's disease (Gusella *et al.*, 1983) and Rett syndrome (Ellison *et al.*, 1992; Amir *et al.*, 1999) (for comprehensive list see Harding and Rosenberg, 1993). Additionally, they revealed disease-related loci for more common NDDs, including ASD (Autism Genome Project Consortium *et al.*, 2007; Freitag, 2007), seizures (Nakayama, 2009), ADHD (Sharp *et al.*, 2009) and SLI (Bartlett *et al.*, 2004; SLI Consortium, 2004). Nevertheless, most signals remain unreplicated and few causal genes have been established, advising Mendelian models under high genetic heterogeneity or polygenic models should be applied to NDDs (Risch, 1990; Risch and Merikangas, 1993; McClellan and King, 2010).

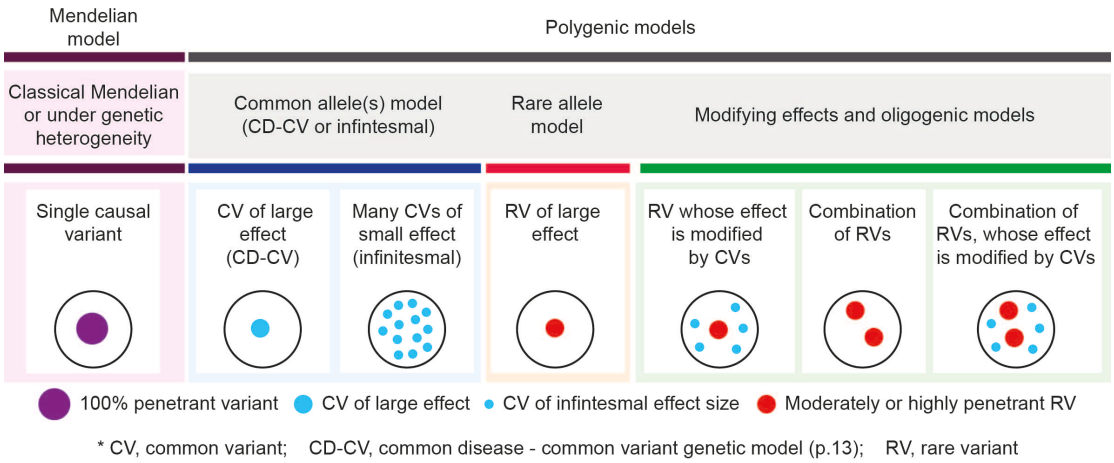


Fig 1.1 | Genetic models of NDDs. Schematic of the expected burden and effect size of disease risk variants under Mendelian and polygenic (i.e. common, rare and oligogenic) models of NDDs. Circles depict variants, whose penetrance is proportional to the circle size; purple, red and blue fills refer to ultra rare or rare, rare and common variants, respectively. CV, common variants; CD-CV, common disease-common variant model; RV, rare variant. (Based on Berg and Geschwind, 2012).

In Mendelian models of high genetic heterogeneity single mutations in any one of a large number of disease genes induce the associated phenotypes. While such models have successfully been applied to conditions including retinitis pigmentosa (Hartong *et al.*, 2006) and hearing impairment (Dror and Avraham, 2010), they have generally been ineffective across broader NDDs. This is likely to be due to the lack of classical Mendelian inheritance patterns associated with the high disease risks in siblings and family members. Instead, polygenic models of NDDs produced stronger evidence.

Polygenic models propose that the joint effects of many independent or interacting variants (i.e. additive and non-additive models, respectively) cause disease (**Fig 1.1**). In additive polygenic models, risk variants are distributed through the population and their segregation defines each individual's vulnerability to disease (termed *liability*), which manifests itself above a certain

threshold of the liability distribution (**Fig 1.2a**) (Falconer, 1965). Conversely, non-additive epistatic models propose that the phenotypic effects of one variant depend on the genotype of a second variant, such that a particular trait arises from the interaction between two or more gene variants (**Fig 1.2c**) (Mackay and Moore, 2014; Hemani *et al.*, 2014). The contribution of such epistatic effects in NDDs is supported by the recent observation that deleterious phenotypes arise from the combined – but not single – inactivation of fly orthologues of autism genes (Grice *et al.*, 2015). Along with these polygenic models, mixed models (i.e. involving both additive and non-additive genetic effects) and multifactorial models (i.e. both genetic and non-genetic factors contribute to disease) (**Fig 1.2d**) represent the prevailing paradigms in the genetics of NDDs. An extension of these models derives from the decomposition of NDDs into continuous components, termed *endophenotypes*, such that disease onset is instigated when several traits (governed by the previously discussed models) coincide (Gottesman and Gould, 2003). Moreover, particular allelic combinations of genes could predispose towards diverse forms of disease and responses to therapies (Webber, 2017), hence supporting the observed high clinical heterogeneity.

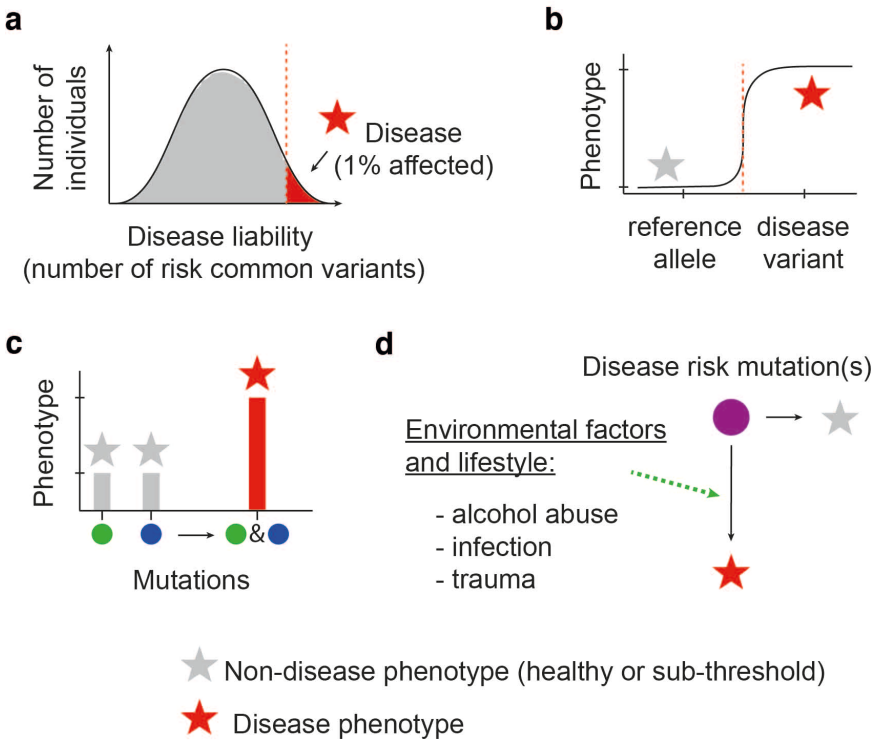


Fig 1.2 | Mechanisms of pathogenicity. The genetic architecture of NDDs manifests heterogeneous mechanisms of causality, four of which are described here. (**a**) Disease liability, governed by the segregation of common variants, is normally distributed across the population, such that above a threshold burden of risk common variants, the disease state is reached. This occurs in approximately 1% of the population for ID and ASD. (**b**) Penetrant rare variants exert dominant or recessive effects, mediated by haploinsufficiency, loss of function or gain of gene function, which translate into abnormal phenotypes. (**c**) Epistatic models propose that the interaction between two (or more) genes is necessary to trigger a phenotype. Under this model, the disorder is only expressed in individuals displaying variants in both genes while carriers of single gene variants have undetectable phenotypes. (**d**) Under multifactorial models, *gene x environment* effects underlie a disease state. Specifically, genetic variants make individuals more susceptible to the effects of environmental factors, such that particular combinations of risk genes and environmental factors result in disease.

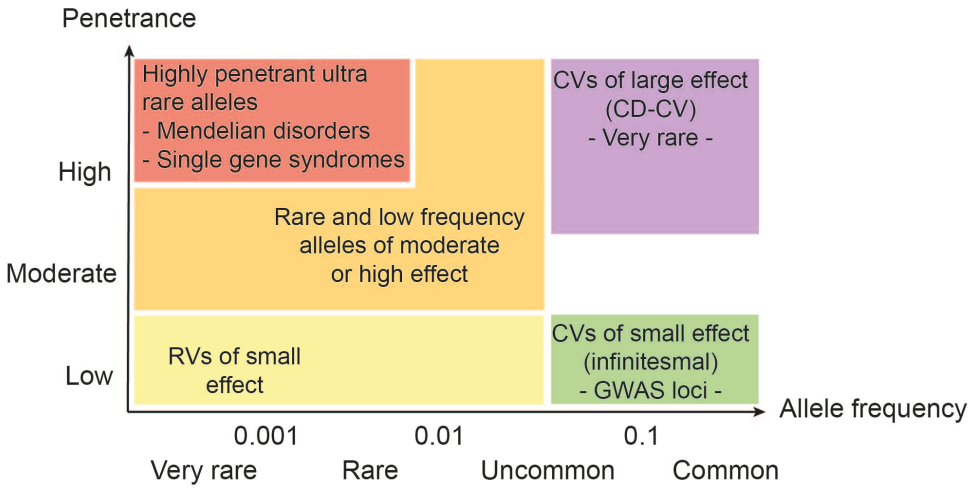
1.1.2 Frequency of alleles: common versus rare variants

Polygenic disorders implicate three genetic origins, namely (1) Common Variants (CVs) occurring in over 1% of the population and termed Single Nucleotide Polymorphisms (SNPs), (2) Rare Variants (RVs) occurring in less

than 1% of the population and (3) the combination of genetic, epigenetic and environmental interactions (**Box 1**).

The *Common Disease – Common Variant* (CD-CV) hypothesis, developed following the launch of the Human Genome Project, proposes that common disorders arise from the combination of one or several CVs of moderate or large effects (**Figs 1.1 and 1.3**) (Lander, 1996). Nevertheless, the CD-CV model can not be effectively applied to NDDs. Firstly, from an evolutionary standpoint, given the increased mortality and the reduced fecundity rates observed in NDDs (Barbaresi *et al.*, 2013; Dalsgaard *et al.*, 2015; Bilder *et al.*, 2013; Hirvikoski *et al.*, 2016; Hosking *et al.*, 2016; Fairthorne *et al.*, 2014; Keller and Miller, 2006; Power *et al.*, 2013), selection should remove genetic variants associated with such reductions in fitness. Several arguments have been proposed to explain the lack of adherence of NDDs to this evolutionary model, including (1) balancing selection, where variants are deleterious or beneficial in different contexts (Keller and Miller, 2006). This scenario is typically exemplified by the *HBB* gene, such that gene mutations cause sickle cell anaemia but protect from malaria (Hedrick, 2006). Additionally, (2) ancestral neutrality suggests the causal variant was neutral or beneficial in a different environment or among our ancestors. Nonetheless, these arguments lack supporting evidence (Bubb *et al.*, 2006; Keller and Miller, 2006). Secondly, Genome-Wide Association Studies (GWAS), which aim to uncover disease-associated SNPs (**Box 2**), revealed that particular loci explain only a small fraction of disease susceptibility, thus raising the question of *missing heritability* (Gibson, 2012). The CD-CV hypothesis was thereupon abandoned

in favour of a more promising genetic model involving a large number of CVs of small effect sizes.



CV, common variant
CD-CV, common disease - common variant genetic model
RV, rare variant

Fig 1.3 | Variant frequency and penetrance in disease. Mendelian and other single gene disorders are caused by extremely penetrant very rare or rare (frequencies of < 0.1% and < 1% in the population, respectively) variants (red). Conversely, several alternative models have been proposed for NDDs, wherein disease results from the combination of a large number of common variants (CVs) of individually small effects (infinitesimal model; green) or from one or few rare variants (RVs) of moderate to high effects (orange). Models involving single CVs of large effect (CD-CV model; purple) or RVs of small effect (yellow) are rare. CD-CV, common disease-common variant genetic model.

The *infinitesimal* model proposes that a large number of common loci contribute to disease susceptibility, such that individually, each displays an infinitesimal effect size, invisible to selection (**Figs 1.1 and 1.3**). This model builds on the view that that the proportion of genetic variance explained for a given trait is proportional to the number of loci considered (Gibson, 2012). Correspondingly, studies have shown that (1) nominally significant GWAS loci collectively explain more variance than genome-wide significant SNPs

(International Schizophrenia Consortium *et al.*, 2009) and that (2) genetic variance is proportionately distributed across chromosomes, such that larger chromosomes capture more variance (Davies *et al.*, 2011). The success of GWAS in identifying hundreds of trait- and disease-associated loci using large numbers of individuals, such as for height (Wood *et al.*, 2014) and schizophrenia (Ripke *et al.*, 2014), is consistent with the infinitesimal model and suggests failures of GWAS are attributable to technical limitations (i.e. sample sizes too small to detect variants of such small effect sizes). Moreover, disease-associated CVs have been shown to collectively influence NDD endophenotypes, such as social cognition and verbal reasoning in schizophrenia (Germine *et al.*, 2014) and social communication in ASD (Robinson *et al.*, 2016). These findings underline the relevance of infinitesimal models in describing threshold-dependent endophenotype models of NDDs. Nevertheless, multiple arguments against the infinitesimal model have also been proposed. Most notably, the majority of risk variants uncovered through GWAS have not been functionally validated, raising the question of false positives (i.e. is the variant truly causal or is it synthetically associated with a causal variant) (Gibson, 2012). Furthermore, this genetic model is not consistent with some observed demographic effects, such as the differences in disease risk across geographic areas, which are not caused by genetic ancestry. While this could be a consequence of *gene x gene* and *gene x environment* interactions, GWAS are underpowered to detect such synergies (Gibson, 2012). Finally, most GWAS undertaken for NDDs failed to uncover genome-wide significant loci, replicate previously observed signals and identify novel candidate suggesting infinitesimal models are incomplete and

capture only a limited fraction of genetic variance. An alternate view of the infinitesimal model suggests that while CVs alone have negligible phenotypic effects, they act as modifiers of rare penetrant mutations. This model will be further discussed in **section 1.2.4**.

The rare allele model proposes that the major genetic component of disease can be attributed to moderately or highly penetrant RVs occurring at frequencies of less than 1% in the population (**Box 1; Figs 1.1 and 1.3**). Accordingly, RVs exert a dominant effect, mediated by HaploInSufficiency (HIS), Loss of Function (LoF) or gain of function mechanisms, resulting in disease (**Fig 1.2b**). This model does not assume full penetrance of each RV, such that several RVs (or RVs and CVs) can combinatorially or additively contribute to risk and phenotypic expressivity. Furthermore, disease risk and expressivity can be modulated by additional CVs (**section 1.2.4**). The rare allele model is consistent with evolutionary genetics, such that the existence of disease-causing variants reveals a balance between DNA mutation rates and purifying selection, which rapidly removes deleterious alleles or maintains them at low frequencies. This view is supported by the observation that mutations which alter a gene's amino acid sequence (i.e. non-synonymous mutations) are more likely to be deleterious and occur at lower frequencies than those predicted to be neutral (i.e. synonymous mutations or mutations without any functional impacts) (Cargill *et al.*, 1999; Kryukov *et al.*, 2007; Marth *et al.*, 2011; Zhu *et al.*, 2011; Subramanian, 2012). The validity of the rare allele model is further reinforced by the identification of rare causal gene mutations in many X-linked ID syndromes (Tarpey *et al.*, 2009; Lubs *et al.*,

2012) and of rare deletions or duplications of DNA segments larger than 1kb (termed Copy Number Variants as they alter the number of copies of the overlapped genes; CNVs) in ID and ASD syndromes (Moreno-De-Luca *et al.*, 2013). In this thesis, we only consider RVs affecting coding DNA sequence regions and neglect the contribution of regulatory variants as these remain difficult to identify and their impact is not readily interpretable. Undoubtedly, however, large-scale projects such as the ENCODE project (ENCODE Project Consortium, 2012) will help elucidating the roles of non-coding DNA regions and whole-genome sequencing studies will evaluate their contribution to disease, as it has recently been done for ASD (Turner *et al.*, 2016).

In the rare allele model, the deleteriousness of a variant is inversely proportional to its frequency such that rarer and more recent mutations, which have not yet been subject to purifying selection, are more penetrant. These primarily refer to mutations arising for the first time in a family member (i.e. *de novo*; *dn*) as a result of mutations in germ cells. Consequentially, research has primarily focused on rare gene point mutations (i.e. Single Nucleotide Variants; SNVs) (**Box 1**) arising *dn*, such that there are 74 *dn* SNVs per genome, of which 1 to 4 affect the exome (Bourgeron, 2015; Yuen *et al.*, 2017). Their study led to the identification of causal variants or genes for multiple NDDs, including ASD and ID and their associated syndromes (**Table 1.1**) (Veltman and Brunner, 2012). In particular, *dn* LoF SNVs, defined as nonsense, frameshift and splice-site mutations with high probabilities of functional impacts on protein products, represent a highly deleterious class of variants. These variants occur twice as often in ASD patients compared to

their unaffected siblings (Iossifov *et al.*, 2012; Neale *et al.*, 2012; O'Roak *et al.*, 2012; Sanders *et al.*, 2012; Iossifov *et al.*, 2014), suggesting they target disease-associated genes. More generally, both *dn* LoF and/or missense mutations are more common in NDD cases compared to healthy controls (Iossifov *et al.*, 2014; Hoischen *et al.*, 2014; Shohat *et al.*, 2017; Stessman *et al.*, 2017), such that they collectively (i.e. Likely Gene-Disrupting variants; LGD) encompass an expanded class of likely pathogenic mutational events. Recently, the study of private inherited SNVs, defined as those present in unique families, has gathered attention as an excess of LoF mutations, affecting evolutionarily constrained genes, was observed in ASD patients compared to their unaffected siblings (Krumm *et al.*, 2015). Nonetheless, this thesis neglects their contribution, focusing primarily on *dn* events. In addition to SNVs, the rare allele model incorporates disease susceptibility roles for rare CNVs. This thesis examines both *dn* and inherited CNVs as (1) many CNVs associated with neurological syndromes are inherited (Moreno-De-Luca *et al.*, 2013; Veltman and Brunner, 2012; Krumm *et al.*, 2015). Additionally, (2) inherited CNVs are generally more likely to have functional impacts, compared to inherited SNVs, as they disrupt larger DNA segments.

Table 1.1 | High confidence ASD genes and their pleiotropic effects.

Gene	Description	Locus	Estimated percentage of ASD patients with variant	ASD penetrance	Neuropsychiatric pleiotropy	Somatic pleiotropy
KATNAL2	katanin p60 subunit A-like 2	18q21.1	0.08	Unknown	Unknown	Unknown
POGZ	pogo transposable element with ZNF domain	1q21.3	0.08	Incomplete	ID, speech delay, language delay, schizophrenia	Microcephaly, obesity, impaired vision
TBR1	T-boc brain 1	2q24.2	0.08	Unknown	ID	Unknown
ADNP	activity-dependent neuroprotector homeobox protein	20q13.13	0.1	Complete	ID, ADHD	Recurrent infections, short stature, heart defect, hypotonia, hypermetropia, epilepsy, hyperlaxity
SYNGAP1	Synaptic Ras GTPase-activating 1	6p21.32	0.1	Unknown	ID	Epilepsy
GRIN2B	glutamate ionotropic receptor NMDA type subunit 2B	12p13.1	0.13	Unknown	ID	Epilepsy
ANK2	ankyrin 2	4q25-q26	0.13	Unknown	None reported	Heart arrhythmia
ARID1B	AT-rich interactive domain-containing 1B	6q25.3	0.13	Incomplete	ID, speech impairment	Short stature, hypertrichosis, cryptorchidism, epilepsy, vision impairment
SCN2A	sodium voltage-gated channel α subunit 2	2q24.3	0.13	Incomplete	ID, schizophrenia	Epilepsy, episodic ataxia
DYRK1A	dual specificity tyrosine-phosphorylation-regulated kinase 1A	21q22.13	0.13	Incomplete	ID, speech impairment, ADHD, anxiety	Microcephaly, epilepsy, vision impairment, short stature, gastrointestinal symptoms or feeding difficulties
CHD8	chromodomain helicase DNA-binding 8	14q11.2	0.21	Incomplete	ID, schizophrenia, speech delay, sleep problems	Macrocephaly, gastrointestinal symptoms

Genes, identified through exome sequencing studies, which are recurrently disrupted in ASD patients. These genes also occur in other neuropsychiatric conditions and often involve additional phenotypes. (Adapted from Vorstman *et al.*, 2017).

1.2 The genetic architecture of NDDs

This section presents the main genetic findings from GWAS, CNV and gene studies of NDDs and reviews several integrative models of NDDs.

1.2.1 GWA studies of NDDs

Aiming to predict genetic susceptibilities for a trait of interest, GWAS have been developed to examine allelic frequencies at hundreds of thousands of genomic loci. Specifically, GWAS assess whether particular nucleotides, termed Single Nucleotide Polymorphisms (SNPs), are more common in individuals displaying the trait of interest. SNPs represent ancient mutations at

specific chromosomal locations, which have spread through an appreciable proportion (> 1%) of the population as a result of random genetic drift. Due to uneven patterns of recombination, SNPs are organised in blocks, termed haplotypes, which are co-inherited (International HapMap Consortium, 2003; Hardy and Singleton, 2009). These can be used to accurately infer the allelic variants at adjacent loci, hence enabling the genetic examination of millions of DNA variants using a substantially smaller set of genotyped markers. Accordingly, based on the assumption that disease susceptibility variants are more frequent in cases than controls, GWAS have been successful in identifying SNPs predisposing to disorders such as schizophrenia (Ripke *et al.*, 2014) or coronary artery disease (Nikpay *et al.*, 2015; McPherson and Tybjaerg-Hansen, 2016). More generally, close to 25,000 reliable SNP-trait associations have been made so far and can be accessed from the US National Human Genome Resource Institute and European Molecular Biology Laboratory-European Bioinformatics Institute (NHGRI-EBI) GWAS catalogue (MacArthur *et al.*, 2017). On the other hand, GWAS of NDDs have mostly failed to detect or replicate significant signals, hence questioning the genetic contribution of CVs to ASD (Ma *et al.*, 2009; Wang *et al.*, 2009; Weiss *et al.*, 2009; Anney *et al.*, 2010; Anney *et al.*, 2012; Kuo *et al.*, 2015; Chaste *et al.*, 2015; Liu *et al.*, 2016a; Mitra *et al.*, 2016), language impairment (Eicher *et al.*, 2013; Nudel *et al.*, 2014; Gialluisi *et al.*, 2014), GTS (Scharf *et al.*, 2013; Paschou *et al.*, 2014), OCD (Stewart *et al.*, 2013; Mattheisen *et al.*, 2015), ADHD (Neale *et al.*, 2010; Lantieri *et al.*, 2010; Stergiakouli *et al.*, 2012; Zayats *et al.*, 2015) as well as NDD endophenotypes (Rommelse *et al.*, 2008; Bakken *et al.*, 2011; Fanous *et al.*, 2012; Connolly *et al.*, 2013; Jones *et al.*,

2013; Luciano *et al.*, 2013; Bramon *et al.*, 2014; Goodbourn *et al.*, 2014; Chaste *et al.*, 2015; Burton, 2016; Crosbie, 2016; Tao *et al.*, 2016; Cantor *et al.*, 2017). Moreover, focusing on SNPs affecting specific classes of genes, no increased polygenic risk was detected for SNPs nearby genes from 32 NDD-related functional classes (Jansen *et al.*, 2017). Only SNPs in proximity of genes regulated by the Fragile X Mental Retardation Protein (FMRP) showed an association with ASD status (Steinberg and Webber, 2013; Jansen *et al.*, 2017), although this signal was not replicated using auxiliary sets of FMRP target genes (Jansen *et al.*, 2017).

A possible explanation for the lack of significant and reliable GWAS findings involves the use of relatively small and statistically underpowered patient cohorts. In schizophrenia (O'Donovan *et al.*, 2008; International Schizophrenia Consortium *et al.*, 2009; Ripke *et al.*, 2011; Ripke *et al.*, 2014) and bipolar disorder (Baum *et al.*, 2008; Ferreira *et al.*, 2008; Sklar *et al.*, 2011; Chen *et al.*, 2013; Mühleisen *et al.*, 2014; Hou *et al.*, 2016), for example, the examination of increasing numbers of subjects led to the discovery of growing numbers of disease-associated loci, suggesting GWAS of NDDs remain promising. In fact, a cross-disorder GWAS of patients with ASD, ADHD, schizophrenia, bipolar and major depression disorder identified several statistically significant loci, highlighting the gain in power of detection resulting from the interrogation of large cohorts (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013). Additionally, a recent ASD meta-analysis identified a genome-wide significant SNP in proximity of the *EXT1* gene, in which rare mutations have previously been observed in two ASD

patients (Mitra *et al.*, 2016). The strong influence of cohort sizes on the performance of GWAS has also been demonstrated in the study of endophenotypes. Indeed, the advent of large-scale studies, such as the UK Biobank (Sudlow *et al.*, 2015), enabled the identification of many statistically significant loci associated with general cognitive and psychological measures correlated with NDDs (Davies *et al.*, 2016; Lane *et al.*, 2016; Okbay *et al.*, 2016; Smith *et al.*, 2016). Furthermore, recent estimates propose that approximately 50% of the narrow-sense heritability of autism and other NDDs derives from CVs, hence corroborating their contribution to disease (Anney *et al.*, 2012; Klei *et al.*, 2012; Cross-Disorder Group of the Psychiatric Genomics Consortium *et al.*, 2013; Gaugler *et al.*, 2014). Nevertheless, the lack of positive signals in recent NDD GWAS, comprising reasonably large numbers of subjects, suggests that even if identified, individual SNPs minimally increase disease risk. This underscores the need to uncover more informative genetic (i.e. RVs) and non-genetic risk factors.

While the study of sex differences in NDDs goes beyond the scope of this thesis, we briefly note that since the strong gender bias is presumed to be of genetic origin, the combined evaluation of males and females may provide an inadequate strategy. In fact, a recent ASD meta-analysis, which performed sex-specific GWA studies, identified different genome-wide significant signals in the male- and female-specific studies, which, nevertheless, remain to be replicated (Mitra *et al.*, 2016). This study investigated several genetic models revealing male preponderance in ASD is unlikely to be due to a higher mutational load needed to elicit disease in females or to sex-specific

autosomal susceptibility factors (Mitra *et al.*, 2016). On the other hand, X chromosome variants and SNPs governing anthropometric traits were associated with increased male ASD risk, suggesting the lack of a compensatory X chromosome contributes to disease and general mechanisms of sexual dimorphism are implicated in ASD (Mitra *et al.*, 2016).

The biological interpretation of GWAS results is difficult, as the affected genes are not readily detectible given that (1) SNPs rarely cause amino acid changes and may have regulatory impacts, and (2) SNPs do not necessarily constitute the risk variants (see below). Furthermore, the Odds Ratios (ORs), used to calculate SNP effect sizes, generally estimate the increase in disease risk to be less than 1.2-fold. Although this observation is consistent with the infinitesimal model, where thousands of loci individually contribute minimally to disease liability, alternate hypotheses are conceivable. For example, as the haplotypic structure of the human genome entails any variants co-inherited with the significant SNP can be causal, the signal may actually be synthetically driven by RVs of larger effects associated with that particular haplotype (Dickson *et al.*, 2010; Wang *et al.*, 2010; Sun *et al.*, 2011; Rivas *et al.*, 2011; Chang and Keinan, 2012; Beaudoin *et al.*, 2013; Saunders *et al.*, 2014). Nevertheless, such mechanisms cannot explain the majority of GWAS signals (Orozco *et al.*, 2010; Anderson *et al.*, 2011; Wray *et al.*, 2011; Hunt *et al.*, 2013; Fuchsberger *et al.*, 2016). A second possibility, discussed in **section 1.2.4**, is that common loci act as genetic modifiers of RVs, such that their effects are negligible in absence of such mutations.

1.2.2 CNV studies of NNDs

Classical cytogenetic screening of patients displaying analogous behavioural and physical abnormalities led to the identification of many rare syndromes associated with particular chromosomal anomalies. With the development of microarray technologies, systematic CNV screens across whole genomes revealed that *dn* CNVs arise significantly more often in ASD patients compared to their unaffected siblings or healthy controls (Sebat *et al.*, 2007; Marshall *et al.*, 2008; Itsara *et al.*, 2010; Pinto *et al.*, 2010; Levy *et al.*, 2011; Sanders *et al.*, 2011; Pinto *et al.*, 2014; Sanders *et al.*, 2015). Moreover, these studies led to the discovery of ASD-predisposing CNVs, recurrently observed in patients (**Table 1.2**) (Kumar *et al.*, 2008; Marshall *et al.*, 2008; Weiss *et al.*, 2008; Glessner *et al.*, 2009; Levy *et al.*, 2011; Sanders *et al.*, 2011; Sanders *et al.*, 2015; Leppa *et al.*, 2016; Yuen *et al.*, 2017) (see Moreno-De-Luca *et al.*, 2013 for review). Furthermore, the over-representation of rare large CNVs among individuals with DD (Coe *et al.*, 2014), ID (Cooper *et al.*, 2011; Girirajan *et al.*, 2011; Vulto-van Silfhout *et al.,* 2013a), ADHD (Williams *et al.*, 2010; Williams *et al.*, 2012) and GTS (Nag *et al.*, 2013; Huang *et al.*, 2016) supports their implication in disease.

Studies of rare CNVs have identified many *neurodevelopmental CNVs*, which occur at higher frequencies in NDD cases than controls, such that, despite their low individual frequencies (i.e. < 1%), they collectively account for approximately 10% of ID and ASD cases (Grayton *et al.*, 2012; Vissers *et al.*, 2016) (a review of neurodevelopmental CNVs is available in Nevado *et al.*,

2014). Furthermore, these CNVs are moderately or highly penetrant as clinical abnormalities manifest in between 10% and 53% of carriers, hence making them clinically significant (Rosenfeld *et al*, 2013). Intriguingly, the same neurodevelopmental CNVs are associated with multiple conditions, pointing to an overlapping NDD aetiology (**Table 1.2**) (Grayton *et al.*, 2012; Gershon and Grennan, 2015; Torres *et al.*, 2016). Accordingly, disruption of common biological pathways globally predisposes towards NDDs, although the exact pathogenic mechanisms may vary between disorders. This scenario is classically exemplified using the 16p11.2 locus, whose microdeletion and reciprocal microduplication are associated with autism and autism or schizophrenia, respectively (Weiss *et al.*, 2008; McCarthy *et al.*, 2009; Stefansson *et al.*, 2014).

Although the mechanisms of CNV formation go beyond the scope of this thesis and are reviewed in Hastings *et al.*, 2009, we conclude this section by discussing the presumed mechanisms of pathogenicity (Hastings *et al.*, 2009). The most likely consequence of CNVs is gene dosage change as the loss, reduction or increase in protein quantity inevitably impacts molecular processes either directly or indirectly by affecting homeostasis (i.e. the balance of various protein levels) (Lupski and Stankiewicz, 2005). Gene dosage change itself results either from whole gene duplications and deletions (i.e. HIS genes for deletions) or from gene interruptions, which can induce nonsense-mediated messenger RNA decay or gene fusions. Alternatively, CNVs can have positional and transvection effects, whereby

they do not overlap genes but alter the dosage of regulatory regions, which exert their effects on genes located on the same or different chromosomes, respectively (Lupski and Stankiewicz, 2005). Moreover, the observation that many pathogenic CNVs overlap multiple genes raises the inherent question of the contribution of each gene to the underlying pathogenic mechanisms. While it has long been acknowledged that genes are functionally clustered in the human genome (Lee and Sonnhammer, 2003; Makino and McLysaght, 2008; Al-Shahrour *et al*. 2010), two recent studies showed that *dn* CNVs from patients with ASD and DD affect functionally-related genes (Noh *et al.*, 2013; Andrews *et al.*, 2015a). These studies suggest that copy number changes of genes from the same CNVs combinatorially contribute to disease pathogenesis via additive or epistatic (Nei, 1967; Makino and McLysaght, 2008) effects. The existence of epistatic effects is further supported by a recent study, which found that the simultaneous knock-down of orthologues of genes mapping to the same ASD-associated CNV, was necessary to cause neurological phenotypes in fly models (Grice *et al.*, 2015).

Table 1.2 | Neurodevelopmental CNVs recurrently associated with ASD and their pleiotropic effects.

Locus	Direction	Approximate penetrance (rate of ASD in carriers; %)	Neuropsychiatric pleiotropy	Somatic pleiotropy
1q21.1	Deletion	8	ID, ADHD, schizophrenia	Microcephaly, heart defect, eye abnormalities, short stature, epilepsy
1q21.1	Duplication	36	ID, schizophrenia	Epilepsy, macrocephaly, heart defect
2q23.1	Deletion	100	ID, ADHD, language disorder, motor delay	Epilepsy, obesity, brachycephaly, microcephaly, short stature
2q37	Deletion	25-42	ID, ADHD	Epilepsy, short stature, obesity, heart defect
3q29	Deletion	27	ID, speech delay, language disorder, anxiety, schizophrenia, bipolar disorder	Gastrointestinal problems, heart defect, feeding problems, recurrent ear infections, abnormal dentition
5q14.3	Deletion	43	ID, absent speech	Epilepsy, capillary malformation
7q11.23	Duplication	41	ID, ADHD, anxiety disorders, oppositional defiant disorders, speech delay	Epilepsy, macrocephaly, brachycephaly, dilatation of aorta, patent ductus arteriosus, chronic obstipation, kidney abnormalities
8p23	Deletion	Unknown	ID, ADHD	Heart defect, congenital diaphragmatic hernia
15q11–q13	Duplication	69	ID, ADHD	Epilepsy, heart defect, muscle hypotonia, short stature
15q11.2	Deletion	32	ID, ADHD, OCD, speech delay	Epilepsy, ataxia, heart defect
15q11.2	Duplication	43	ID, ADHD, speech delay	Epilepsy, ataxia, hypotonia
15q13.2–q13.3	Duplication	80	ID, speech delay	Epilepsy, urogenital anomalies, recurrent infections
15q13.2–q13.3	Deletion	60	ID, ADHD	None reported
16p11.2	Deletion	15	ID	Epilepsy, hypotonia, sacral dimples, speech articulation problems
16p11.2	Duplication	Unknown	schizophrenia, bipolar disorder	Epilepsy, hypotonia, tremor, ataxia, sacral dimples, speech articulation problems
16p13.11	Deletion	25	ADHD, speech delay	Epilepsy
17p11.2	Deletion	Unknown	None reported	Epilepsy
17q12	Deletion	Unknown	schizophrenia	Macrocephaly, renal anomalies
22q11.2	Deletion	30	Schizophrenia, ADHD, speech delay, anxiety disorders	Heart defect, palate abnormalities, hypocalcaemia, feeding difficulties, recurrent infections (among others)
22q11.2	Duplication	18	ID, ADHD	Heart defect, hearing loss, urogenital anomalies, palate abnormality
Del22q13.3	Deletion	>50	ID, language disorder	Epilepsy, heart defect, renal anomalies, strabismus

(Adapted from Vorstman *et al.*, 2017).

1.2.3 Gene studies of NDDs

Rare neurological syndromes can be caused by mutations in single genes, such as *FMR1*, *MECP2* and *ABCD1*, associated with Fragile X syndrome, Rett syndrome or adrenoleukodystrophy, respectively (for reviews see Wall *et al.*, 2009, Gauthier and Rouleau, 2012 and Huang *et al.*, 2014). Since, many other genes have been linked to broader NDDs (**Table 1.1**) and the development of NGS technologies is enabling the discovery of increasing numbers of susceptibility genes. Indeed, Whole-Exome Sequencing (WES) studies (**Box 1**), using case-control or simplex (i.e. with one affected member) and multiplex family designs, uncovered many genes more frequently

disrupted by *dn* SNVs in NDD patients compared to unaffected individuals (ASD: O'Roak *et al.*, 2011; Chahrour *et al.*, 2012; Iossifov *et al.*, 2012; Neale *et al.*, 2012; O'Roak *et al*., 2012; Sanders *et al.*, 2012; Lim *et al.*, 2013; Yu *et al.*, 2013; Cukier *et al.*, 2014; De Rubeis *et al.*, 2014; Iossifov *et al.*, 2014; Toma *et al.*, 2014; Krumm *et al.*, 2015; Stessman *et al.*, 2017; Yuen *et al.*, 2017; ID: Hamdan *et al.*, 2011; Najmabadi *et al.*, 2011; de Ligt *et al.*, 2012; Rauch *et al.*, 2012; Alazami *et al.*, 2014; Gilissen *et al.*, 2014; Hamdan *et al.*, 2014; Hu *et al.*, 2016; Riazuddin *et al.*, 2016; Rump *et al.*, 2016; Stessman *et al.*, 2017; epilepsy: Epi4K and EPGP Investigators, 2013; Kodera *et al.*, 2013; EuroEPINOMICS-RES Consortium *et al.*, 2014; Helbig *et al.*, 2016; EuroEPINOMICS-RES Consortium *et al.*, 2017; ADHD: Lyon *et al.*, 2011; Demontis *et al.*, 2016; OCD: Cappi *et al.*, 2016).

Similarly to CNVs, SNVs – detectable by current study designs – are associated with a moderate or high disease risk and are individually rare although they collectively explain a large fraction of genetic variance, estimated at 25% in ASD (Geschwind and State, 2015). Furthermore, conforming to their presumed higher penetrance compared to inherited variants, *dn* SNVs represent an appreciable proportion of pathogenic variants, accounting for as much as 30% of ASD and ID cases (Ronemus *et al.*, 2014; Vissers *et al.*, 2016; Yuen *et al.*, 2017). Specifically, *dn* LoF and missense (and collectively LGD) SNVs, which occur significantly more often in NDDs than in the healthy population (Iossifov *et al.*, 2014; Hoischen *et al.*, 2014; Shohat *et al.*, 2017; Stessman *et al.*, 2017), represent deleterious mutational classes. As for CNVs, the same genes are disrupted across NDDs (Shohat *et*

al., 2017; Stessman *et al.*, 2017) and predispose to a large variety of clinical features, lacking respect towards diagnostic categories (**Table 1.1**). Taken together, these findings are consistent with the rare allele model, wherein NDDs can be caused both by multiple rare inherited or single *dn* mutations, hence reconciling the high heritability of such disorders with the large number of sporadic cases.

1.2.4 Integrative models of NDDs

In addition to patients, NDD-associated mutations also occur in healthy individuals, indicating that additional layers of complexity are necessary to describe the genotype and phenotype relationships in disease. The rejection of Mendelian genetic models of NDDs stems from the dual observation that (1) not all carriers of a particular susceptibility gene or variant display an apparent abnormality (i.e. incomplete penetrance) and that (2) among those presenting with a phenotype, the mutation is not always expressed similarly and covers a large clinical spectrum (i.e. variable expressivity).

Several theoretical considerations must be mentioned when discussing these notions. Firstly, estimates of penetrance vary upon interpretation, such that many neurodevelopmental CNVs are associated with higher penetrance across NDDs compared to individual disorders (Vassos *et al.*, 2010; Shishido *et al.*, 2013; Kirov, 2015). Furthermore, variants may predominantly exert their effect on endophenotypes rather than disease (Stefansson *et al.*, 2014). Secondly, the notion of variable expressivity relies on the initial phenotypic

description of individuals carrying a variant of interest. For example, 22q11.2 microdeletions, initially described in relation to the Velo-Cardio-Facial syndrome, were further implicated in schizophrenia (Grayton *et al.*, 2012). Hence, this CNV may underlie common pathogenic mechanisms between the two conditions, rather than show variable expressivity. Nevertheless, the strong pleiotropic effects described for many neurodevelopmental CNVs and genes suggest this ascertainment bias cannot alone explain their strong association with multiple NDDs (Geschwind, 2011; Grayton *et al.*, 2012). Lastly, comparisons across all gene variants may be inappropriate as phenotypic outcomes are inherently determined by the functional impacts of the associated mutations (Cooper *et al.*, 2013; Walsh and Engle, 2010). For example, the same gene may be targeted by mutations of differing functional impacts in cases and controls, as suggested by the excess of LoF, but not synonymous, *dn* SNVs in ASD patients compared to their unaffected siblings (Iossifov *et al.*, 2012; Iossifov *et al.*, 2014). Nevertheless, given the occurrence of neurodevelopmental CNVs with the same breakpoints in affected and healthy individuals, we conclude allelic heterogeneity alone is unlikely to explain the observed incomplete penetrance and variable expressivity.

Oligogenic models of disease propose that disease is caused by several RVs, whose additive, epistatic or non-additive interactions underlie phenotypic heterogeneity (**Fig 1.1**). Such models have gained recognition in NDDs as multiple gene mutations and CNVs are uncovered in increasing numbers of patients (Jiang *et al.*, 2004; Girirajan *et al.*, 2010; Schaaf *et al.*, 2011; Girirajan

et al., 2012; Leblond *et al.*, 2012; Steinberg and Webber, 2013; Pinto *et al.*, 2014). Furthermore, in line with current views of variant pathogenicity, secondary mutational events are primarily observed for inherited variants, while *dn* variants can presumably solely induce clinical outcomes due to their higher penetrance (Girirajan *et al.*, 2012; Steinberg and Webber, 2013; Pinto *et al.*, 2014).

The presence of genetic modifiers modulating the effects of primary mutations represents a second type of integrative model (**Fig 1.1**). Studies of various Mendelian disorders and monogenic neurological syndromes, such as Rett syndrome (Grillo *et al.*, 2013) and Down syndrome (Del Hoyo *et al.*, 2016), revealed the phenotypic effects of causal mutations can be modified by additional rare or common variants of negligible individual consequences (Cooper *et al.*, 2013).

Such modifying effects can also occur in oligogenic conditions, wherein the combinations of penetrant mutations cannot collectively explain the observed clinical spectrum. This scenario is typically exemplified using the Bradet-Biedl syndrome (Badano *et al.*, 2006), where additional CVs modulate phenotypic variability.

The incomplete concordance of monozygotic twins (i.e. < 100%) indicates that non-genetic components also contribute to the aetiology of NDDs. The most common sources of non-genetic variation include environmental risk factors, such as maternal lifestyle (e.g. alcohol consumption and smoking status),

maternal infections, immune deficiencies and perinatal complications (Hamlyn *et al.*, 2013; Lyall *et al.*, 2014). More generally, the birth season and place have also been linked to NDDs, based on the observation that schizophrenia, bipolar disorder, ASD and major depression are more likely to occur in individuals born in winter or spring and in urban compared to rural areas (Torrey *et al.*, 1996; Torrey *et al.*, 1997; Mortensen *et al.*, 1999). Although such environmental risks can be strongly associated with particular disorders (e.g. alcohol consumption during pregnancy and foetal alcohol syndrome), they typically display low penetrance, suggesting genetic factors modulate the effects and expressivity of such risks (**Fig 1.2d**). Epigenetic and stochastic developmental variation represent additional types of non-genetic risk factors and reflect the neurodevelopmental events leading to the differences between genetically-identical monozygotic twins. Except for the study of maternal autoimmunity and gestational age in ASD, we mostly neglect the existence of non-genetic components in the aetiology of NDDs throughout this thesis.

1.2.5 Conclusions

The genetic architecture of NDDs is multifactorial ranging from single gene disorders to integrative models involving the interplay between inherited, *dn*, rare and common variants. Furthermore, the genetic overlap between disorders as well as the incomplete penetrance and variable expressivity of individual mutations reflect the tremendous level of locus and allelic heterogeneity. As a result, a great diversity of mutations and genes has been implicated in NDDs, inherently adding layers of complexity to their aetiological

understanding. The advent of NGS technologies and their decreasing cost is leading to the identification of increasing numbers of high-risk mutations, hence shaping the genetic landscape of NDDs. Comprehensive analyses of such variants using computational and statistical approaches as well as cellular and animal models of disease, coupled to our growing understanding of brain functioning, will undoubtedly elucidate the pathogenic mechanisms driving the clinical outcomes observed NDD patients.

1.3 Biological insights from gene, cellular and animal studies

Over the past decade, linkage, CNV and exome studies have strongly advanced our understanding of the genetic architecture of NDDs, by accelerating the discovery of disease-associated mutations. Genetic and functional studies of genes targeted by such variants revealed that many are involved in neuronal and synaptic development and function, suggesting that dysfunction of such processes contributes to NDD pathogenesis. This section reviews these processes, the evidence in support for their involvement in the aetiology of NDDs and the importance of cellular and animal models in elucidating the underlying pathogenic mechanisms.

1.3.1 Neurodevelopment and synaptic function

A healthy adult brain requires the differentiation of approximately 100 billion neurons and formation of up to 1000 trillion synaptic connections between them (Herculano-Houzel, 2009). Cortical neurogenesis, defined as the

production of cells destined to form the human brain cortex, starts three weeks post-conception and ends by five months post-conception. Neurons, then, start migrating from the subventricular zone, where their progenitors proliferated, towards the cortical plate acquiring more precise differentiation fates until reaching their final destination. There, they mature by growing dendrites, essential for neurotransmission, and axons to enable synapse formation.

While dendritic growth extends over several years, axonal growth is fast and guided by attractive and repulsive extracellular molecular cues, specifically recognised or secreted by cell surface receptors. Binding of such guidance molecules to receptors triggers downstream signalling pathways, which control the cytoskeletal dynamics of the neuronal growth cone, thus determining axonal trajectories and the target innervations regions, where synapses are formed (Dickson, 2002). A variety of guidance cues have been described, such as netrins and slits, which are involved in the extension of commissural axons, and are reviewed in Dickson, 2002.

Synaptic formation relies on the recognition and adhesion between pre- and post-synaptic cells and the construction of the synaptic structure. These processes are mainly driven by homophilic and heterophilic adhesion molecules, such as immunoglobulins, cadherins, leucine-rich repeats and neurexins (reviewed in Missler *et al.*, 2012), expressed by the two synaptic partners. Specifically, when matched these molecules enable cell adhesion and initiate the functional changes required to construct the synaptic structure.

They recruit cytoplasmic signalling proteins, such as scaffolding proteins, actin cytoskeleton regulators and other transduction pathways, which drive the differentiation and specialisation of pre- and post-synaptic functions.

The number of synapses formed culminates before the age of two years and is followed by a period of refinement, during which biochemical signals and experiences instigate processes of cell death and synaptic pruning to eliminate nonessential cells and connections. Finally, the last stage of brain development, extending until adulthood, involves the production of myelin sheaths around axons to improve neuronal function by enabling rapid and efficient signal propagation.

Throughout adulthood, on-going synaptic plasticity enhances the brain's ability to adjust and adapt to new information, by dynamically modulating the strength of synapses in time. Several types of synaptic plasticities have been described. Firstly, (1) long-term plasticity induces long-lasting changes in synaptic strength and mediates memory and learning. Secondly, (2) short-term plasticity involves transient changes in synaptic strength to enable short-term memory and adaptation to sensory inputs. Lastly, (3) metaplasticity, which refers to the plasticity of synaptic plasticity, controls the ability of synapses to weaken and strengthen over time as a result of previous adaptation events (Abraham, 2008; Citri and Malenka, 2008). Mechanisms of synaptic plasticity rely on the detection of activity levels by postsynaptic receptors, such as ion channels and neurotransmitter receptors, including N-Methyl-D-Aspartate (NMDA) receptors. The flow of ions across these

receptors controls cellular signal transduction pathways regulating (i) messenger RNA translation and the expression of (ii) synaptic genes and (iii) transcription factors and chromatin modifiers. The latter category, in turn, controls the broader expression of genes involved in processes of cell survival, dendritic and axonal growth, synaptic development, and neuronal plasticity (Greer and Greenberg, 2008).

1.3.2 NDD genes affect neurodevelopmental and synaptic processes

Many genes associated with the above-described processes, namely neurodevelopment and synaptic function, have been implicated in NDDs. A detailed review of disease-associated genes and gene classes is available in **section S1**.

Post-mortem neuropathological studies revealed a wide range of cortical abnormalities in ASD brains, including alterations in neuronal size, number, position and emplacement, dendritic abnormalities and reductions in white matter tracts (de la Torre-Ubieta *et al.*, 2016). These observations suggest that dysregulation of foetal cortical development plays an important role in NDD pathogenesis and are supported by evidence at the genetic level. Specifically, genes involved in various processes of neurogenesis and neuronal migration have been uncovered in NDDs, including cell cycle control and mitosis, transcriptional and chromatin regulators, cytoskeletal remodellers and extracellular matrix components (**Table 1.3 and section S1**).

Neuropathological evidence of increased spine density, decreased number of inhibitory GABA-ergic neuronal receptors and downregulation of interneuron markers in post-mortem ASD brains further implicates synaptic dysfunction in disease aetiology (de la Torre-Ubieta *et al.*, 2016). In fact, CNV and exome studies of NDDs uncovered genes participating in many processes of synaptic formation and plasticity, including cell adhesion molecules, scaffolding proteins, ion channels and NMDA receptors, signal transduction components, transcription factors and chromatin modifiers (**Table 1.3 and section S1**).

Table 1.3 | Functions and genes implicated in NDDs.

Gene	Gene Name	Band	Function	Disorder
			NEURONAL PROLIFERATION AND DIFFERENTIATION	
ASPM	Abnormal spindle-like microcephaly-associated protein	1q31.3	Spindle pole protein regulating the mitotic spidle orientation during embryonic neurogenesis	Autosomal recessive primary microcephaly, ID, ASD
CDK5RAP2	Cyclin-dependent kinase 5 regulatory subunit-associated associated protein 2	9q33.2	Pericentriolar material protein involved in the maintainance of centriole engagement and cohesion, thus regulating the mitotic spindle orientation; involved spindle checkpoint activation and microtubule polymerisation	Autosomal recessive primary microcephaly, ID, ASD
CHD8	Chromodomain Helicase DNA Binding Protein 8	14q11.2	ATP-dependent chromatin helicase negatively regulating Wnt signaling pathway; involved in transcriptional regulation, epigenetic remodeling, promotion of cell proliferation, and regulation of RNA synthesis	ASD with accompanying macrocephaly, CHARGE syndrome (multisystemic disorder associated with growth retardation, heart defects, eye, ear and choanal
CENPJ	Centromere protein J	13q12.13	Involved in centrosome biogenesis by controlling centriole length and microtubule function	Autosomal recessive primary microcephaly, Seckel syndrome, ID
CEP135	Centrosomal protein 135 kDa	4q12	Involved in centriole biogenesis and maintenance of centrosome organisation and structure	Autosomal recessive primary microcephaly, ID
CEP152	Centrosomal protein 152 kDa	15q21.1	Centrosomal protein involved in centriole duplication, polarity and motility	Autosomal recessive primary microcephaly, Seckel syndrome, ID
KDM5C	Lysine Demethylase 5C	Xp11.22	Histone demethylase involved in regulation of transcription and chromatin remodeling; acts as a transcriptional repressor of neuronal genes	X-linked intellectual disability, ASD, microcephaly
MCPH1	Microcephalin 1	8p23.1	DNA repair; coupling of the centrosome cycle with mitotic entry	Autosomal recessive primary microcephaly, ID
SMARCC1/ SMARCC2	SWI/SNF Related, Matrix Associated, Actin Dependent Regulator Of Chromatin Subfamily C Member 1/2	3p21.31 / 12q13.2	Component of the neural progenitors-specific chromatin remodeling complex (npBAF complex) and the neuron-specific chromatin remodeling complex (nBAF complex); the transition from proliferating neural stem/progenitor cells to postmitotic neurons requires a switch between npBAF to nBAF	ASD
SMC1A	Structural Maintenance Of Chromosomes 1A	Xp11.2	Component of the cohesin complex, involved in chromosome cohesion during cell cycle and in DNA repair	Cornelia de Lange syndrome with variable phenotpe including ID and ASD-like features
STIL	SCL/TAL1 interrupting locus	1q33	Centrosome-associated protein involved in mitotic spindle checkpoing and regulating apoptosis regulator and cell cycle progression	Autosomal recessive primary microcephaly, ID
WDR62	WD repeat-containing protein 62	19q13.12	Spindle pole protein involved in centriole duplication during cerebral cortical development	Autosomal recessive primary microcephaly, ID
			NEURONAL MIGRATION	
AUTS2	Autism susceptibility candidate 2	7q11.22	Cell proliferation and neuronal cell survival in the developing brain via trannscriptional regulation	ASD, ID, ADHD, speech disorders, DD, microcephaly, epilsepsy and schizophrenia
ASTN2	Astrotactin 1	1q25.2	Involved in neuronal migration by enabling adhesion between neurons and astroglia	ASD, ADHD and OCD
DCX	Doublecortin	Xq23	Regulator of microtubule organisation, stability and function during neuronal migration	X-linked lissencephaly
DLX 1/2	Distal-less homeobox 1/2	2q31.1	Transcription factors crucial for tangential migration of GABAergic neurons from the MGE to the cortex	ASD
LIS1	Platelet Activating Factor Acetylhydrolase 1b Regulatory Subunit 1	17p13.3	Part of LIS1/NDE1/NDEL1/cytoplasmic dynein complex, particpating in neuronal proliferation and migration and involved in control of mitotic spindle orientation	lissencephaly, Miller-Dieker syndrome
NDE1	NudE nuclear distribution E homolog 1	16p13.11	Part of LIS1/NDE1/NDEL1/cytoplasmic dynein complex, particpating in neuronal proliferation and migration	ASD, ID, ADHD, epilepsy, speech and language delay, schizopohrenia, microlissencephaly, microhydraencephaly
TUBA1A	Tubulin Alpha 1a	12q13.2	Cytoskeleton component with crucial roles in all eukaryotic cells; involved in neuronal proliferation, migration, differentiation and trafficking	Microcephaly, lissencephaly, polymicrogyria, agenesis of corpus callosum, pachygyria and other abnormalities
TUBA8	Tubulin Alpha 8	22q11.21		Microcephaly, polymicrogyria, agenesis of corpus callosus, seizures and other abnormalities
TUBB2B	Tubulin Beta 2B Class IIb	6p25.2		Microcephaly, lissencephaly, polymicrogyria, pachyrgyria, agenesis of corpus callosum, ADHD, seizures, DD, ASD
TUBB3	Tubulin Beta 3 Class III	16q24.3		microcephaly, polymicrogyria, agenesis of corpus callosum, ID, DD, epilepsy and other abnormalities
RELN	Reelin	7q22.1	ECM glycoprotein mediating neuronal cell adhesion during migration	ASD, lissencephaly
WDFY	WD repeat and FYVE domain containing 3	4q21.23	Adaptor protein involved in autophagy, possile role in progenitor proliferation and neurogenesis	ASD

CELL-CELL INTERACTIONS AND SYNAPTIC ADHESION				
CNTNAP2	Contactin-associated protein-like 2	7q35-q36.1	Neurexin family member	Cortical dysplasia-focal epilepsy syndrome, associated with ID and ASD; Deletions or chromosomal rearrangements disrupting a single copy of *CNTNAP2* have been reported in patients with ASD, ID, epilepsy, schizophrenia and bipolar disorder as well as healthy subjects; their clinical significance is unknown
IL1RAPL1	Interleukin 1 receptor accessory protein-like 1	Xp21.2-p21.3	Cell adhesion, synapse formation	X-linked ID and ASD, with or without dysmorphic features
L1CAM	L1 cell adhesion molecule	Xp28	Cell adhesion, synapse formation and development	Syndromic X-linked ID with hydrocephalus, adducted thumbs, spastic paraplegia and hypoplasia of the corpus callosum
LGI1	Leucine-rich glioma inactivated-1	10q21	Secreted LRR protein, synapse formation	Autosomal dominant lateral temporal lob epilepsy and sporadic epilepsy
NLGN3	Neuroligin 3	X13.1	Cell adhesion, inhibitory synapse formation	Non-syndromic X-linked ASD and ID
NLGN4X	Neurologin 4, X-linked	Xp22.32-p22.31	Cell adhesion, synapse formation	Non-syndromic X-linked ASD and/or ID
NRXN1	Neurexin 1	2p16.3	Cell adhesion, synapse formation	Disrupted in ASD, ID, schizophrenia, ADHD, GTS and other NDD and psychiatric disorders; deletions also reported in unaffected parents and controls. Pitt-Hopkins-like syndrome-2
PCDH19	Protocadherin 19	Xp22.1	Cell adhesion, synapse formation	X-linked femaled-limited epilepsy, with or without cognitive impairment and ASD; early infantile epileptic encephalopathies
SLITRK1	SLIT and NTRK-like family, member 1	13q31.1	Cell adhesion, neurite outgrowth	GTS, trichotillomania, OCD
SCAFFOLDING PROTEINS				
CASK	Calcium / calmodulin-dependent serine protein kinase	Xp11.4	Adaptor / scaffold protein, synapse plasticity	X-linked ID and microcephaly with pontine and cerebellar hypoplasia; non-syndromic ID with or without nystagmus and ASD in males
DISC1	Disrupted in schizophrenia 1	1q42.2	Scaffold protein, synapse development	schizophrenia, bipolar disorder and depression
DG3	Discs, large homolog 3	Xq13.1	Synaptic scaffolding (MAGUK)	Non-syndromic X-linked ID
SHANK2	SH3 and multiple ankyrin repeat domains 2	11q13.3	Synaptic scaffolding	ID, ASD
SHANK3	SH3 and multiple ankyrin repeat domains 3	22q13.33	Synaptic scaffolding	Phelan-McDermid syndrome, associated with ID, absence or delay in speech, ASD, hypotonia
RECEPTORS AND CHANNELS				
CACNA1C	Calcium channel, voltage-dependent, L type, alpha 1C subunit	12p13.33	Voltage-gated calcium channel	Timothy syndrome
CHRNA7	Cholinergic receptor, nicotinic, apha 7	15q13.3	Ligand gated ion channel, nicotinic acetylcholine receptor	15q13.3 microdeletion syndrome characterised by a variable phenotype including ID, ASD, seizures, schizophrenia and other psychiatric conditions
GRIA3	Glutamate receptor, ionotropic, AMPA3	Xq25	Ligand gated ion channel, ionotropic glutamate receptor, AMPA	X-linked ASD and ID
GRIN2A	Glutamate receptor, ionotropic, N-methyl D-aspartate 2A	16p13.2	Ligand gated ion channel, ionotropic glutamate receptor, NMDA	Focal epilepsy with speech disorder, with or without ID
GRIN2B	Glutamate receptor, ionotropic, N-methyl D-aspartate 2B	12p13.1	Ligand gated ion channel, ionotropic glutamate receptor, NMDA	ID, ASD, seizures
SCN1A	Sodium channel, voltage-gated, type I, alpha subunit	2q24.3	Voltage-gated sodium channel	Severe myoclonic epilepsy of infancy (Dravet syndrome), ID, ASD
SCN2A	Sodium channel, voltage-gated, type II, alpha subunit	2q24.3	Voltage-gated sodium channel	Early infantile epileptic encephalopathy; benign infantile seizures, ID, ASD

Table 1.3 Continued

SIGNAL TRANSDUCTION, TRANSLATION				
CDKL5	Cyclin-dependent kinase-like 5	Xp22.13	Synapse development and plasticity	Early infantile epileptic encephalopathies, ID, ASD
CUL4B	Cullin 4B	Xq24	Ubiquitin E3 ligase subunitm synapse formation	Syndromic X-linked ID, Cabezas type
FMR1	Fragile X mental retardation 1	Xq27.3	Translation repressor of numerous synaptic proteins	Fragile X syndrome associated with ID, ASD, ADHD and epilepsy
IQSEC2	IQ motif and Sec7 domain 2	Xp11.22	Arf6-GEF, synaptic plasticity	Non-syndromic X-linked ID, ASD, epilepsy
NF1	Neurofibromin 1	17q11.2	MAPK-signalling pathway (Ras-GTPase activator protein)	Neurofibromatosis type I, associated with ID and ASD
PTEN	Phosphatase and tensin homolog	10q23.31	AKT-mTOR signalling pathway, synaptic development, activity-dependent synaptic plasticity	PTEN-hamartoma tumour syndrome; macrocephaly, autism
RAB39B	RAB39B, member RAS oncogene family	Xq28	Synapse formation and maintenance	X-linked ID associated with ASD, epilepsy and macrocephaly
STXBP1	Syntaxin binding protein 1	9q34.11	Synapic vesicle fusion	Early infantile epileptic encephalopathy, nonsyndromic ID
SYN1	Synapsin 1	Xp11.23	synaptic vesicle phosphoprotein, modulates neurotransmitter release, synaptic vesicle trafficking and synaptogenesis	X-linked epilepsy with variable learning disabilities and behaviour disorders
SYNGAP1	Synaptic Ras GTPase activating protein 1	6p21.32	RAS signalling pathway, regulation of glutamate receptor trafficking and synaptic efficacy, maturation of excitatory synapses	Non-syndromic ID, ASD and epilepsy; epileptic encephalopathy
TSC1 / TSC2	Tuberous complex 1/2	9q34.13 / 16p13.3	mTOR pathway inhibitor	Tuberous sclerosis complex, associated with ID, ASD and epilepsy
UBE3A	Ubiquitin protein ligase E3	15q11.2	Ubiquitin ligase; synapse formation	Angelman syndrome associated with ID, ASD lack of speech and seizures; 15q11-q13 duplication syndrome associated with ID, ASD, schizophrenia, language impairment and seizures
TRANSCRIPTION AND CHROMATIN REGULATION				
ARID1A	AT rich interactive domain 1A	1p36.11	Chromatin remodelling	Coffin-Siris syndrome (ID with language impairment and finger abnormalities)
ARID1B	AT rich interactive domain 1B	6q25.3	Neuron-specific chromatin remodelling, synaptic plasticity	ID, speech impairment, ASD, corpus callosum abnormalities, Coffin-Siris syndrome
CREBBP / EP300	CREB binding protein / E1A binding protein p300	16p13.3 / 22q13.2	Transcription modulator	Rubinstein-Taybi syndome (ID with facial and finger features)
EHMT1	euchromatic histone-lysine N-methyltransferase 1	9q34.3	Transcription regulation, synapse development	Kleefstra syndrome / 9q subtelomeric deletion syndrome (ID with microcephaly and facial abnormalities), ASD, schizophrenia
HDAC4	Histone deacetylase 4	2q37.3	Transcription regulation, synapse plasticity	Brachydactyly mental retardation syndrome (2q37 microdeletion syndrome) associated with ID and ASD
MECP2	Methyl-CpG-binding protein 2	Xq28	Transcription factor	Rett syndrome (females); congenital encephalopathy (males), non-syndromic ID (males), ASD
MEF2C	Myocyte enhancer 2C	5q14.3	Regulation of synapse formation and function	5q14.3 microdeletion syndrome, ID, stereotypic movement, epilepsy, cerebral malformations and ASD
SMARCA2 *SMARCA4* *SMARCB1* *SMARCE1*	SWI/SNF-related matrix associated, actin dependent regulator of chromatin, subfamilies a, b and e, member 2, 4 and 1	9q24.3 19p13.2 22q11.23 17q21.1	Neuron-specific chromatin remodelling, synaptic plasticity	Nicholaides-Baraitser syndrome (severe ID with language impairment, microcephaly, seizures, short staturem facial coarseness), Coffin-Siris syndrome, ID, ASD
SRCAP	Snf2-related CREBBP activator protein	16p11.2	Neuron-specific chromatin remodelling, synaptic plasticity	Floatin-Habor syndrome (short stature, laguage delay and facial features), ID, ASD, ADHD
TBR1	T-box, brain-1	2q24.2	Transcription factor, neuronal specification and synaptic plastcity	ID, ASD

Table 1.3 Continued

(Compiled from LaSalle, 2013; Betancur and Mitchell, 2015; Homem *et al.*, 2015; Chakraborti *et al.*, 2016 and Reiner *et al.*, 2016).

1.3.3 Cellular and animal models

Based on epidemiological studies (i.e. gene frequencies in cases compared to control), mutational types (i.e. LoF compared to synonymous), modes of inheritance (i.e. *dn* compared to inherited) and our current knowledge of brain development and functions, it is possible to implicate genes in the aetiology of NDDs. Nevertheless, the understanding of the underlying pathogenic mechanisms requires molecular, cellular, circuit and system characterisation of the consequences of the evaluated mutations. In this respect, cellular and animal models provide powerful research tools.

Traditionally, cellular *in vitro* models enabled the study of specific molecular processes and interactions, including cell division, cell migration or synaptic function. As a result, they expanded our understanding of the fundamental biological pathways governing growth and cognition and how they may be involved in disease. More recently, human induced pluripotent stem cells have emerged as powerful cellular systems. These models involve the examination of cells derived from patients, which have been re-programmed to the cell types of interest (e.g. neurons), hence enabling the direct investigation of the pathogenic mechanisms underlying the observed symptomatic presentation. Human stem cell models of strong ASD candidate genes, such as *CDH8*, *SHANK3*, *MECP2* and *CACNA1C*, have been created and point to the dysregulation of neurogenesis, neuromorphogenesis and synaptic function in disease (de la Torre-Ubieta *et al.*, 2016). Stem cell models can also be used to test the efficacy of various compounds in reversing the phenotypes

observed, although it is unclear how *in vitro* models translate *in vivo*. Other methodological limitations of stem cell-based technologies involve the inability to create mature non-foetal neurons and reproduce the circuit-level cellular organisation. Ardhanareeswaran *et al.*, 2017 provides a detailed review of the use of stem cell technologies in NDDs.

Animals represent powerful models to investigate gene function as they enable the interrogation of large numbers of subjects and the use of controlled experiments, tailored to the underlying hypothesis. Under the *orthologue conjecture*, decidedly orthologous genes (i.e. present at 1:1 ratios between species) are inclined to share function suggesting that the roles of human genes can be extrapolated from animal studies (conditional on the existence of orthologous genes).

The use of animal models overcomes certain limitations associated with human studies. These include (1) environmental factors, as experiments are performed in the same controlled and well-characterised environment, and (2) genetic background, as most research uses genetically identical and homozygous laboratory animal strains. Furthermore, animal models are particularly appealing for the study of human disease and NDDs as (3) they enable the use of invasive approaches (e.g. brain electrophysiological recordings used to study epilepsy), and the extraction of brain developmental tissue. Lastly, (4) animal models can be used in drug discovery research to assess the efficacy and safety of various compounds.

Based on the differences in genetics, anatomy and physiology between species, three criteria are used to assess the relevance of model systems. Firstly, (1) face validity evaluates the ability of the model to recapitulate human phenotypes. Secondly, (2) construct validity examines the ability of the model to recapitulate the human disease-causing dysfunction (e.g. gene mutation or physical abnormality). Thirdly, (3) predictive validity tests whether the model responds similarly to humans to a specific treatment. Effective mouse models are expected to incorporate all three criteria, such that many successful models of NDD-associated genes have been constructed so far (Silverman *et al.*, 2010).

Mus musculus, simply referred to as mouse, is the preferred model organism and transgenic mice encompassing various NDD susceptibility genes have been constructed (Blake *et al.*, 2017) (the targeting mechanisms to produce gene knockins and KnockOuts (KOs) are reviewed in Doyle *et al.*, 2013). Phenotypic assessments of these KO models, facilitated by the development of well-characterised behavioural assays (Silverman *et al.*, 2010), have been particularly valuable in validating the role of many genes in the aetiology of NDDs. Furthermore, they expanded our understanding the underlying molecular, cellular and circuit-level pathogenic mechanisms (i.e. many described in **Table 1.3** are supported by mouse studies).

Many successful mouse models of ASD have been generated so far and recapitulate the disease-associated social impairments and repetitive behaviours. For example, the targeted disruption of the mouse orthologues of

the *SCN1A* (i.e. *Scn1a*) and *CNTNAP2* (i.e. *Cntnap2*) genes, encoding a sodium channel and a neurexin-like adhesion protein, respectively, leads to neuronal abnormalities, impairments in social interaction and grooming, as well as seizures, conforming to the association between *SCN1A* and epilepsy (de la Torre-Ubieta *et al.*, 2016). Another example involves the *CHD8* gene, which encodes a master transcriptional regulator and is considered a major ASD risk gene, with *dn* LoF variants affecting 0.2% of patients (Bourgeron, 2015). Mice with heterozygous *Chd8* mutations present ASD-like symptoms, including abnormal social behaviours, repetitive behaviours and anxiety (Katayama *et al.*, 2016; Platt *et al.*, 2017). Embryos have a delayed neurodevelopment, accompanied by gene expression changes in the brain and particularly downregulation of neuronal genes, as well as genes encoding chromatin remodellers, histone modifiers and cell cycle regulators (Katayama *et al.*, 2016; Platt *et al.*, 2017). Moreover, similarly to humans (Barnard *et al.*, 2015), *Chd8+/-* mice display macrocephaly and craniofacial abnormalities (Platt *et al.*, 2017). Overall, these results support a strong contribution of *CHD8* haploinsufficiency to ASD, implicating neurodevelopmental delay as a mechanism of pathogenicity. Mouse KOs have also been constructed to model disruptions of the *SHANK3* gene, which encodes a synaptic scaffolding molecule associated with the 22q13 microdeletion syndrome (i.e. Phelan-McDermid syndrome), ASD and ID (Wang *et al.*, 2011). Mice with homozygous mutations or *Shank3* KOs display ASD-like symptoms, including impairments in social behaviours and communication and repetitive behaviours (Peça *et al.*, 2011; Wang *et al.*, 2011). Further analyses of these mice revealed striatal dysfunction, abnormalities in long-term hippocampal

potentiation and morphological dendritic spine defects (Peça *et al.*, 2011; Wang *et al.*, 2011). Altogether, these results suggest disruption of the *SHANK3* gene impairs the development of neuronal connectivity.

More generally, efforts have been undertaken to perform large-scale mouse mutagenesis screens coupled to detailed phenotypic assessments in order to interrogate the function of all protein-coding genes and evaluate their contribution to traits and disease. As a result, many consortia have been formed and are summarised in Ayadi *et al.*, 2012 (**Table 1.4**). Comprehensive mouse models are of particular importance to human disease, as they can be used to prioritise and validate susceptibility genes, as well as investigate their functions and underlying pathogenic mechanisms. In fact, the mouse phenotypes ensuing the disruption of 112 ASD-associated orthologues collectively recapitulate the neurobiological and behavioural abnormalities observed in humans, thus supporting the face and construct validity of mouse models (Buxbaum *et al.*, 2012; Gai *et al.*, 2012). The majority of our current knowledge of the genetics, genomics and phenotypes of the laboratory mouse is integrated in a community data resource and can be accessed from the Mouse Genome Informatics database (MGI; http://www.informatics.jax.org) (Blake *et al.*, 2017). As of 2017, the MGI provides phenotypic annotations for over 9,500 human orthologues (Blake *et al.*, 2017).

Based on the observation that many fundamental human molecular pathways are preserved across species, other model organisms can also be used to study NDDs, namely *Caenorhabditis elegans* worms, *Drosophila*

melanogaster fruit flies and *Danio rerio* zebrafish. At the expense of the nervous system complexity, these organisms present several advantages to mice including shorter generation time, lower maintenance costs and simpler genetic manipulations. In particular, *D. melanogaster* models have successfully been used to study the neural circuits underlying rare neurological syndromes, such as Angelman syndrome and Fragile X syndrome (Wu *et al.*, 2008; O'Connor *et al.*, 2017). Furthermore, supporting the face and construct validities of fly models, heterozygous mutants of the *DYRK1A* orthologue display a reduced brain volume, recapitulating the phenotypes observed in mice and humans (i.e. microcephaly) (Hoischen *et al.*, 2014). More recently, Grice *et al.* used single and double fly mutants to provide evidence of oligogenic effects between orthologues of genes from the same neurodevelopmental CNVs (Grice *et al.*, 2015).

Table 1.4 | Large-scale mouse phenotyping initiatives

Consortium	Website
The European Conditional Mouse Mutagenesis Programme (EUCOMM)	www.eucomm.org
The EUCOMMTools: Tools for Functional Annotation of the Mouse Genome (EUCOMMTools)	www.eucommtools.org
The European Mouse Phenotyping Resource of Standardised Screens (EMPRESS)	www.empress.har.mrc.ac.uk
The European Union Mouse Research for Public Health and Industrial Applications (EUMORPHIA)	www.eumorphia.org
The European Mouse Disease Clinic (EUMODIC)	www.EUMODIC.org
The EuroPhenome Mouse Phenotyping Resource	www.europhenome.org
The European Mouse Mutant Archive (EMMA)	www.emmanet.org
The European Infrastructure for Phenotyping and Archiving of model mammalian genomes (INTRAFRONTIER)	www.infrafrontier.eu
The International Knockout Mouse Consortium (IKMC)	www.knockoutmouse.org
The International Mouse Phenotyping Consortium (IMPC)	www.mousephenotype.org
The Knockout Mouse Project (KOMP)	www.komp.org
The Knockout Mouse Phenotyping Programme (KOMP2)	http://commonfund.nih.gov/KOMP2/
The Wellcome Trust Sanger Institute Mouse Genetics Project (SANGER-MGP)	http://www.sanger.ac.uk/mouseportal/

(Adapted from Ayadi *et al.*, 2012).

1.3.3 Conclusion

This section provides an overview of the mutational spectrum of NDDs, wherein disease-associated genes encompass a diverse set of functional classes and affect neuronal proliferation, differentiation and migration as well as synaptic formation and plasticity. We, additionally, highlight the significance of cellular and animal models to study gene function, identify disease genes and investigate the underlying pathogenic mechanisms. The variety of affected molecular functions coupled to the discovery of increasing numbers susceptibility genes highlights the need for comprehensive approaches to enable the simultaneous assessment and prioritisation of disease variants.

1.4 Comprehensive approaches to the study NDDs

The discovery of genes causing well-defined clinical syndromes and high confidence NDD genes coupled to the examination of cellular and animal models has begun unravelling the aetiology of complex disease. In this genomics era, however, more appropriate strategies are necessary to examine the thousands of loci predicted to underlie disease risk (Gratten *et al.*, 2014). While large-scale genetic studies are uncovering ever growing numbers of NDD susceptibility genes (e.g. over 800 and 750 genes are currently associated with ASD and ID, respectively (Basu *et al.*, 2009; Casanova *et al.*, 2016; Kochinke *et al.*, 2016)), their significance often remains obscure. Indeed, (1) the highly variable evidence underpinning individual variants, as a result of their incomplete penetrance and variable

expressivity, challenges their interpretation. Furthermore, (2) the observation that while rare and *dn* SNVs or CNVs are strongly associated with NDDs, their occurrence (and frequency) is not necessarily proof of causation.

The observation that NDD-associated genes recurrently affect the same functional classes (**Table 1.3; section S1**) instigated the hypothesis that genetic variants in NDD patients target genes, which participate in shared molecular processes. Accordingly, the disruption of common biological processes should underlie the development of the same disorders. This view is supported by functional studies of high confidence disease genes, such as *FMR1* (Steinberg and Webber, 2013) or *CHD8* (Cotney *et al.*, 2015), which revealed they often encode proteins regulating or interacting with other risk genes. These findings encouraged a shift in focus from genes to pathways, with the aim of elucidating the underlying pathogenic disease mechanisms. This, in turn, led to the development of methods evaluating how genes operate at circuit, network and system levels. The goal of such approaches is to evaluate whether genes disrupted in patients (1) converge at the functional level, (2) are involved in common biological processes and (3) display signature spatiotemporal expression patterns.

1.4.1 Annotation datasets and functional enrichments

Several databases aggregate our current knowledge of gene function, biological pathways and disease relationships, namely the Gene Ontology (GO) (Gene Ontology Consortium, 2000), the Kyoto Encyclopaedia of Genes

and Genome Elements (KEGG) (Kanehisa and Goto, 2000) and the Online Mendelian Inheritance in Man (OMIM) (McKusick, 1998), respectively.

The aim of Functional Enrichment Analyses (FEAs) is to evaluate the biological themes over-represented among sets of genes using annotations from the aforementioned databases. FEAs of high confidence ASD, ID and ADHD genes implicate nervous system development, synaptic transmission and learning in the aetiology of NDDs (Cristino *et al.*, 2014). Genes targeted by rare or *dn* mutational events in ASD patients are enriched for similar processes including synaptic functions, chromatin remodelling and transcriptional regulation, Wnt and Mitogen-Activated Protein Kinase (MAPK) signalling as well as the broad interaction network of FMRP (Vorstman *et al.*, 2017). Both rare and *dn* point and structural (in particular deletions) variants affect these processes, suggesting that different mutational types target common pathogenic ASD processes (Pinto *et al.*, 2010; Sanders *et al.*, 2011; Iossifov *et al.*, 2012; O'Roak *et al.*, 2012; Ben-David and Shifman, 2013; De Rubeis *et al.*, 2014; Iossifov *et al.*, 2014; Pinto *et al.*, 2014; Toma *et al.*, 2014; Stessman *et al.*, 2017). Moreover, small *dn* CNVs and *dn* LoF SNVs have been shown affect a shared set of genes, enriched for chromatin remodelling and to a lesser extent synaptic functions (Sanders *et al.*, 2015). Although FEAs have not consistently examined sets of genes disrupted in other NDD cohorts, similar enrichments have been observed in epilepsy (i.e. FMRP targets and synaptic plasticity via ion channels) (Epi4K and EPGP Investigators, 2013) and ADHD (i.e. glutamate receptors, ion channels and learning) (Elia *et al.*, 2010; Elia *et al.*, 2011; Lima Lde *et al.*, 2016). As

functional gene studies have implicated the same processes, we can speculate penetrant mutational events observed across NDDs converge on the aforementioned biological pathways. This is further supported by the observation that many variants from patients with ASD, epilepsy and GTS overlap NDD-associated genes (Pinto *et al.*, 2010; Sundaram *et al.*, 2010; Fernandez *et al.*, 2012; Epi4K and EPGP Investigators, 2013; EuroEPINOMICS-RES Consortium *et al.*, 2014; De Rubeis *et al.*, 2014; Iossifov *et al.*, 2014; Pinto *et al.*, 2014; EuroEPINOMICS-RES Consortium *et al.*, 2017)

The MGI database, comprising phenotypic annotations for mouse KOs, can be used to predict the human abnormalities arising from the disruption of orthologous genes. Accordingly, FEAs using MGI annotations of human orthologues revealed that genes mapping to *dn* CNVs and SNVs from ADHD and ASD patients converge onto neurological and behavioural phenotypes, consistent with their neurodevelopmental nature (Noh *et al.*, 2013; De Rubeis *et al.*, 2014; Taylor *et al.*, 2015).

1.4.2 Transcriptomic analyses of NDDs

The systematic evaluation of molecular gene signatures has been enabled by recent large-scale collaborative efforts, which led to the development of spatial and temporal brain and whole body transcriptome maps by the Allen Brain Institute and Brainspan (Sunkin *et al.*, 2013) and the Genotype-Tissue Expression (GTEx) project (Lonsdale *et al.*, 2013), respectively. Using such

datasets, gene co-expression networks have been constructed to reflect the similarity of expression patterns between pairs of genes (**Box 3**). Under the dual assumption that (1) shared patient symptoms arise from the disruption of the same biological pathways and that (2) genes simultaneously expressed in the same tissues are more likely to work in common functional processes, these networks can be used to infer functional gene couplings (**Table 1.5**).

Strong co-expression across the brain has been observed among genes mapping to CNVs from ADHD patients and whose orthologues are associated with hyperactivity phenotypes in mouse models (Taylor *et al.*, 2015). This suggests that disease susceptibility genes are highly interconnected (i.e. cluster) in such networks and can thus be used to predict new susceptibility genes based on the *Guilt By Association* principle (Wolfe *et al.*, 2005). Using the GBA principle, Krishnan *et al.* recently developed a machine learning approach to uncover new ASD risk genes by prioritising those with similar brain expression patterns to ASD-associated genes (Krishnan *et al.*, 2016). The validity of this model is demonstrated by the excess of top-ranked genes (1) previously implicated in NDDs, (2) associated with NDD themes (i.e. postsynaptic density, FMRP and CHD8 target genes) and (3) affected by *dn* LGD variants in ASD patients but not their unaffected siblings (Krishnan *et al.*, 2016). Furthermore, as many top-ranked genes map to large patient CNVs, this approach can also be used to predict causal CNV genes. Lastly, examination of brain spatiotemporal gene expression dynamics revealed top-ranked genes are strongly expressed during prenatal development (Krishnan *et al.*, 2016). These results corroborate previous findings, which showed that

ASD susceptibility genes converge onto mid-foetal layer 5/6 cortical projection neurons (Willsey *et al.*, 2013). More recently, Huynh and Hormozdiari developed a framework to rank genes affected by *dn* LoF variants based on their frequency in ASD and co-expression with other genes disrupted in patients (Huynh and Hormozdiari, 2017). They uncovered 391 genes likely to be associated with NDDs and showed these were affected by *dn* LoF and missense mutations in independent sets of DD, ID and ASD patients (Huynh and Hormozdiari, 2017). Moreover, this methodology was used to construct a disease prediction model, able to accurately detect ASD (i.e. negligible false positive rate) among individuals with *dn* LoF variants, although the associated false negative rate was high (Huynh and Hormozdiari, 2017).

Weighted Gene Co-expression Network Analysis (WGCNA) is a complementary approach to network analysis, which identifies groups (termed *modules*) of co-expressed genes, presumed to encompass distinct pathogenic processes. WGCNA of ASD susceptibility genes and those targeted by *dn* LoF SNVs in patients uncovered two major classes, namely (1) synaptic genes upregulated in the postnatal brain and (2) embryonically-expressed transcriptional regulators (Ben-David and Shifman, 2013; Parikshak *et al.*, 2013). These subpopulations are recapitulated when applying WGCNA to genes targeted by FMRP, such that *dn* LoF SNVs primarily affect embryonic genes while inherited variants target synaptic genes (Steinberg and Webber, 2013). More generally, *dn* LoF and missense variants in ASD and ID have been shown to predominantly affect embryonic chromatin remodellers and synaptic genes, respectively (Shohat *et al.*, 2017). These observations

suggest that variants of differing severities underlie distinct pathogenic disease mechanisms. In addition to WGCNA, other community detection methods, with similar performances, can be used. For example, hierarchical clustering of ASD risk genes based on their co-expression patterns, or k-nearest means algorithms identified analogous embryonic and postnatal gene modules (Chang *et al.*, 2015; Mahfouz *et al.*, 2015; Krishnan *et al.*, 2016; Yuen *et al.*, 2017).

RNA sequencing of post-mortem case and control brains provides a powerful approach to evaluate transcriptional disease signatures and identify genes displaying altered expression dynamics. Differential expression analysis coupled with WGCNA revealed that genes upregulated or downregulated in ASD compared to control brains are enriched for synaptic functions but also for immune and inflammatory processes and glial markers, implicating immunity in ASD pathogenesis (Purcell *et al.*, 2001; Garbett *et al.*, 2008; Voineagu *et al.*, 2011; Gupta *et al.*, 2014).

Box 3 | Gene networks and graph theory

Functional associations between genes or their protein products can be modelled using networks, in which nodes represent genes and edges (also termed links) describe some type of relationship between them. Edge specification defines the type of relationship evaluated, such that different types of gene networks can be constructed including:

• Co-expression network, wherein edges depict the similarity of expression patterns of pairs of genes (i.e. nodes).

• Protein-protein interaction (PPI) networks, wherein edges depict the presence of physical interactions between the proteins encoded by pairs of genes.

• Functional networks, wherein edges depict functional gene similarities. Functional relatedness can be defined by the similarity of molecular processes and pathways pairs of genes participate in or by the similarity of phenotypes resulting from gene disruptions.

• Integrative networks, wherein edges represent functional gene connections, as suggested by multiple lines of evidence (eg: combination of the aforementioned edges).

Such gene networks present several properties:

• Degree: estimate of network centrality of each node, measured as the number of, or cumulative weight of direct interactors associated with that node.

• Shortest path: path between two nodes, measured as the maximum or minimum sum of constitutive edge weights, when edges depict similarity and distance, respectively.

• Betweenness centrality: estimate of network centrality of each node, measured as the number of shortest paths passing through that node.

• Clustering coefficient: measure of how grouped nodes from a network are. More generally clustering can be used to evaluate to interconnectedness of a set of genes (i.e. nodes) by measuring the sum of all direct edge weights between them.

• Modularity: structural measure depicting how cohesive a network is. Modularity examines the extent to which the network's nodes can be separated into groups, by estimating the proportion of edges connecting inside the groups compared to connecting between groups. A modular network is divisible into groups of highly interconnected nodes, called modules. Various algorithms, including the Louvain algorithm (Blondel *et al.*, 2008), have been developed to optimise the partitioning of a network, by aiming to maximise its modularity.

1.4.3 Protein-protein interaction

Over the past decade, many physical protein interactions datasets (i.e. Protein-Protein Interactions; PPIs) have been compiled based on (1) low throughput experiments, such as co-immunoprecipitation, (2) high throughput, yeast two-hybrid screens and (3) computational predictions (Salwinski *et al.*, 2004; McDowall *et al.*, 2009; Kerrien *et al.*, 2012; Licata *et al.*, 2012; Orchard *et al.*, 2012; Schaefer *et al.*, 2012; Chatr-Aryamontri *et al.*, 2013; Zhang *et al.*, 2013; Baspinar *et al.*, 2014; Orchard *et al.*, 2014; Rolland *et al.*, 2014; Huttlin *et al.*, 2015; Szklarczyk *et al.*, 2015; Li *et al.*, 2017). While the rate of true positive is highest for PPIs with low throughput experimental evidence, such datasets contain ascertainment biases and are often incomplete. Conversely, high throughput genome-wide screens overcome these limitations, although they generate large numbers of false positives (**Table 1.5**).

PPI studies of ASD, ID, ADHD and schizophrenia susceptibility genes revealed they encode physically interacting proteins, which form distinct clusters encompassing NDD-related functions, such as synaptic transmission, transcriptional regulation, cell cycle, cell adhesion and communication (Cristino *et al.*, 2014). Such foundational studies demonstrated the ability of PPI networks (**Box 3**) to recapitulate the biological processes affected in disease, suggesting they are powerful indicators of functional convergence and can be used to address relationships between genes.

Table 1.5 | Properties and considerations of co-expression and PPI gene networks.

	Gene co-expression	Protein-protein interaction
Edge Relationship	Statistical association (correlation or mutual information)	Physical binding (interacting or not interacting)
Main advantages	Indirectly predicts co-regulation, physical interactions and cell type specificity; easiest to measure from tissue of interest	Based on direct physical interactions; predicts protein complexes and signalling pathways
Completeness of data across the genome	Most genes are similarly covered genome-wide	Incomplete assessment for most interactions; biased towards most well-studied molecules
Tissue specificity	Primary data are often tissue specific	Primary data are rarely tissue specific
Module-level interpretation	Reflects cell types and transcriptionally co-regulated biological processes	Protein complexes; signalling cascades; subcellular structures
Interpretation of hubs	Cell type-specific markers; molecular regulators such as transcription factors or RNA-binding proteins	Key proteins in complexes; converging points of signalling cascades
Sources of bias	Technical artefacts (RNA quality and batch effects); biological confounders (age and sex); post-mortem artefacts (cause of death)	Literature-curated data contain biases towards more well-studied interactions, which tend to be non-neuronal
Examples of bioinformatic validation	Preservation of co-expression in independent data; enrichment of physical interactions in modules	Enrichment of co-expression from independent data
Examples of experimental validation	Showing cell type specificity of hubs by *in situ* hybridization; demonstrating regulatory potential of hubs by hub gene knockdown	Co-immunoprecipitation of proteins of interest; disruption of protein complexes when hubs are targeted

(Adapted from Parikshak *et al.*, 2015).

Similarly to disease-associated genes, genes targeted by *dn* LoF and missense variants in ASD, ID and epilepsy patients display an excess of PPIs between genes with synaptic and transcriptional functions (Najmabadi *et al.*, 2011; Neale *et al.*, 2012; O'Roak *et al.*, 2012; Epi4K and EPGP Investigators, 2013; De Rubeis *et al.*, 2014; Hamdan *et al.*, 2014; Krumm *et al.*, 2014; Pinto *et al.*, 2014; Sanders *et al.*, 2015; EuroEPINOMICS-RES Consortium *et al.*, 2017; Shohat *et al.*, 2017; Yuen *et al.*, 2017). In fact, consistent with findings from transcriptomic studies, PPI analyses identified two principal ASD sub-networks encompassing synaptic and chromatin remodelling functions (Krumm *et al.*, 2014; Sanders *et al.*, 2015). Moreover, an excess of PPI connections was observed between genes disrupted by *dn* LGD SNVs, mapping to *dn* CNVs, and high confidence NDD genes (Najmabadi *et al.*,

2011; Neale *et al.*, 2012; Epi4K and EPGP Investigators, 2013; Hamdan *et al.*, 2014; Pinto *et al.*, 2014; EuroEPINOMICS-RES Consortium *et al.*, 2017). This suggests distinct mutational classes impact common functional processes across disorders, supporting the notion of aetiological overlap.

To overcome the incompleteness of PPI datasets, two groups have recently screened for PPIs between brain-expressed genes and ASD susceptibility genes, defined as syndromic genes or genes mapping to *dn* CNVs (Sakai *et al.*, 2012; Corominas *et al.*, 2014). The constructed interactome maps revealed proteins interacting with ASD-associated gene products are enriched for synaptic functions and have previously been implicated in other NDDs, supporting an aetiological overlap between disorders (Sakai *et al.*, 2012; Corominas *et al.*, 2014). Furthermore, many of the uncovered PPI genes are affected by CNVs from independent ASD cohorts, supporting the view that dispersed variants from patients affect common molecular processes.

1.4.4 Integrative approaches

Functional convergence can be inferred from various shared biological characteristics, including functional annotations, gene expression patterns, physical interactions and phenotypes resulting from the disruption of gene orthologues in animal models. As these data types vary in adequacy and competence, due to incompleteness, biases and specificity, they can be combined to increase the confidence of functional associations (**Box 3**).

One example of integrative network is NETwork-Based Analysis for Genetic association (NETBAG) (Gilman *et al.*, 2011). By incorporating various types of literature-curated data, such as KEGG, GO and PPIs, NETBAG assigns a score to pairs of genes reflecting their likelihood of participating in shared genetic phenotypes (Gilman *et al.*, 2011). This network has successfully been applied to ASD, such that the accompanying greedy search algorithm uncovered a module of highly interconnected genes mapping to *dn* patients CNVs; these genes were associated with synaptic functions and an appreciable number had previously been implicated in ASD and ID (Gilman *et al.*, 2011; Pinto *et al.*, 2014). The NETBAG approach was further extended to evaluate different types of mutational events, including CNVs, SNVs and GWAS loci (NETBAG+) (Gilman *et al.*, 2012). When applied to the combined set of genes affected by these variants in ASD patients, NETBAG+ identified the same modules involving synaptic or chromatin remodelling functions (Chang *et al.*, 2015). A drawback of NETBAG derives from the sole incorporation of literature-curated data, making the network highly susceptible to ascertainment biases and challenging its efficacy for the study of unknown genes.

The Phenotypic Linkage Network (PLN) provides a less biased alternative to NETBAG to examine functional gene associations (Honti *et al.*, 2014). The concept of the PLN stems from the observations that (1) thousands of unique orthologues of human genes have been knocked out in mouse models and assigned phenotypes and that (2) the similarity of mouse phenotypes is correlated to that of human diseases (Honti *et al.*, 2014). Functional datasets

(i.e. GO, KEGG, Reactome), PPIs and unbiased genome-wide expression datasets are evaluated and re-scored using mouse phenotypic annotations from the MGI as a benchmark (Honti *et al.*, 2014). The re-scored datasets are further integrated to construct scores, reflecting the likelihood of genes to cause similar phenotypes when disrupted, based on several lines of evidence. Genes targeted by *dn* LGD SNVs in patients with ASD, ID and epilepsy are highly interconnected in the PLN, suggesting it is able to recapitulate the functional convergence underlying NDDs and outperforms individual annotation, co-expression or PPI networks (Honti *et al.*, 2014).

More recently, a new approach, specifically tailored to identify groups of genes associated with a disorder of interest, has been developed and named Merging Affected Genes into Integrated networks (MAGI) (Hormozdiari *et al.*, 2015). MAGI considers small sets of high confidence disease genes, which are co-expressed and whose protein products interact and recursively adds genes to these *seed pathways* based on PPIs and brain co-expression data. Ultimately, gene modules are formed by maximising the inclusion of genes disrupted by *dn* mutations in patients but not controls (Hormozdiari *et al.*, 2015). Consistent with previous reports, the application of MAGI to ASD and ID uncovered two modules comprising (1) embryonically-expressed chromatin remodellers and (2) postnatal genes with synaptic functions (Hormozdiari *et al.*, 2015; Stessman *et al.*, 2017). Moreover, this study showed that while low functioning ASD patients (IQ < 70) were enriched for *dn* LoF variant in embryonic chromatin remodellers, high functioning individuals display an excess of missense mutations affecting synaptic genes (Hormozdiari *et al.*,

2015). These findings support the notion of genotype and phenotype convergence, wherein individuals presenting with similar genetic insults present more similar phenotypic abnormalities (**section 1.5**).

More generally, a statistical framework evaluating the likelihood of pathogenicity of genes uncovered in WES studies has been developed and named Transmission And *De novo* Association (TADA) (He *et al.*, 2013). By examining gene mutation rates and the nature and frequency of mutational events in trios, cases and controls, TADA identifies sets of genes more likely to be involved in disease pathogenesis (He *et al.*, 2013). This approach has successfully been applied to ASD, ID, epilepsy and schizophrenia, such that prioritised genes included NDD-associated genes and were associated with synaptic and chromatin remodelling functions (De Rubeis *et al.*, 2014; Sanders *et al.*, 2015; Li *et al.*, 2016). A novel approach, termed Detection Association With Networks (DAWN), integrates brain co-expression data with TADA scores to uncover sub-networks of genes associated with increased disease risk (Liu *et al.*, 2014). When applied to ASD, DAWN identified a subset of risk genes, enriched for high confidence NDD genes and PPIs, which were disrupted by *dn* LoF variants in an appreciable number of patients (De Rubeis *et al.*, 2014; Liu *et al.*, 2014).

Genome-wide frameworks examining gene essentiality have also been developed to facilitate the prioritisation of disease variants and genes. By combining various gene metrics and examining variant frequencies in healthy populations, these methods interrogate the likelihood of genes to induce

phenotypic abnormalities upon mutational disruptions. Such approaches include selection coefficients (Cassa *et al.*, 2017), haploinsufficiency predictions (Huang *et al.*, 2010; Steinberg *et al.*, 2015), mutational constraint predictions (Samocha *et al.*, 2014), Residual Variation Intolerance Scores (RVIS) (Petrovski *et al.*, 2013) and LoF Intolerance probabilities (pLI) (Lek *et al.*, 2016). Moreover, the functional impacts of individual variants can be predicted by assessing sequence, structural and evolutionary metrics, leading to the development of scoring methods, such as Sorting Intolerant From Tolerant (SIFT) (Ng and Henikoff, 2003), Genomic Evolutionary Rate Profiling (GERP) (Cooper *et al.*, 2005), Polymorphism Phenotyping v2 (PolyPhen-2) (Adzhubei *et al.*, 2013) and Combined Annotation-Dependent Depletion (CADD) (Kircher *et al.*, 2014).

1.4.5 Conclusion

Systems approaches address the underlying mechanisms of NDDs by providing organisational frameworks, which examine various types of gene and protein relationships, in order to uncover the changes in molecular systems leading to disease. While these methodologies have predominantly been applied to ASD to date, the emergence of large-scale genetic studies will inevitably require the use of such genome-wide, integrated, pipelined and reproducible approaches to develop a more comprehensive understanding of across NDDs.

1.5 Genotype and phenotype relationships in NDDs

The traditional phenotype-first approach aims to collect patients with the same clinical presentations and to subsequently identify the causal mutation(s). However, with the emergence of high throughput screening, it became evident that the heritability of complex disorders cannot be explained by genetic variants of large effect sizes. In fact, while GWAS failed to uncover susceptibility loci for many NDDs, microarray and WES studies identified large numbers of rare and *dn* structural and point variants, suggesting the genetic architecture of NDDs is complex and multifactorial. Furthermore, this compelling genetic heterogeneity is mirrored in an equally striking phenotypic heterogeneity, as illustrated by the qualitative variability of core impairments and the diversity of co-occurring conditions and abnormalities. As a result, more appropriate and accurate approaches are needed to link genotype to clinical phenotypes.

In 1990, Percy postulated that ASD is an umbrella term for a large number of syndromes with distinct aetiologies (Percy, 1990). By generalising this view, the genotype-first approach proposes that individuals with the same genetic insults form distinctive clinical subtypes of disease (**Fig 1.4a**) (Stessman *et al.*, 2014). Accordingly, single gene causes have been identified for many phenotypically well-characterised neurological syndromes (Huang *et al.*, 2014). Even as increasing numbers of individuals are identified with abnormalities beyond those described for the syndrome of interest (Grayton *et al.*, 2012), the causal genes remain recurrently and significantly associated

with the expected clinical manifestations. Indeed, despite the variable expressivity and pleiotropic effects of syndromic genes, they globally converge at the phenotypic level, given a sufficiently large set of patients to evaluate. In fact, the genotype-first approach led to the characterisation of distinct clinical entities associated with a significant number of genes recurrently observed in NDDs (see Stessman *et al.*, 2016 for examples). For example, loss of function mutations in the *CHD8* gene are associated with ASD with accompanying macrocephaly (i.e. abnormally large head compared to body) and gastro-intestinal problems. Conversely, loss of function mutations in the *ADNP* gene are linked to ASD with loss of expressive language (Bernier *et al.*, 2014; Helsmoortel *et al.*, 2014; Barnard *et al.*, 2015; Stessman *et al.*, 2016).

Stepping away from the study of single genes, the genotype-first approach can be extended to ask whether physically interacting or functionally related genes are associated with distinct clinical disease subtypes (**Fig 1.4b**). This approach relies on the assumption that analogous phenotypes result from the disruption of common molecular processes. Compared to single gene studies, a considerable gain in statistical power can be achieved through the interrogation of larger number of individuals with hits in the same biological pathways. Recently, two studies revealed that DD and ID patients with functionally related or co-expressed *dn* CNVs display more similar clinical abnormalities than expected by chance (Andrews *et al.*, 2015a; Andrews *et al.*, 2015b). While these findings support the notion of genotype and phenotype convergence, they highlight the need to investigate this coherence

and characterise the distinguishing phenotypic features underlying different molecular aetiologies.

The implementation of the genotype-first approach has been enabled by two major advancements, namely (1) the decreasing cost of NGS technologies and (2) the development of large consortia aiming to collect large numbers of comprehensively phenotyped patients. Genetic screening of thousands of individuals revealed dozens of genes recurrently disrupted in NDDs, whose phenotypic effects can be investigated in clinically characterised patient cohorts, such as the Simons Simplex Collection (SSC) of ASD patients (Fischbach and Lord, 2010). The SSC provides various measures of core ASD symptoms, cognitive and motor impairments, data on comorbid conditions and abnormalities as well as a comprehensive medical history of patients and families, which comprises information regarding diseases, physical malformations, food intolerances and associated problems, medications, immunisations, pregnancies and others (Fischbach and Lord, 2010). The need for standardised and systematic phenotypic comparisons of patients also led the development of the Human Phenotype Ontology, which comprises over 10,000 phenotypic terms organised in a hierarchy (Robinson *et al.*, 2008; Köhler *et al.*, 2017). Clinical characterisation using this ontology has been achieved for several patient cohorts, including the DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resource (DECIPHER), which curates individuals with chromosomal imbalances (Firth *et al.*, 2009; Bragin *et al.*, 2014) and a cohort of individuals with ID, DD or

congenital abnormalities, collected at the Radboud University Nijmegen Medical Centre and described in Vulto-van Silfhout *et al.*, 2013a.

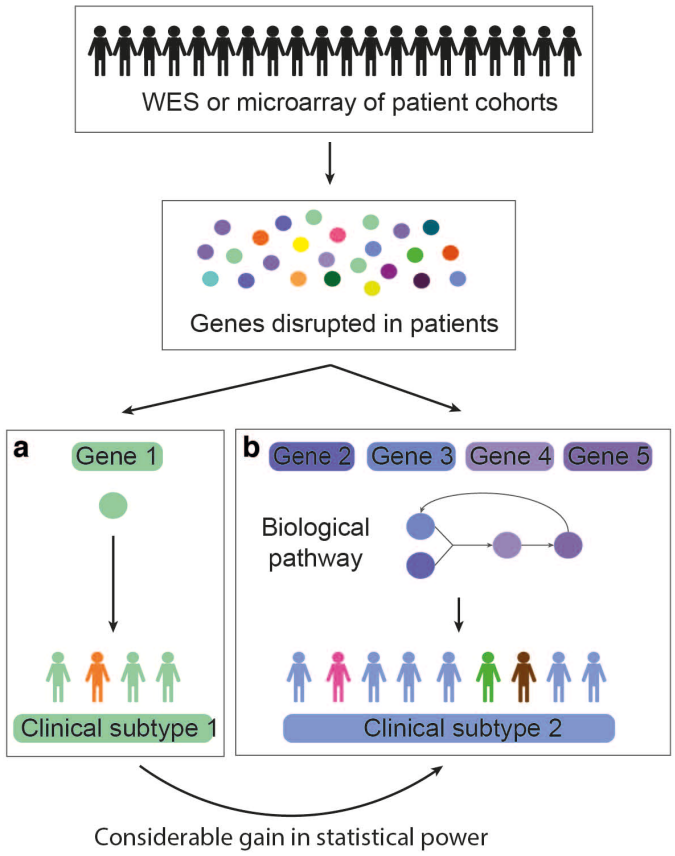


Fig 1.4 | Genotype-first approach. The application of whole-exome sequencing or array technology to NDD cohorts leads to the identification of genes recurrently disrupted by rare or *de novo* variants (i.e. SNVs and CNVs) in patients. (**a**) Thorough clinical examination of individuals carrying the same mutations aims to identify a shared set of phenotypic abnormalities specifically linked to each candidate (i.e. clinical subtype). Nevertheless, as individual variants are rare (incidence of < 1% in the population), relatively small numbers of carriers are available. (**b**) To gain statistical power, this approach is extended to consider multiple genes simultaneously. Under the assumption that the disruption of common biological pathways underlies the development of analogous phenotypes, genes are assembled in pathways, based on their co-expression patterns, functional relatedness or physical interactions of the associated protein products. Larger numbers of individuals with hits in a pathway are available and used to assess phenotypic convergence and identify distinguishing clinical subtypes.

1.6 Conclusions and thesis aims

Disentangling the genetic architecture of NDDs has pivotal implications for future study designs, the aetiological understanding of disease and nosology, as well as the development of effective treatment strategies.

Firstly, with regard to future study designs, the establishment of adequate genetic models is necessary to direct our focus towards the search for rare and *dn* gene variants in simplex families or the use of large-scale case-control and multiplex family association studies. Secondly, the discovery and functional studies of genetic risk factors is crucial in elucidating the aetiology of disease and providing a mechanistic foundation to develop treatments. Additionally, the identification of causal genetic lesions in patients has strong implications for counselling and the surveillance of family members. Lastly, the genetic heterogeneity of NDDs indicates no single treatment can be applied to all patients and individualised therapies are not medically and economically viable. Instead, the identification of biological pathways disrupted in disease provides a solid grounding for the drug discovery process. Accordingly, several distinct pharmacological therapies would be developed and applied to NDD patients, based on their clinical and genetic characteristics.

The aim of this thesis is to expand our genetic and aetiological understanding of NDDs, by tackling the aforementioned theoretical underpinnings, namely genetic models of disease, identification of risk genes and genetic

convergence at the pathway level. Such advances, in turn, will suggest potential therapeutic avenues, tailored to treat the underlying causes of disease, rather than the generic symptoms.

In **chapter 3**, we explore the rare variant model in GTS by performing a genome-wide CNV analysis of a Danish cohort of GTS patients and healthy controls. Specifically, we seek to uncover disease-associated variants, genes and molecular pathways.

In **chapters 4 and 5**, we investigate the genotype and phenotype relationships in ID, DD and ASD by deconstructing the associated genetic heterogeneity into several biological pathways. We, then, search for convergence at the phenotypic level by examining whether individuals with similar genetic disruptions present with analogous clinical abnormalities. Additionally in **chapter 5**, we interrogate the validity of oligogenic models with modifying effects in ASD, by examining the influence of CVs on phenotypic variability. For these studies, we use a cohort of individuals with ID, DD or CA, collected at the Radboud University Nijmegen Medical Centre (Vulto-van Silhout *et al.*, 2013a; Andrews *et al.*, 2015b) and the SSC cohort of ASD patients (Fischbach and Lord, 2010).

Lastly, in **chapter 6**, we analyse over 1,000 gene KO mouse models, generated by the Wellcome Trust Sanger Institute Mouse Genetics Project. We seek to uncover functionally converging genes, which cause mouse

neuroanatomical defects and propose candidate genes, possibly implicated in human brain abnormalities and ID.

CHAPTER 2

Common materials and methods

2.1 Gene lists

The gene lists constructed are recapitulated in **Table 2.1**.

2.1.1 Gene haploinsufficiency scores

We downloaded probabilities of HaploInSufficiency (HIS) for 11,448 protein-coding genes with Ensembl Identifiers (Ids) from Huang *et al*., 2010. Probabilities were estimated using various gene features including genomic (e.g. gene length), functional (expression patterns) and evolutionary properties (Huang *et al*., 2010). A second set of genome-wide HIS scores was downloaded for 19,701 genes with Ensembl Ids from Steinberg *et al*., 2015. These HIS scores incorporate data relating to gene co-expression with known HIS genes, evolutionary conservation, prenatal and postnatal temporal gene expression specificity and depletion of common variants.

2.1.2 Human functional gene lists

We extracted 842 human orthologues of 842 mouse FMRP target genes from Darnell *et al*., 2011 (mouse FMRP targets are used in **chapter 6**). Moreover, we considered a subpopulation of (1) 287 embryonically upregulated FMRP target genes, enriched for penetrant gene mutations in ASD patients and of (2) 239 nonembryonic synaptic FMRP targets, enriched for less penetrant disruptions in ASD patients (Julia Steinberg private correspondence) (Steinberg and Webber, 2013). These sub-groups are referred to as (1)

embryonic and (2) synaptic modules, respectively, and were identified using the set of 842 human orthologues of mouse FMRP targets uncovered in Darnell *et al.*, 2011. Additionally, we downloaded 152 and 1,080 human 1:1 orthologues of mouse synaptosome and PostSynaptic Density (PSD) genes, respectively (*Genes2Cognition* database; Croning *et al*., 2009). By mining the literatures, we further retrieved 1,003 constraint genes intolerant to mutations (Samocha *et al*., 2014), 432 chromatin remodelling genes (Huang *et al*., 2013) and 149 histone modifying genes (HIstome database; Khare *et al*., 2011). Using the QuickGo online tool, we formed a list of 3,071 Transcription Factors (TFs) by extracting genes manually annotated with the GO term *regulation of transcription, DNA-templated (*GO:0006355) based on human experimental evidence (codes IMI, IGI, IPI, IDA, IEP, EXP). Finally, we downloaded a set of 1,223 brain-specific genes from the Human Protein Atlas, with at least five-fold higher mRNA levels in the cerebral cortex compared to the average expression across all 32 tissues analysed (Uhlen *et al*., 2015).

2.1.3 High confidence disease gene lists

We downloaded a list of 6,116 disease genes annotated in the Online Mendelian Inheritance in Man (OMIM) database from Ensembl Biomart (MIM Morbid list; Cunningham *et al*., 2015) and a list of 167 Autism Spectrum Disorder (ASD) genes from the SFARI Gene database (scored 1, 2 or 3) (Abrahams *et al*., 2013). Additionally, three non-exclusive lists of 292, 429 and 650 Intellectual Disability (ID) genes were extracted from Pinto *et al.,* 2014, Casanova *et al*., 2016 and Kochinke *et al*., 2016, respectively.

Table 2.1 | Description of genes lists formed.

Gene set type	Gene set	Alternate name	Number genes	Description
Genes disrupted in NDD patients	ASD penetrant mutations (Iossifov *et al.*, 2015)	ASD Iossifov	421	Genes disrupted *dn* LoF SNVs (nonsense, frameshift and splice-site) in SSC patients (Iossifov *et al.*, 2014)
	Nijmegen dn CNVs (Vulto-van Silfhout *et al.*, 2013a)	Nijmegen	3,598 (160)	Genes mapping to *dn* CNVs < 5 mb from patients with ID, DD or CA from a cohort collected at the Radboud University Nijmegen Medical Centre (Vulto-van Silfhout *et al.*, 2013a)
	Decipher dn CNVs (Bragin *et al.*, 2014)	Decipher	14,302 (1,640)	Genes mapping to *dn* CNVs < 5 mb from DECIPHER patients (Bragin *et al.*, 2014)
	GTS CNV genes (Fernandez *et al.*, 2012)	Fernandez	642 (84)	Genes mapping to CNVs > 20 kb and < 10 mb from GTS patients (Fernandez *et al.*, 2012)
Gene intolerance scores	Selection coefficient	shet	15,877	Gene scores downloaded from Cassa *et al.*, 2017
	Probability of haploinsufficiency	HIS	11,448	Gene scores downloaded from Huang *et al.*, 2010
	Residual variation intolerance score	RVIS	17,047	Gene scores downloaded from Petrovski *et al.*, 2013
	Probability of intolerance to loss of function (nonsense, splice-site, frameshift) variants	pLI	17,816	Gene scores downloaded from Lek *et al.*, 2016
NDD-associated gene classes	MIM Morbid	-	6,116	List of genes associated with OMIM disorders downloaded from Ensembl (Cunningham *et al.*, 2015)
	Transcription factors	TFs	3,071	List of genes manually annotated with the GO term *regulation of transcription, DNA-templated* - GO:0006355, based on human experimental evidence (QuickGO)
	Chromatin remodelling genes	Chromatin remodellers	432	List of chromatin emodelling genes downloaded from Huang *et al.*, 2013
	Histone modifying genes	Histone modifiers	149	List of histone modifying genes downloaded from the Histome database (Khare *et al.*, 2011)
Human brain genes	Brain-expressed	-	13,306	List of genes with RPKM values > 1 in at least one region and during at least one developmental point in the brain extracted from Brainspan (Sunkin *et al.*, 2013)
	Brain-specific	-	1223	List of genes with ≥ 5-fold higher mRNA levels in the cerebral cortex than the average across 32 tissues downloaded from the Human Protein Atlas (HPA; Uhlen *et al.*, 2015)
	FMRP targets	-	842	List of FMRP target genes downloaded from Darnell *et al.*, 2011
ID genes	ID (Pinto *et al.*, 2014)	ID Pinto	292	List of high confidence ID genes downloaded from Pinto *et al.*, 2014
	ID (Kochinke *et al.*, 2016)	ID Kochinke	650	List of high confidence ID genes downloaded from Kochinke *et al.*, 2016
	ID (Casanova *et al.*, 2016)	ID Casanova	429	List of high confidence ID genes downloaded from Casanova *et al.*, 2016
Mouse genes	Postsynaptic density	PSD	1,040	List of mouse and human 1:1 orthologues of mouse post-synaptic density genes downloaded from the Genes2Cognition database (Croning *et al.*, 2009)
	Synaptosome	-	150	List of mouse and human 1:1 orthologues of mouse synaptosome genes downloaded from the Genes2Cognition database (Croning *et al.*, 2009)

The Gilles de la Tourette syndrome (GTS) CNV gene set (*Fernandez*) and the *Gene intolerance scores* datasets are described in **section 3.3.7** and **6.3.8**, respectively.

2.1.4 Genes and phenotypes from patients with NDDs

We downloaded a list of 421 genes with Ensembl Ids disrupted by *de novo* (*dn*) Loss of Function (LoF) SNVs in 2,498 ASD patients from the Simons Simplex Collection (SSC) (Iossifov *et al.*, 2014). LoF variants, encompassing mutations affecting protein functionality, were defined as nonsense, splice-site and frameshift *dn* SNVs, absent from unaffected siblings. The SSC cohort has been deeply and rigorously phenotyped, such that a large variety of clinical data is available including the presence of physical and morphological abnormalities.

Recognising that *dn* CNVs are strongly associated with NDDs including ASD and ID, we extracted genes mapping to *dn* CNVs of sizes < 5 mb from several cohorts. We considered (1) a cohort of patients with ID, Developmental Delay (DD) or Congenital Abnormalities (CA) collected at the Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands termed *Nijmegen* (n = 3,598 genes) (Vulto-van Silfhout *et al.*, 2013a; Andrews *et al.*, 2015b) and (2) a cohort of individuals with chromosomal imbalances from the Database of GenomiC varIation and Phenotype in Humans using Ensembl Resources (DECIPHER) (n = 14,302 genes) (Bragin *et al.*, 2014; Firth *et al.*, 2009). When assessing over-representations, we only consider genes mapping to CNVs overlapping ≤ 5 genes; the number of DECIPHER and Nijmegen genes is then reduced to 1,640 and 160, respectively.

The Nijmegen and DECIPHER cohorts have been systematically phenotyped by clinicians using terms from the Human Phenotype Ontology (HPO) (Robinson *et al.*, 2008; Köhler *et al.*, 2017), such that absence of specific terms translates into absence of clinical abnormalities, rather than lack of assessment. Of note, DECIPHER annotations are not as consistent as those for the Nijmegen cohort. In total, 77%, 41%, 10% and 18% of Nijmegen patients have varying degrees of ID, DD, CA and ASD, respectively. These proportions are less for DECIPHER patients (i.e. 49%, 22%, 6% and 7%, respectively). Overall, the cohorts are annotated with 1,395 and 350 distinct HPO terms, respectively, which we remapped to fit the newest HPO release.

2.2 Functional gene annotation

2.2.1 GO annotations

We downloaded Gene Ontology (GO) Biological Process (GO-BP), Molecular Function (GO-MF) and Cellular Compartment (GO-CC) annotations for 18,469 mouse and 20,165 human genes from the GO database (Gene Ontology Consortium, 2000). To reduce uninformative results, we considered GO terms annotated with > 100 and < 4000 genes, resulting in 1,659 and 1,761 GO pathways annotated with 17,968 mouse and 19,435 human genes, respectively.

2.2.2 MGI annotations

We downloaded annotations regarding mouse phenotypes caused by the disruption of protein-coding genes from the Mouse Genome Informatics (MGI) database (Blake *et al*., 2017). We annotated 6,169 human genes with the mouse phenotypes associated with the disruption of their 1:1 mouse orthologues. Each gene was further annotated with all parent terms of the phenotype it influenced, as given by the mammalian phenotype ontology (Smith and Eppig, 2009). Ultimately, we considered 6,867 mouse phenotypic terms annotated with > 100 human genes.

2.3 Expression datasets

2.3.1 Body-wide gene expression map

We downloaded RNA sequencing gene expression data for 54,031 human genes across 51 bodily tissues from the Genotype Tissue Expression (GTEx) project (Lonsdale *et al.*, 2013). The dataset was filtered to consider only 16,611 protein-coding genes with 19,625 Ensembl Ids, as given by Ensembl Biomart (Cunningham *et al*., 2015). The expression level of each sample was measured as the number of Fragments Per Kilobase of transcript per Million mapped reads (FPKM). For each gene, tissue expression was calculated as the median FPKM value across all samples mapping to the given tissue (average of 83.70 samples per gene). To increase statistical power, we assembled the 51 available tissues into 30 *structures*, such that expression

for each structure was calculated as the median expression across all associated tissues (**Table 2.2**).

2.3.2 Spatiotemporal brain expression map

We downloaded normalised RNA sequencing expression data for 52,376 human genes across 27 brain tissues spanning twelve developmental stages (from post-conception to adulthood) from Brainspan (Sunkin *et al*., 2013). The dataset was filtered to consider only 19,672 protein-coding genes, as given by Ensembl Biomart (Cunningham *et al*., 2015). The expression level of each sample was measured as the number of Reads Per Kilobase of transcript per Million mapped reads (RPKM).

To evaluate temporal brain expression patterns, we defined six developmental stages as suggested by the experimental protocols of Brainspan. These are (1) early/mid prenatal development (0 to 18 weeks post-conception), (2) late prenatal development (19 to 38 weeks post-conception), (3) infancy (birth to 18 months), (4) childhood (19 months to 11 years), (5) adolescence (12 to 19 years) and (6) adulthood (≥ 20 years). For each gene, we calculated the overall brain expression level at each developmental point as the median RPKM value across all tissues and samples mapping to the given stage.

To evaluate spatial brain expression patterns, we assembled the 27 brain tissues into six brain regions as follows: (1) neocortex (formed by the prefrontal cortex (dorsolateral and vetrolateral), neocortex (temporal, occipital

and parietal) and cortex (visual, somatosensory, motor, motor-sensory, auditory, caudal superior temporal, inferior parietal, orbital frontal, inferolateral temporal and anterior cingulate)), (2) amygdala (formed by the amygdaloid complex), (3) medulla (formed by the dorsal thalamus and mediodorsal nucleus of thalamus), (4) striatum (formed by striatum and the caudal, lateral and medial ganglionic eminences), (5) cerebellum (formed by the cerebellar cortex, cerebellum and upper (rostral) rhombic lip) and (6) hippocampus. For each gene, the brain region expression level was defined as the median RPKM value across all samples mapping to the given region, independently of the sample's time point during development.

Brain spatiotemporal expression patterns were calculated for each gene as the median RPKM value across gene samples mapping to all combinations of brain regions and developmental points. We identified 13,306 brain-expressed genes, with RPKM values > 1 in at least one brain region during one developmental stage (**Table 2.1**). This dataset was not used to interrogate gene expression dynamics due to reduced statistical power (n = 36 tissue-stage combinations to evaluate).

Table 2.2 | Description of bodily structures constructed using the GTEx expression dataset.

Structure	Tissues included
Adipose	Sub-cutaneous, Visceral (Omentum)
Adrenal Gland	-
Artery	Aorta, Coronary, Tibial
Bladder	-
Brain	Amygdala, Anterior cingulate cortex (BA24), Caudate (basal ganglia), Cerebellar hemisphere, Cerebelum, Cortex, Frontal cortex (BA9), Hippocampus, Hypothalamus, Nucleus accumbens (basal ganglia), Putamen (basal ganglia), Spinal cord (cervical c-1), Substantia nigra
Breast	Mammary tissue
Cervix	Ectocervix, Endocervix
Colon	Sigmoid, Transverse
Esophagus	Gastroesophageal junction, Mucosa, Muscularis
Fallopian tube	-
Heart	Atrial appendage, Left ventricle
Kidney	Cortex
Liver	-
Lung	-
Minor salivary gland	-
Muscle	Skeletal, Tibial
Nerve	-
Ovary	-
Pancreas	-
Pituitary	-
Prostate	-
Skin	Sun exposed (lower leg), Not sun exposed (supra-pubic)
Small intestine	Terminal ileum
Spleen	-
Stomach	-
Testis	-
Thyroid	-
Uterus	-
Vagina	-
Whole blood	-

The GTEx expression dataset provides expression measurements across 51 bodily tissues, which were assembled into 30 structures.

2.4 Gene networks

The gene networks constructed or downloaded are summarised in **Table 2.3**.

2.4.1 GTEx and Brainspan co-expression networks

Genes with FPKM or RPKM values < 1 in > 95% of samples were excluded from the GTEx and Brainspan expression datasets downloaded (Lonsdale *et al.*, 2013; Sunkin *et al.*, 2013). The Brainspan co-expression network was derived by estimating the correlation of expression patterns of pairs of protein-coding gene across all associated samples, using Pearson's coefficients (Steinberg and Webber, 2013). For the GTEx co-expression network, we used a weighted Pearson's coefficient in order to control for sample redundancy, as described in Obayashi *et al.*, 2008 (the expression similarity between two genes is defined the maximum similarity value of corresponding probes).

In these networks, nodes are genes and edges represent the correlation (r) of their expression patterns. To reduce uninformative results, we restricted subsequent analyses to gene links with $r > 0.6$ or $r > 0.8$, although results were confirmed using networks with $r > 0.5$, $r > 0.7$ and $r > 0.9$.

2.4.2 Protein-protein interaction networks

A human Protein-Protein Interaction (PPI) network was downloaded from String (Szklarczyk *et al.*, 2015) and filtered to consider experimentally validated results only.

Table 2.3 | Description of gene networks.

Organism	Network Type	Name	Alternate name	Threshold	Number genes	Number edges	Description
Mouse	Functional network	MouseNet v2	-	all links	15,937	641,661	Downloaded from Kim *et al.*, 2016
Mouse	Co-expression network	Mouse developing CNS and adult brain co-expression network	-	r>0.5	19,496	38,165,080	Co-expression network, wherein pairs of genes are assigned a score, reflecting the correlation of their expression patterns across the developing CNS, based on RNA sequencing data
				r>0.6	19,496	30,437,026	
				r>0.7	19,496	23,052,649	
				r>0.8	19,496	15,759,232	
				r>0.9	19,496	8,433,834	
Human	Functional network	Phenotypic linkage network	PLN	all links	20,642	4,158,191	Constructed as described in Honti *et al.*, 2014
				top 25% links	19,229	1,039,548	
				top 10% links	16,937	415,819	
				top 1% links	9,523	41,582	
Human	PPI network	String	-	all links	15,992	1,561,410	All PPIs with experimental evidence were downloaded from String (Szklarczyk *et al.*, 2015)
Human	Co-expresison network	GTEx	-	r>0.5	15,870	6,943,284	Co-expression network, in which pairs of genes are assigned a score, reflecting the correlation of their expression patterns, based on GTEx and Brainspan RNA-seq expression data.
				r>0.6	15,371	2,962,276	
				r>0.7	13,486	963,763	
				r>0.8	8,680	222,165	
				r>0.9	2,721	27,837	
		Brainspan	BS	r>0.5	14,886	19,182,009	
				r>0.6	14,678	11,421,142	
				r>0.7	13,869	5,428,039	
				r>0.8	11,414	1,610,975	
				r>0.9	5,246	124,603	

Mouse gene networks are described in **section 6.3.11**.

2.4.3 Phenotypic linkage network

We combined various genomic datasets, which evaluate PPIs, biological pathways and processes, co-expression patterns and phenotypes observed in animal models, in an integrated gene network as described in Honti *et al.*, 2014 (**Fig 2.1; Table 2.4**). This network, denoted a Phenotypic Linkage Network (PLN), assigns scores to gene pairs reflecting their likelihood of functional interaction, based on multiple lines of evidence. Details regarding the construction of the PLN are found in **section S2**. Overall, this network

consists of 4,158,191 links between 20,642 genes. While the whole network is used for most analyses, results are confirmed in sub-networks consisting of the top 1%, 10% and 25% strongest links.

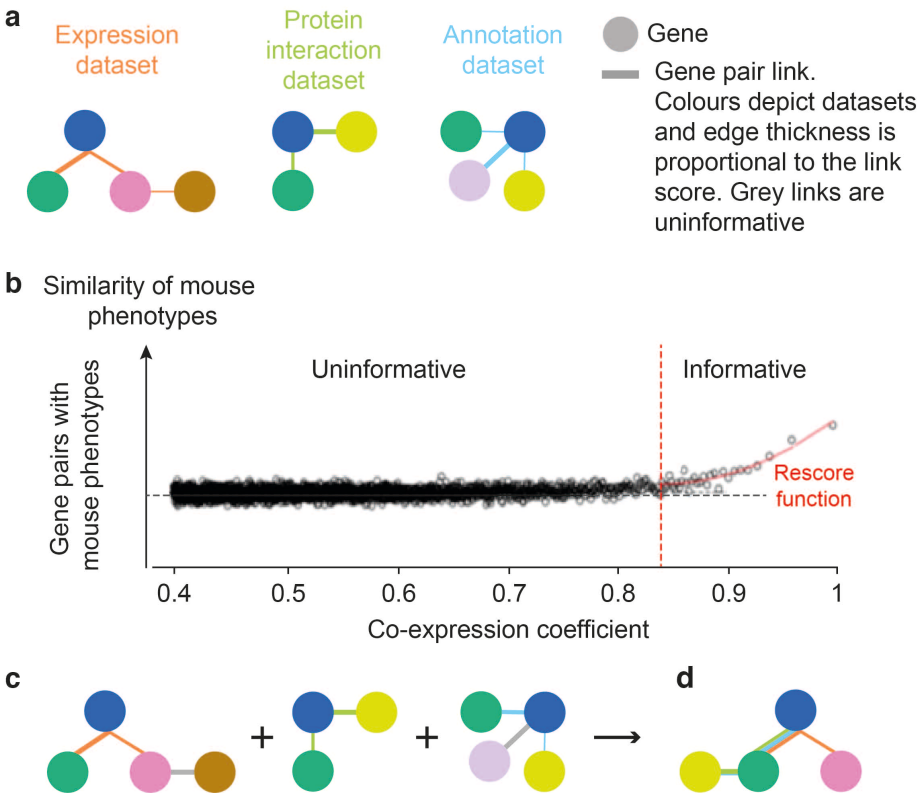


Fig 2.1 | Phenotypic linkage network (PLN) construction. (**a**) Genomics datasets (e.g. expression datasets, wherein links depict co-expression coefficients; annotation datasets, wherein links depict the semantic similarity of annotations and physical interaction datasets) are processed to extract gene pair linkages. (**b**) Each dataset is assessed on the MGI benchmark, which reveals the similarity of the phenotypes resulting from the targeted disruption of 1:1 gene orthologues in mouse models (y-axis). When gene pairs influence mouse phenotypes, which are more similar than expected by chance (given as the upper 45th percentile of all mouse phenotype similarities for the given dataset) (horizontal black dotted line), they are said to be informative (vertical red dotted line) and are re-scored according to a linear regression function. (**c**) Once datasets have been re-scored, they are integrated together and when several re-scored datasets suggest functional linkage between two genes, (**d**) they are combined.

Table 2.4 | Genomics datasets used to build the PLN.

Data Type	Dataset	Score Type	Description
Gene Annotations	Mouse genome database (MGI) (Blake *et al.*, 2017) Gene ontology (GO) – biological process (Gene Ontology Consortium, 2000) Gene ontology (GO) – molecular function (Gene Ontology Consortium, 2000) Gene ontology (GO) – cellular component (Gene Ontology Consortium, 2000) InterPro (Hunter *et al.*, 2012) KEGG (Kanehisa *et al.*, 2000; Kanehisa *et al.*, 2017) Reactome (Fabrecat *et al.*, 2014; Croft *et al.*, 2014)	Semantic similarity	Score characterizing the similarity of annotations of pairs of genes
Physical interactions	BioGRID (Stark *et al.*, 2011), IntAct (Kerrien *et al.*, 2012) Corum (Ruepp *et al.*, 2010), DICS (Dietmann *et al.*, 2009) Reactome (Fabrecat *et al.*, 2014; Croft *et al.*, 2014)	Score	All interactions predicted from the same experiment are assigned the same score
Gene Expression	Brainspan (Sunkin *et al.*, 2013; Steinberg and Webber, 2013), GNF2 (Su *et al.*, 2004), GSE3594 (Zapala *et al.*, 2005) MTAB-62 (Lukk *et al.*, 2010) and Shyamsundar *et al.*, 2005)	Pearson's correlation coefficient	
Co-citation	STRING (Szklarczyk *et al.*, 2015)	Score	Human genes are assigned the scores of their mouse orthologues

2.5 Network analysis

2.5.1 Gene matching

The Coding DNA Sequence lengths (CDS) and network connectivities (i.e. degree) of genes influence their network associations (Honti *et al.*, 2014). We controlled for these confounding effects during clustering tests and expression analysis. CDS annotations for 22,513 human protein-coding genes were downloaded from Ensembl BioMart (Cunningham *et al.*, 2015). Genes were sorted by the features of interest (CDS and/or degree) and separated them into bins of 100 (Honti *et al.*, 2014; Steinberg and Webber, 2013). For a given gene set, we matched each gene to the bin with the closest CDS and/or degree and sampled without replacement a gene from each bin, thus forming a matched randomised gene set.

2.5.2 Gene set clustering

Based on the assumption that functionally related genes sit closely together in gene networks, we use clustering tests to examine their interconnectedness. The strength of connections between genes from a particular set is calculated as the sum of their link scores in the network (in the PPI network, all links have a score of 1) (**Fig 2.2a**). We then run 1,000 simulations estimating the sum of links of randomised gene sets, constructed by randomly sampling the same number of genes, matched for CDS and degree, from the genome or from a background of choice. By determining the proportion of simulations during which genes from randomised sets are more interconnected than genes of interest, we derive an empirical p-value, reflecting the likelihood of observing such an interconnected gene set by chance (**Fig 2.2c**). Gene sets and backgrounds are restricted to genes with at least one edge in the evaluated network.

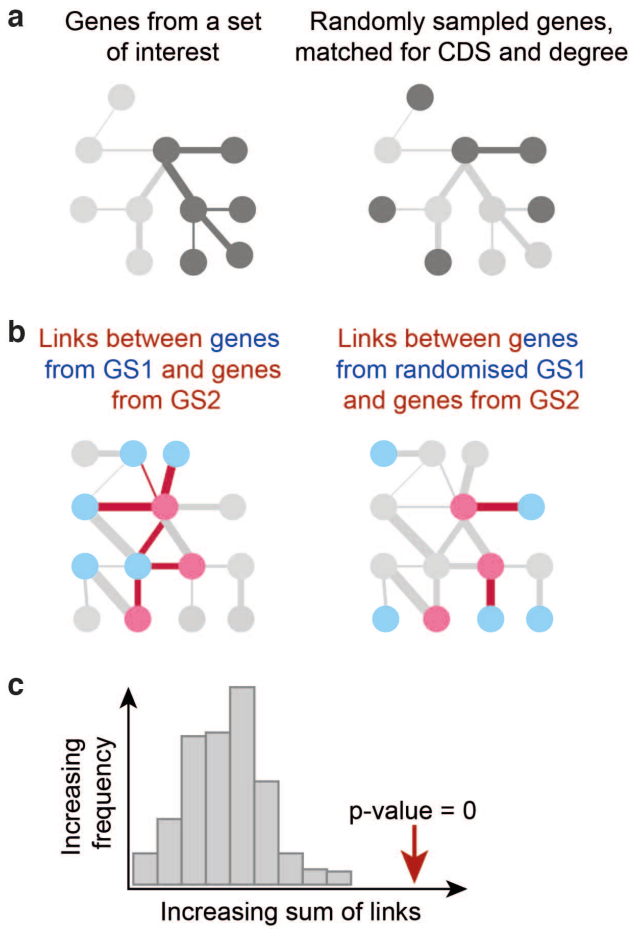


Fig 2.2 | Clustering gene sets. (**a**) To evaluate the interconnectedness of a gene set, the sum of links between genes from the given set is calculated (left panel; dark grey lines) and compared to that of randomly sampled genes, controlled for number, CDS and degree (right panel; dark grey lines). (**b**) To evaluate whether genes from a gene set, GS2 (red) are more interconnected with genes from gene set, GS1 (blue), we measure the sum of links between genes from the two sets (left panel; red lines) and compare it to that of a permuted GS1, formed by randomly sampling the same number of genes controlled for CDS and degree (right panel; red lines). (**c**) 1,000 randomisations are run and an empirical p-value is derived, reflecting the proportion of randomisations, during which permuted sets are more interconnected with (**a**) each other or (**b**) GS2 genes. GS1, gene set 1; GS2, gene set 2.

2.5.3 Clustering two gene sets

A slight variation of the method described in **section 2.5.2** is applied when clustering two gene sets, termed *GS1* and *GS2* (e.g. high confidence disease genes and a gene set of interest to evaluate) (**Fig 2.2b**). Following the

removal of genes present in both sets, we examine the network links between genes from the two sets as opposed to those between genes from the same set, such that the sum of links between genes from GS1 and GS2 is given by:

$$w = \sum_{i=a}^{b} \sum_{j=k}^{l} w(i,j)$$

where *[a...b]* and *[j..k]* denote genes from GS1 and GS2, respectively and *w(i,j)* represents the link weight between genes *i* and *j*, which belong to *[a...b]* and *[k...l]*, respectively. One thousand randomised GS1 sets are constructed, by randomly drawing the same number of genes, matched for CDS and degree, from the genome or background of choice. An empirical p-value is calculated as the proportion of randomisations during which genes from randomised GS1 sets are more strongly connected to GS2 genes than genes from the initial GS1 set (**Fig 2.2c**).

2.5.4 Gene module identification

To identify groups of strongly interconnected genes, termed *modules* or *communities*, gene networks are partitioned using the Louvain algorithm (Blondel *et al*., 2008). This greedy optimization method attempts to maximise the modularity of a network division (strength of connections inside modules as compared to between modules), based on two steps (**Fig 2.3**):

1. Each network node represents its own community. For a given node *i*, we calculate the change in modularity associated with removing *i* from its own community and adding it to the community of each of its neighbours. Node *i* is then added to the community, which provides the highest gain in modularity or remains in its own community if no gain is obtained. This step is performed for all nodes until no further increase in modularity can be achieved.
2. A new network is formed wherein communities from step 1 become nodes and edges are calculated as the sum of links between genes from the associated community.

These two steps are re-iterated until no further modularity improvement can be achieved, resulting an optimal network division involving one large or several smaller gene modules. Of note, the order in which nodes are considered can affect the final partitioning although it should not significantly influence the obtained modularity (Blondel *et al*., 2008). The algorithm was implemented in Gephi using a resolution parameter of 1.

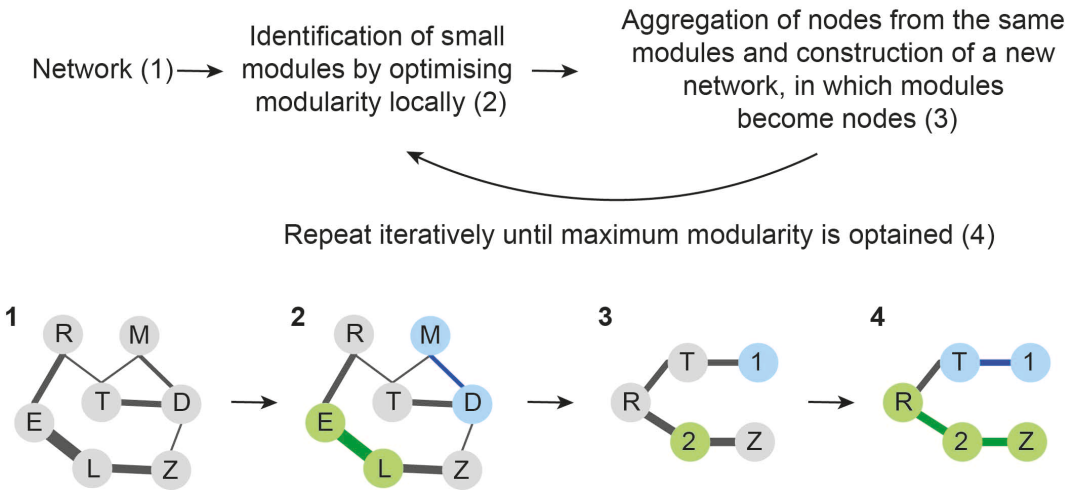


Fig 2.3 | Module detection. The Louvain algorithm is implemented in order to identify modules of densely interconnected genes from a starting network: (1) each node represents its own module and (2) nodes are merged into the same community when this leads to a modularity gain; (3) a new network is then formed wherein modules become nodes. (4) These steps are re-iterated until no further increase in modularity can be achieved.

2.6 Phenotypic similarity

Phenotypic similarity between subjects (i.e. patients or mouse models in **chapters 4/5 and 6**, respectively) is calculated using the Goodall3 index (Boriah *et al*., 2008). The Goodall3 index evaluates the shared presence or absence of a feature, weighted by its frequency, in order to assign a higher weight to the shared presence or absence of rare and common phenotypes, respectively. Computationally it is implemented as suggested by Boriah *et al*., 2008:

$$S(i,j)_k = \begin{cases} 1 - f_k^2 \ , & if\ k\ is\ present\ in\ both\ patients \\ 1 - (1 - f_k)^2 \ , & if\ k\ is\ absent\ i\ both\ patients \\ 0 \ , & if\ k\ is\ present\ in\ only\ one\ patient \end{cases}$$

where $S(i,j)_k$ is the Goodall3 measure between subjects *i* and *j* for phenotype *k* and f_i is the proportion of subjects from the whole sample displaying phenotype *k*.

The overall phenotypic similarity between pairwise combinations of subjects is estimated as the cumulative Goodall3 similarity across all phenotypes tested (or a subset of interest), normalised by the number of variables assessed. This results in a phenotypic similarity network, wherein nodes are subjects and edges represent the similarity of their phenotypes.

2.7 Statistical tests

2.7.1 Gene set enrichments and comparisons

The excess of genes from particular lists is tested using one-sided right-tailed Fisher's tests. HIS, gene intolerance scores (**section 5.2**) and phenotypic similarities are compared between two sets using one-sided right-tailed Mann-Whitney U tests. During functional enrichments using GO or MGI, the over-representation of each term is assessed using one-sided right-tailed hypergeometric tests. A 5% False Discovery Rate (FDR) is applied according the Benjamini-Hochberg (BH) procedure to control for the number of terms tested in each category (i.e. 6,867 mouse phenotypes; 1188 mouse GO-BP, 265 mouse GO-MF and 206 mouse GO-CC terms; and 1288, 265 and 208 human terms, respectively). The backgrounds used consist of all protein-

coding genes annotated in GO or MGI or only those present in the background of choice.

To ensure signals did not arise fortuitously, we formed n = 100 permuted gene sets by drawing the same number of genes from the genome or background of choice and performed the same tests to built a null distribution of p-values. As no p-values were significant post BH correction in any analyses, we concluded significant enrichments represented true biological signals (not shown).

2.7.2 Gene expression

We calibrated the GTEx and Brainspan expression datasets (**section 2.3**), by scaling expression between 0 and 1 for each body structure, brain region or developmental stage. The expression of each gene was then normalised across tissues, regions or stages using Euclidean distances in order to provide specificity. These normalised datasets are used to evaluate whether sets of genes are preferentially expressed in particular tissues or developmental stages and to compare expression levels between gene sets.

Tissue specificity: To evaluate tissue, brain region or developmental stage specificity, we use a permutation-based approach. We measure the cumulative tissue, region or stage expression of genes from a set of interest and compare it to that of randomly sampled genes from the genome (and annotated in the expression dataset considered) matched for CDS (**section**

2.5.1). By estimating the proportion of randomisations (n = 1,000), during which randomised sets have a higher cumulative expression than the set of interest, we derive an empirical p-value reflecting the likelihood of the genes of interest to be more strongly expressed in the tissue, region or stage tested than expected by chance. A correction for multiple testing is applied using the BH method to account for the number of tissues or stages tested (n = 30 for GTEx and n = 6 for Brainspan datasets).

Expression comparison between two gene sets: For a given tissue, brain region or developmental stage, gene expression between genes from two sets is compared using one-sided right-tailed Mann-Whitney U tests. A correction for multiple testing is applied using the BH method to account for the number of tissues or stages tested.

2.7.3 Phenotypic similarity comparisons

To assess whether subjects from one module are more phenotypically similar than expected by chance, we compare the pairwise phenotypic similarities between subjects from that module and those between remaining subjects (i.e. *remaining* = *background – module subjects*). Significance is derived using one-sided right-tailed Mann-Whitney U tests. We find consistent results when comparing the phenotypic similarities between module subjects and those between module subjects and remaining subjects using one-sided right-tailed Mann-Whitney U tests (not shown).

CHAPTER 3

Gilles de la Tourette syndrome:

a CNV study

3.1 Introduction

Gilles de la Tourette Syndrome (GTS) is a NeuroDevelopmental Disorder (NDD), characterised by the presence of multiple motor and at least one vocal tic. Tics typically start during late childhood or adolescence (mean age of 7 years) (Robertson *et al.*, 2009) and manifest themselves as involuntary movements and vocalisations, preceded by premonitory urges. Patients often present additional comorbidities such as Attention Deficit Hyperactivity Disorder (ADHD) (i.e. 21-90% of patients), Obsessive Compulsive Disorder (OCD) (i.e. 11-80% of patients), depression (i.e. 13-76% of patients) and to a lesser extent Autism Spectrum Disorder (ASD), learning disabilities and epilepsy (Paschou, 2013; Mol Debes; 2013).

GTS affects around 1% of the population although prevalence rates vary between reports and populations (Robertson *et al.*, 2009). As for many NDDs, the condition affects males more frequently than females, at a 4:1 ratio. GTS has a strong genetic basis, as suggested by high concordance rates in monozygotic (i.e. 53-56%) compared to dizygotic (i.e. 8%) twins and a family history of tics in up to 60% of cases (Price *et al.*, 1985; Pauls *et al.*, 1991; O'Rourke *et al.*, 2009). In fact, GTS has one of the highest familial recurrence risks with a ten-fold increased risk in first-degree relatives of affected individuals. Non-genetic environmental aspects, such as prenatal stress and insults, perinatal complications and microbial infections, have also been suggested (Robertson, 2000). Additionally, mixed models have been proposed, wherein gene disruptions are necessary for disease although onset

and severity may by modulated by environmental factors. Most research has, nonetheless, focused on genetics and the search for causal genes, loci and alleles. It has, however, proven difficult and no variants of major effect have been uncovered to date. This introductory section outlines our genetic understanding of the disorder and introduces the aims of this chapter.

3.1.1 Chromosomal aberrations

Chromosomal aberrations are abnormalities of larger DNA segments. These include inversions, translocations, deletions and duplications and have important roles in NDDs, such as ASD and ID. Using karyotyping, cytogenetics and microarray profiling, several loci have been linked to GTS with varying degrees of confidence. Four of these map to protein-coding genes: 7q22-q31, 7q35-q36, 13q31.1 and Xp22.3 (Deng *et al.*, 2012).

Translocations and duplications involving the 7q22-q31 region have been observed in four unrelated families or individuals and functional disruption of the affected *IMMP2L* gene was confirmed in two cases (Deng *et al.*, 2012). The exact function of the protein encoded by *IMMP2L* is unknown but appears to be related to inner mitochondrial membrane targeting. This suggests mitochondrial dysfunction, which has been linked to various psychiatric conditions (Jou *et al.*, 2009), may also be implicated in the aetiology of GTS.

Re-arrangements of the 7q35-q36 region, overlapping the *CNTNAP2* gene, have been observed in three unrelated GTS families or individuals (Deng *et al.*, 2012). The protein encoded by the *CNTNAP2* gene belongs to the neurexin family of adhesion molecules and receptors and localises to the juxtaparanodes of myelinated axons. *CNTNAP2* represents a good candidate gene given its association with ASD (Arking *et al.*, 2008; Alarcón *et al.*, 2008) and language disorders (Vernes *et al.*, 2008; Newbury *et al.*, 2011). Nonetheless, its role in GTS remains unclear, as it is not causal in one of the families with 7q35-q36 translocations.

A *de novo* inversion of the q31.1 region of chromosome 13, nearby the *SLITRK1* gene, was reported in a GTS patient (Deng *et al.*, 2012). Additionally, rare and common variants in this gene have also been identified (Abelson *et al.*, 2005; Miranda *et al.*, 2009). The protein encoded by *SLITRK1* has roles in neuronal survival, neurite outgrowth and synaptic formation (Deng *et al.*, 2012), supporting an association with GTS. Nevertheless, while mouse knockout models of the orthologue of *SLITRK1* display behavioural abnormalities, such as elevated anxiety and immobility (Katayama *et al.*, 2010), these cannot readily be linked to GTS.

Finally, a deletion of the Xp22.3 region overlapping the *NLGN4* gene was observed in a family with two brothers, one with ASD and chronic tics and the other with ADHD and GTS (Lawson-Yuen *et al.*, 2008). *NLGN4* encodes a postsynaptic cell adhesion molecule from the neuroligin family, which is involved in synaptic remodelling and synaptogenesis. *NLGN4* has been linked

to ASD, ID and developmental delay, suggesting a shared aetiology between GTS and other NDDs. Hence, these disorders could converge on synaptic connectivity dysfunction, resulting from impaired targeting of cell adhesion proteins, such as neurexin and neurologins, to pre- and postsynaptic membranes, respectively.

3.1.2 Linkage studies

A linkage study of a two-generation pedigree family (Ercan-Sencicek *et al.*, 2010) identified a nonsense mutation in the *HDC* gene, which encodes a protein involved in histamine biosynthesis. Although the role of histamine biosynthesis in the brain is poorly understood, histamine is produced by neurons and signals primarily through G-coupled receptors, which are abundant in the human striatum and cortex. Changes in size and activity of the putamen and caudate (Stern *et al.*, 2000; Peterson *et al.*, 2003) and disruptions of the cortico-striatal circuitry (Albin and Mink, 2006) have been reported in GTS patients, suggesting a role for both brain regions and histamine signalling dysfunction in disease. Nevertheless, no clear GTS-related behaviours have been observed in the *hdc* knockout mice, which display decreased locomoter activity at night (Kubota *et al.*, 2002). Moreover, mutations in *HDC* have not been observed outside the aforementioned family, hence challenging the link between *HDC* and GTS.

3.1.3 Candidate gene studies

Neuroimaging, post-mortem and pathophysiological studies have implicated the dopaminergic and serotonergic systems in GTS aetiology. Dopamine is a neurotransmitter involved in signalling from the substantia nigra to the striatum. The observation of increased dopamine release in GTS patients suggests a pathogenic role for abnormal nigrostriatal dopaminergic activity. This association is further supported by the observation that tics intensify following dopamine stimulation by agents such as L-dopa and are alleviated following reduction of dopamine levels at synapses or blockage of D2 dopamine receptors (Leckman *et al.*, 2010).

The serotonergic system also makes a plausible suspect in the aetiology of GTS, based on its possible involvement in the pathogenesis of OCD. A role for this system is supported by the observation that the post-mortem basal ganglias from GTS patients display low levels of serotonin and other related neurotransmitters (Leckman *et al.*, 2010). Nonetheless, the evidence is less than for the dopaminergic system and other post-mortem studies find no link between the serotonergic system and GTS (Nikolaus *et al.*, 2009; Yoon *et al.*, 2007).

The dopaminergic and serotonergic systems are considered the most promising suspects for a role in the pathophysiology of GTS although others, such as the histaminergic (see above), non-adrenergic and cholinergic

systems, have also been suggested and underpinned by variable lines of evidence (Leckman *et al.*, 2010).

Several genes involved in the aforementioned systems have been investigated as susceptibility genes. These include dopamine receptor genes (*DRD1*, *DRD2*, *DRD4* and *DRD5*), for which evidence has been conflicting and positive results unreplicated. The *MAO-A*, *DAT1* and *TPH2* genes, encoding proteins with roles in serotonin/dopamine inactivation, dopamine neurotransmission and serotonin brain synthesis, respectively, have gathered more promising findings. Indeed, gene polymorphisms in *MAO-A* and *DAT1* (1 Single Nucleotide Polymorphism; SNP) have been associated with GTS and ADHD, whereas a variable number of tandem repeats in *TPH2* has been linked to GTS and OCD (Paschou, 2013). Studies of candidate genes have, nevertheless, been inconsistent, failing to uncover genes of large effects, possibly owing to limited study power associated with small sample size and number of variants explored.

3.1.4 Genome-wide association studies

The first GTS Genome-Wide Association Study (GWAS) (Scharf *et al.*, 2013), undertaken in a sample of 1,285 cases and 4,964 ancestry-matched controls found no genome-wide significant signals. Nonetheless, top-scoring nominal signals included an intergenic SNP located between the *SLITRK1* and *SLITRK6* genes, further supporting an association between *SLITRK1* and GTS (Scharf *et al.*, 2013). The 42 top-scoring SNPs (with smallest nominal p-

values) were subsequently genotyped in an independent cohort of over 600 cases and controls and the associated meta-analysis validated a SNP in the proximity of the *NTN4* gene, encoding an axon guidance molecule, expressed in the developing striatum (Paschou *et al.*, 2014). More recently, an epigenome-wide association study was undertaken in a sample of over 1,600 individuals, but no association was genome-wide significant (Zilhao *et al.*, 2015). Nonetheless, top-scoring nominal signals were identified within methylation sites nearby genes implicated in neurological disorders, such as *GABBR1*, *BLM*, and *ADAM10* (Zilhao *et al.*, 2015).

3.1.5 Copy number variant studies

Structural variation, such as Copy Number Variation (CNVs; deletions or duplications of DNA segments > 1 kb), underlies a significant proportion of genetic variation and its importance in the aetiology of NDDs has been increasingly recognised. In the last few years, several genome-wide CNV scans revealed that GTS patients harbour an excess of CNVs recurrently observed across NDDs (termed neurodevelopmental CNVs as they predispose to many NDDs) compared to healthy individuals, suggesting an overlapping aetiology between GTS and ASD, OCD and ADHD (Sundaram *et al.*, 2010; Fernandez *et al.*, 2012; Nag *et al.*, 2013; McGrath *et al.*, 2014). Additionally, rare recurrent and *de novo* (*dn*) CNVs observed in GTS patients overlap several NDD-associated genes. These include *NRXN* (encodes a neurexin-like cell adhesion molecule implicated in ASD, mental retardation and schizophrenia), *CNTNAP2*, *CTNNA3* (encodes a cell adhesion protein

linked to ASD and schizophrenia), *CTTNBP2* (encodes a cell adhesion protein involved in ASD) as well as the 22q22.12 region, which disrupts the *KCNE1*, *KCNE2* and *RCAN1* genes, implicated in Down's syndrome (Sundaram *et al.*, 2010; Fernandez *et al.*, 2012; Nag *et al.*, 2013; McGrath *et al.*, 2014; Huang *et al.*, 2016). Novel recurrent CNVs have also been identified in these patients, although most do not reach statistical significance or remain unreplicated. Nonetheless, CNV deletions overlapping the *AADAC* deacetylase gene, initially observed in two GTS patients (Sundaram *et al*, 2010), were subsequently identified in eight individuals from an independent cohort (Bertelsen *et al.*, 2016), suggesting *AADAC* is a promising susceptibility gene.

3.1.6 Whole-exome sequencing studies

D*e novo* gene point mutations, termed Single Nucleotide Variants (SNVs), have strongly been associated with NDDs. Recently, the first whole-exome sequencing study of GTS cases revealed patients harbour an excess of likely damaging *de novo* SNVs compared to healthy individuals, in line with observations made for ASD and ID (Iossifov *et al.*, 2014; Hoischen *et al.*, 2014; Shohat *et al.*, 2017; Stessman *et al.*, 2014; Willsey *et al.*, 2017). Specifically, SNVs predicted to disrupt the functionality of the proteins encoded by the affected genes, namely frameshift and missense mutations, were more frequently observed in cases (Willsey *et al.*, 2017). These results were further replicated in an independent GTS cohort (Willsey *et al.*, 2017). In fact, examination of the combined cohort predicted likely damaging *de novo*

SNVs (i.e. nonsense, frameshift, splice-site and missense mutations) contribute to GTS risk in up to 12% of cases and identified four potential disease susceptibility genes (Willsey *et al.*, 2017). These include (1) *WWC1*, which encodes a phosphoprotein involved in memory performance and cognition (Zhang *et al.*, 2014), (2) *NIPBL*, associated with the Cornelia de Lange syndrome, as well as (3) *CELSR3* and (4) *FN1* encoding a cadherin and glycoprotein, respectively, with possible roles in neurite growth (Willsey *et al.*, 2017).

3.1.7 Conclusions

Over the past decade, pedigree, chromosomal, genome-wide and exome sequencing studies of GTS patients led to the identification of putative genes and genomic loci associated with the disorder. Such studies have helped shaping the genetic landscape of GTS and inspired many neurophysiological analyses. However, the overall interpretation of the results has remained difficult, rendering the genetic understanding of the disorder incomplete. The nature of the studies, small sample sizes, ethnic heterogeneity, phenotypic variability and the complex, polygenic nature of the disorder may explain the lack of reproducibility of most findings. Nonetheless, the recent advent of high throughput methods enables the screening of large numbers of patients, making the discovery of susceptibility genes and loci as well as the establishment of their effect sizes more feasible. This will undoubtedly broaden our understanding of the genetic aetiology of GTS, which in turn may lead to the development of effective new treatment strategies.

3.2 Chapter aims

To provide insights into the genetic architecture of GTS, we performed a genome-wide CNV analysis using a Danish GTS cohort of 261 patients and 1,053 healthy controls. We aimed to (1) compare the occurrence of rare CNVs between cases and controls, (2) identify CNVs more frequently harboured by patients than controls and (3) assess the presence of loci and genes previously observed in GTS and other NDDs. Additionally, we sought to (4) characterise the functions of genes mapping to GTS CNVs and (5) evaluate their possible involvement in disease.

3.3 Specific Materials and Methods

The work described in **sections 3.3.1 and 3.3.2** was performed by Zeynep Tümer's group, Kennedy Center, Department of Clinical Genetics, The Juliane Marie Centre, Copenhagen University Hospital, Denmark and their collaborators. The work described in **sections 3.3.3 to 3.3.8** was performed in collaboration with Shanmukha Sampath Padmanabhuni, Department of Molecuar Biology and Genetics Democritus, University of Thrace, Alexandroupoli, Greece.

3.3.1 Study Subjects

In total, 261 patients were recruited as part of a large study at the Tourette Clinic, Copenhagen University Hospital, Herlev (Mol Debes *et al.* 2008). Clinical diagnosis of GTS was established according to criteria from the Diagnostic and Statistical Manual of mental disorders, fourth edition, Text Revision (DSM-IV-TR) (American Psychiatric Association 2000). Control subjects comprise 1,053 unrelated Danish healthy adults, genotyped in deCODE (Mol Debes *et al.*, 2008; Stefansson *et al.* 2009).

GTS cases comprise 208 males and 53 females. Controls include 649 males and 403 females, revealing a male bias among cases, consistent with an increased male compared to female prevalence of GTS ($p = 1.88 \times 10\text{-}8$, Odds Ratio, OR = 2.44; two-sided Fisher test).

3.3.2 Genotyping

Genomic DNA was extracted from peripheral blood using standard protocols. GTS subjects were genotyped using the Affymetrix CytoScanHD array (Affymetrix Santa Clara, USA), which includes 750,000 SNP markers and ~1.9 Million single-locus Copy Number markers. Of the 1,053 control subjects, 1,035 and 18 were genotyped using Illumina HumanOmniExpress-12 v1.0 and v1.1 chips, respectively. Both chips cover approximately 730,000 SNPs.

3.3.3 Genotype data pre-processing

Raw CEL files of Danish GTS cases were pre-processed using Affymetrix Power Tools (APT, version 1.6) to estimate the median raw intensity of single-locus copy number markers and SNPs and perform quantile normalisations. Intensities were further summarised using logarithms of quantile-normalised intensity values. The APT software performs SNP Quality Control (SNPQC) and estimates genotype calling rates. To minimise noise (i.e. processing batch effects), samples with *median absolute values of all pairwise differences* > 0.25 and with *waviness standard deviation* > 0.15 were removed. Based on these criteria, ten cases and no controls failed pre-processing Quality Control (QC) metrics, leaving 251 and 1,053 cases and controls (**Table 3.1**).

Table 3.1 | Per sample quality control.

	Danish GTS	Danish Controls
Total pre QC	261	1053
Pre-processing QC	10	0
Discordant sex	0	0
Genotyping call rate < 98%	0	2
Heterozygosity ± 0.05 sd	5	8
Duplicated or related individuals	1	26
European Outliers	4	11
Individuals with > 65 CNV calls	1	8
Individuals with cumulative CNV length > 10Mb	0	1
Total post QC	240	997

sd = standard deviation

Number of Danish GTS cases and controls removed at each QC step.

3.3.4 Per individual quality control

To circumvent the introduction of false positives and noise, we removed individuals with (1) discordant sex information, (2) low genotyping success rates and abnormal heterozygosity rates, (3), duplicated or related subjects and (4) those showing divergent ancestry (**Table 3.1**). QC steps were performed in PLINK (Purcell *et al.*, 2007).

The sex of an individual can be determined by evaluating the heterozygosity of SNPs located on the X chromosome. Because males only detain one copy of the X chromosome, they should be homozygous for all markers tested, while females are expected to have homozygosity rates of less than 0.2. We measured the F-statistic for each sample and to account for genotyping errors, F values < 0.25 and > 0.75 were assigned to females and males,

respectively. F values between 0.25 and 0.75 were indicative of discordant sex relative to genetic data. No gender ambiguity was observed in the cohort.

Low genotyping call rates and abnormal heterozygosity rates are indicative of poor sample quality or contamination. The average genotyping call rates of cases and controls were 99% and 96%, respectively. We removed samples with > 2% of missing markers (two controls) and with heterozygosity rates ± 0.05 standard deviations from the mean (five cases and eight controls) (**Table 3.1**).

The presence of duplicated or related individuals in a cohort may lead to the over-representation of genotypes from the associated families. As a result, genotype frequencies in the evaluated cohort may no longer be representative of the general population. The degree of shared ancestry between pairwise combinations of individuals was measured as the *Identity By Descent* (IBD) metric. We removed one sample from each pair of individuals showing relatedness, as given by PLINK metrics Z1 > 0.2 and pi-hat > 0.1 (under these thresholds, IBD < 0.185, suggesting individuals from our cohorts are beyond second-degree relatives). Z1 estimates the probability that the two individuals share 1 allele by descent (Z1 = 0.5 for full siblings) while pi-hat provides a global measure of IBD, such that unrelated individuals have pi-hat values close to 0. In total, one Danish case and 26 controls were removed from downstream analyses (**Table 3.1**).

As particular alleles occur at different frequencies between populations, we sought to only consider individuals of European ancestry. We performed principal component analysis of Danish cases and controls, separately, using the EIGENSTRAT software (Price *et al*., 2006). We removed four cases and eleven controls with significant deviations from the European cluster identified using 1000 Genome reference samples (1000 Genomes Project Consortium, 2015). Outliers were defined as those with by principal component 2 scores < 0.05 for cases and principal component 1 scores > 0 and 2 scores > 0.01 for controls (**Fig 3.1; Table 3.1**).

In total, ten Danish GTS cases and 47 controls were removed, leaving 241 cases and 1,006 controls for downstream analyses (**Table 3.1**).

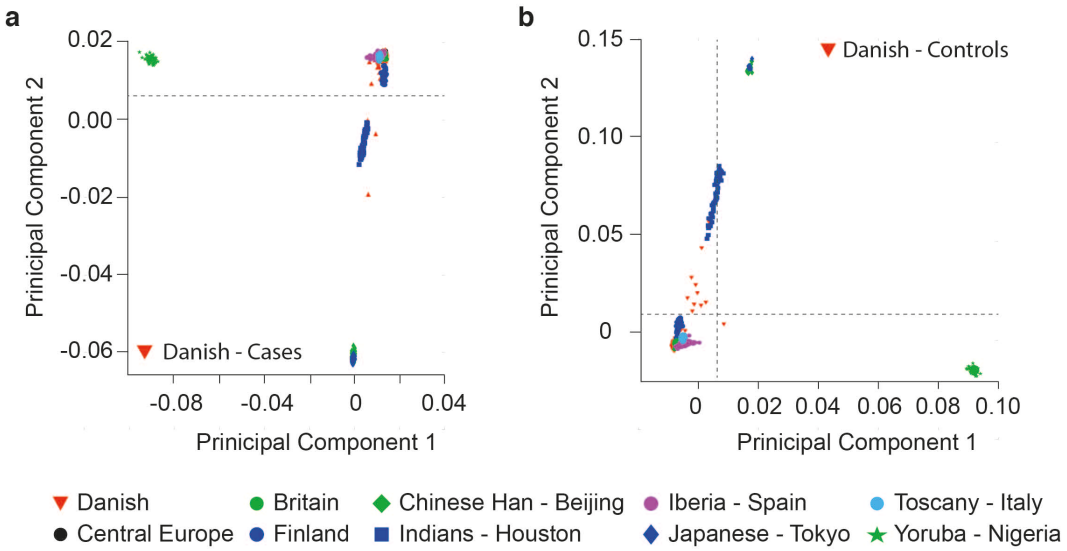


Fig 3.1 | Ancestry clustering. Principal component analysis of genotype data from Danish (**a**) GTS cases and (**b**) healthy controls (red triangles). Populations from the 1000 Genome are used as a reference panel (Genomes Project Consortium, 2015). Individuals lying outside of the primary European cluster are removed, as given by (**a**) principal component 2 < 0.05 and (**b**) principal component 1 > 0 or principal component 2 > 0.01.

3.3.5 Per SNP quality control

By implementing the same QC measures in PLINK (Purcell *et al.*, 2007), as Scharf *et al.*, 2013, we removed SNP markers, which could generate false positives (**Table 3.2**). These included (1) monomorphic SNPs, (2) SNPs lacking chromosomal location information, (3) SNPs with AT/CG alleles, lacking strand information, (4) SNPs with genotyping rates < 98% and (5) SNPs with significant deviations from the Hardy-Weinberg equilibrium. The Hardy-Weinberg equilibrium states that allelic and genotype frequencies in the population remain constant across generations. This means that for a given SNP with alleles **A** and **a** of frequencies p and q, respectively, the frequency of each genotype is given by freq(**AA**) = p^2, freq(**aa**) = q^2 and freq(**Aa**)= $2pq$. While deviations from these frequencies could indicate the associated alleles are subject to purifying selection, they are more likely to be due to genotype calling errors. We, therefore, removed all SNPs with significant deviations at p < 1 x 10-5. Lastly, (6) we chose a minor allele frequency cut-off of 0.15 as our power to detect associations at low allelic frequency is largely reduced owing to the small sample size.

3.3.6 CNV calling and quality control

The PennCNV software (Wang *et al.*, 2007) was used to predict CNVs with different default settings for cases and controls, which passed QC. For GTS cases, the *Population frequency of B Allele* file, required by PennCNV, was compiled using a publicly available cohort of 820 French-Canadian healthy

controls of European ancestry, genotyped on the same array (i.e. Affymetrix CytoScan HD array) (Uddin *et al.*, 2014).

PennCNV uses a hidden Markov model to call CNVs, which incorporates a (1) *Log R Ratio* providing a measure of magnitude of the combined fluorescence intensity signals across probes and a (2) *B allele frequency* providing the relative ratio of fluorescence signals between allelic probes. Deletions are called when the Log R Ratio is lower relative to the normal copy number state of two and loss of heterozygosity is indicated by a deviation from the heterozygous value of B allele frequency = 0.5. Conversely, duplications are associated with an increase in Log R Ratio relative to the expected copy number state of two and B allele frequencies of 0, 1/3, 2/3 and 1 depict the combination of **AAA**, **AAB**, **ABB** and **BBB** genotypes, given alleles **A** and **B**.

We used *Log R Ratio standard deviation* thresholds of > 0.32 for cases and > 0.35 for controls, leading to the exclusion of three controls. No additional individuals failed other PennCNV metrics, namely -0.1 < *waviness factor* < 0.1 and *B allele frequency drift* > 0.01. Furthermore, we removed individuals, which had over 65 CNV calls (one case and eight controls), and if the aggregate lengths of their CNVs was larger than 10 mb (i.e. one control) (**Table 3.1**). Ultimately, our final study sample consisted of 240 Danish GTS cases and 997 controls.

We, additionally, performed QC on individual CNVs and maintained only CNVs larger than 1 kb and smaller than 10 mb, with a minimum of five SNPs

per CNVs for deletions and ten markers per CNV for duplications. We merged adjacent CNVs calls with gaps spanning less than 10% of their combined lengths, as these were likely to have been artificially split. Furthermore, we removed CNVs, if > 50% of their length overlapped immunoglobulin regions, centromeres, telomeres and human gaps larger than 200 kb, as described by the UCSC genome browser (Kent *et al.*, 2002).

CNVs were classified as rare if < 50% of their length overlapped regions present at > 1% in the Database of Genomic Variants (MacDonald *et al.*, 2014) or in the cohort itself. Conversely, common CNVs are present at frequencies > 1% in the Database of Genomic Variants. In total, 7,272 CNVs were defined as rare.

We considered all protein-coding and RNA genes, as given by Ensembl Biomart (Cunningham *et al*., 2015) and assigned genes to CNVs when these overlapped > 2 bases from each protein-coding or RNA transcript of that gene, based on hg18 genome coordinates from the UCSC genome browser (Kent *et al.*, 2002). In total, rare CNVs ≥ 20 kb from cases and control affect 300 and 1,250 distinct genes, respectively.

Table 3.2 | Per SNP quality control.

	Danish GTS	Danish Controls
Platform	Affymetric	Illumina
Total	749,157	719,665
SNPs on sex chromosomes	22,104	19,868
Monomorphic SNPs	73,639	41,553
Indels	0	1
Genotyping call rates < 98%	62,023	15,534
Significant deviation from Hardy-Weinberg equilibrium (p < 1 x 10-5)	389	592
Common SNPs	128,527	
AT/CG SNPs	1,671	
Absoulte MAF > 0.15 between platforms	38	
Total remaining	460,766	511,881

Number of SNPs removed at each QC step.

3.3.7 Rare CNVs and platform bias check

A publicly available cohort of French-Canadian healthy controls of European ancestry genotyped on the Affymetrix CytoScanHD platform (Uddin *et al.*, 2014) was analysed using the same QC and CNV calling steps as the Danish GTS and control cohorts.

Table 3.3 | Number of common CNV calls stratified by size and platform biases.

CNV size	Danish GTS Cases		Danish Controls		Test	P-value
	CNVs per Sample Mean (SD)	Samples	CNVs per Sample Mean (SD)	Samples		
All common CNVs	23.26 (6.32)	--	5.9 (3.51)	--	t-test	< 2.2E-16
size > 50 kb	6.04 (1.84)	--	2.38 (1.89)	--	t-test	< 2.2E-16
size > 100 kb	3.65 (1.46)	--	1.66 (1.08)	--	t-test	< 2.2E-16
size > 500 kb	--	81	--	34	χ^2	0.0001
size > 600 kb	--	37	--	21	χ^2	0.0001
size > 1 mb	--	15	--	6	χ^2	0.0001

To ensure our analysis is not biased by the use of cohorts genotyped on different platforms, we compared the number of CNV calls between the Danish and the French-Canadian controls. We observed a heavy platform bias (**Fig 3.2a-b**). The separate examination of common and rare CNVs revealed that the observed platform bias is primarily associated with common CNVs and persists when varying the minimum CNV size considered (**Fig 3.2d; Table 3.3**). On the other hand, the platform bias is largely reduced for rare CNVs and disappears for rare CNVs ≥ 20 kb (**Fig 3.2d; Table 3.4**). We, thus, performed the study using the 2,350 rare autosomal CNVs ≥ 20 kb. Case CNVs consist of 205 deletions and 299 duplications. Conversely, control CNVs comprise 1,055 losses and 791 gains, revealing an excess of deletions ($p = 6.62 \times 10\text{-}11$, Odds Ratio, OR = 1.94; two-sided Fisher test). The origin of this bias is, however, unclear.

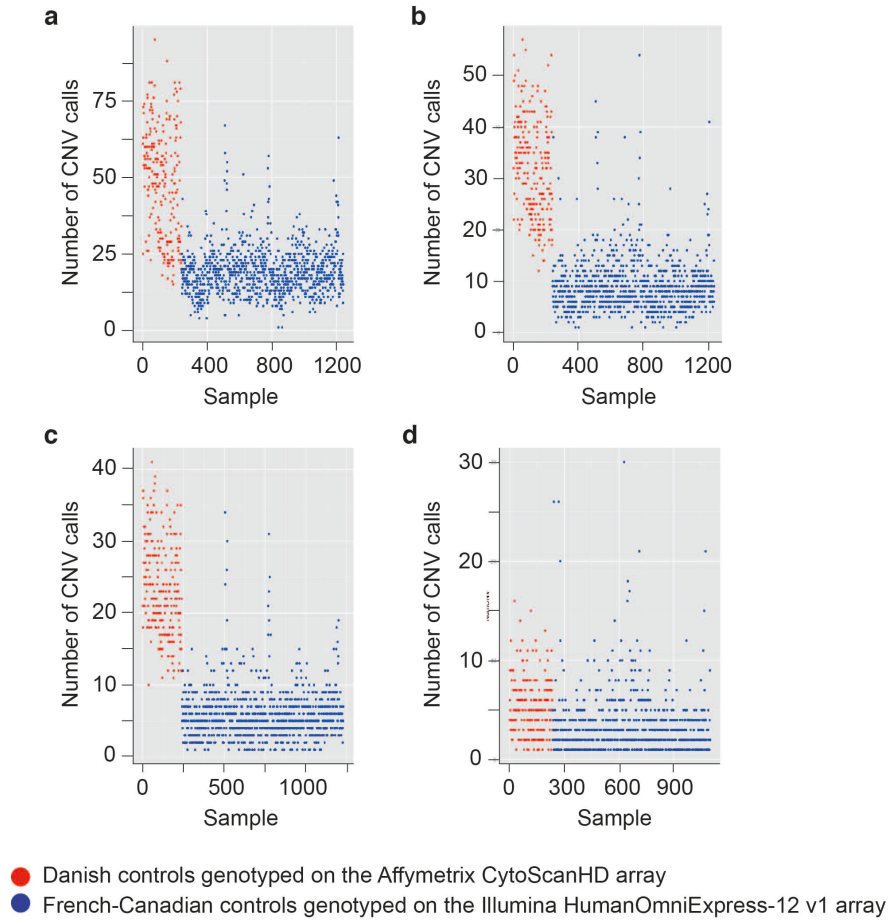


Fig 3.2 | CNV calls and platform bias. (**a and b**) The number of CNVs calls in controls, genotyped on the (red) Affymetrix CytoScanHD and (blue) Illumina HumanOmniExpress-12 v1 arrays, when evaluating (**a**) all pre- or (**b**) post-QC samples. (**c and d**) The number of calls when considering post-QC (**c**) common and (**d**) rare CNVs only, defined as those occurring in > 1% and < 1% of individuals from the database of genomic variants, respectively (MacDonald *et al.*, 2014).

3.3.8 CNV burden and association analyses

Burden analysis was performed using a permutation-based method in PLINK implemented using the *-cnv-indiv-perm* command. We considered only CNVs ≥ 20 kb as no global CNV burden was observed (differences in rare CNV burden, size and gene density have not been reported in GTS (Fernandez *et al.*, 2012; McGrath *et al.*, 2014; Huang *et al.*, 2016)) (**Table 3.4**). Segmental

association and regional tests were carried out in PLINK using a permutation-based implemented using the *-mperm* command.

Reasoning that a true biological signal observed between the Danish cases and controls should not be significant between Danish and French-Canadian controls, we cross-validated the results from the association analysis and regional tests using French-Canadian and Danish controls. Specifically, we ensured that case-associated signals were not significant when repeating the analyses using French-Canadian controls instead of GTS cases.

Table 3.4 | Burden of rare CNVs ≥ 1 kb and ≥ 20 kb.

	All Rare CNVs			Rare CNVs > 20kb		
	Cases	Controls	p-value	Cases	Controls	p-value
Number of CNVs	4.32	3.109	1.00E-05	2.10	1.85	0.525
Proportion of samples with ≥ 1 CNV	1	1	1	0.943	0.947	0.653
Total kb length spanned	253.8	267.1	0.637	259.9	279.1	0.693
Average SNVsize	62.24	93.25	0.999	46	56	0.753
Number of genes spanned by CNVs	4.14	4.02	0.385	0.88	1.0	0.671
Number of CNVs with ≥ 1 gene	0.93	0.798	1.00E-05	0.810	0.727	0.0058
Number of genes per total CNV Kb	0.04	0.029	0.055	0.019	0.019	0.542

3.2.9 GTS CNVs from Fernandez *et al.*, 2012

We downloaded all rare autosomal CNVs ≥ 20 kb and < 10 mb from GTS patients from an independent cohort described in Fernandez *et al.*, 2012. In total, 642 distinct genes mapped to these CNVs (herein termed *Fernandez CNV set*), of which 84 occurred within *dn* CNVs (**Table 2.1**).

3.4 Results

3.4.1 Summary of data

Of the 261 Danish GTS cases and 1,053 Danish controls constituting the study sample, 240 cases and 997 controls were maintained following genotyping and CNV quality control. In total, 31,037 CNVs called on the post-QC sample passed quality control. Of these, we considered 2,350 rare autosomal CNVs ≥ 20 kb, occurring in less than 1% of the population. Overall, cases and controls harbour on average 2.1 and 1.85 rare CNVs, respectively. These do not differ in size and span a median of 46 and 56 kb, respectively (**Table 3.4**). Furthermore, the gene density of case and control CNVs is similar, such that case and control CNVs overlap up to 16 and 26 genes with an average of 0.88 and one gene per CNV, respectively.

3.4.2 Recurrence and disease gene analysis

One genome-wide significant segment was detected on chromosome 14, present in five cases and no controls (Benjamini-Hochberg correction, BH p = 0.0127). This CNV overlaps a long non-coding RNA gene, *LINC00871* (BH p = 0.0037), with no prior association to disease. Regional tests on rare autosomal CNVs showed four genome-wide significant genes (**Table 3.5**). These include (1) *LINC00871*, (2) the mitochondrial membrane protein encoded by *TARP* (BH p = 0.032), (3) *ZNF277* (BH p = 0.032), encoding a putative transcriptional regulator with a possible association with specific

language impairment (Ceroni *et al.*, 2014) and (4) *DLG2* encoding a postsynaptic density protein (BH p = 0.047). *DLG2* represents a plausible candidate gene, as it has previously been observed in ASD (Egger *et al.*, 2014) and schizophrenia (Kirov *et al.*, 2012; Fromer *et al.*, 2014). Nevertheless, the relevance of this signal is unclear given the lack of functional validation. Lastly, *MYT1L*, a gene previously implicated in ID, is nominally more commonly affected by CNVs in GTS cases than controls (nominal p = 1 x 10-3, n = 3 patients), suggesting a possible comorbidity with GTS.

Table 3.5 | Recurrently disrupted genes in GTS patients.

Chr	Gene	CNV start	CNV stop	# Cases	# Controls	*p* - only Danish	*p* – French Canada
7	TARP	38299243	38313248	4	0	0.0323	1
7	ZNF277	111846642	111983989	4	0	0.0323	1
11	DLG2	83166055	85338314	4	0	0.047	1
14	LINC00871	46533361	46971104	5	0	0.0037	0.48

GTS patients from our cohort did not carry CNVs overlapping the *IMMP2L*, *SLITRK1*, *HDC*, or *CNTNAP2* genes, previously implicated in GTS through linkage or CNV studies (Deng *et al.,* 2012). Similarly, CNVs overlapping the *WWC1*, *NIPBL*, *CELSR3* or *FN1* genes, recurrently mutated by *de novo* SNVs in GTS patients, were not observed in our cohort (Willsey *et al.*, 2017). Similarly, CNVs disrupting the *NRXN1* and *CNTN6* genes*,* which were recurrently observed in an independent European GTS cohort (Huang *et al.*, 2016), were not found in our cases. Nonetheless, 62 genes mapping to CNVs from our Danish GTS (herein termed *case genes*) are present in an

independent Caucasian GTS cohort (i.e. *Fernandez* dataset; **Table 2.1**) (Fernandez *et al.,* 2012). Although this does not represent a significant enrichment (p = 0.46, OR = 1.13; two-sided Fisher's test), two case genes, *PANK2* and *RNF24,* map to a *dn* CNV (20p13) from the Fernandez dataset. *PANK2* encodes a mitochondrial panthothenate kinase, which is involved in co-enzyme A biosynthesis and, thus, has various roles in energy, fatty acids, neurotransmitter and glutathione metabolisms (Hayflick, 2003). Various types of *PANK2* mutations cause Pantothenate Kinase-Associated Neurodegeneration (PKAN; MIM:234200) and Hypoprebetalipoproteinemia, Acanthocytosis, Retinitis pigmentosa and Pallidal degeneration (HARP syndrome, MIM:607236; less severe form of PKAN) in an autosomal recessive manner. PKAN is a progressive neurodegenerative disorder, characterised by progressive iron accumulation in the basal ganglia. Symptoms often include gait, motor and speech abnormalities, as well as variable psychiatric and cognitive symptoms including depression, ID and OCD. Tourette motor and vocal tics have been observed in early stages of PKAN (Gregory *et al.*, 2009). As little is known about the putative transcription factor encoded by *RNF24* and no disease associations have been made to date, *PANK2* currently constitutes a better CNV candidate gene. Furthermore, a *dn* CNV deletion overlapping *PANK2* was observed in an ADHD patient (Williams *et al.*, 2010). We note that, in our cohort, the associated CNV deletion overlaps a third gene, *MAVS*, which encodes a protein involved in the innate immune defense (interferon signalling; Seth *et al.*, 2005) and lacks known disease associations.

3.4.3 Functional pathway and expression analyses

To elucidate the molecular functions affected in GTS, we performed functional enrichment analyses of genes overlapped by CNVs observed in our GTS cohort (case genes) using Gene Ontology (GO) annotations. While no GO terms are over-represented compared to the whole genome, case genes are associated with transcriptional regulation compared to genes disrupted by CNVs in Danish controls (herein termed *CT* genes) (**Fig 3.3a**). Accordingly, enriched GO terms include *regulation of transcription, DNA-templated* (GO:0006355; BH p = 0.0325), *DNA binding* (GO:0003677; BH p = 0.00158) and *RNA biosynthetic process* (GO:0032774; BH p = 0.0325). Given functional clustering in the genome (Andrews *et al.*, 2015a), we sought to ensure enrichments are not driven by large multigenic CNVs. We repeated the analysis using only genes overlapped by singleton CNVs and observed similar enrichments (**Fig 3.3b**). Conservatively, comparing the number of GTS patients and CT controls with hits in these pathways revealed an excess of patients with ≥ 1 CNV genes annotated with each term ($0.01 < p < 0.15$ and $1.2 < OR < 1.6$) (**Table S3.1**). Taken together, these results indicate that dysfunction of transcriptional regulation is a theme globally observed in this GTS cohort.

Stratifying CNVs by size and direction revealed the observed excess of transcriptional regulators is specific to CNV duplications ranging between 100 and 500 kb (**Table S3.2**). Conversely, CNVs ≤ 100 or > 500 kb and deletions of all sizes do not overlap genes enriched for any GO terms.

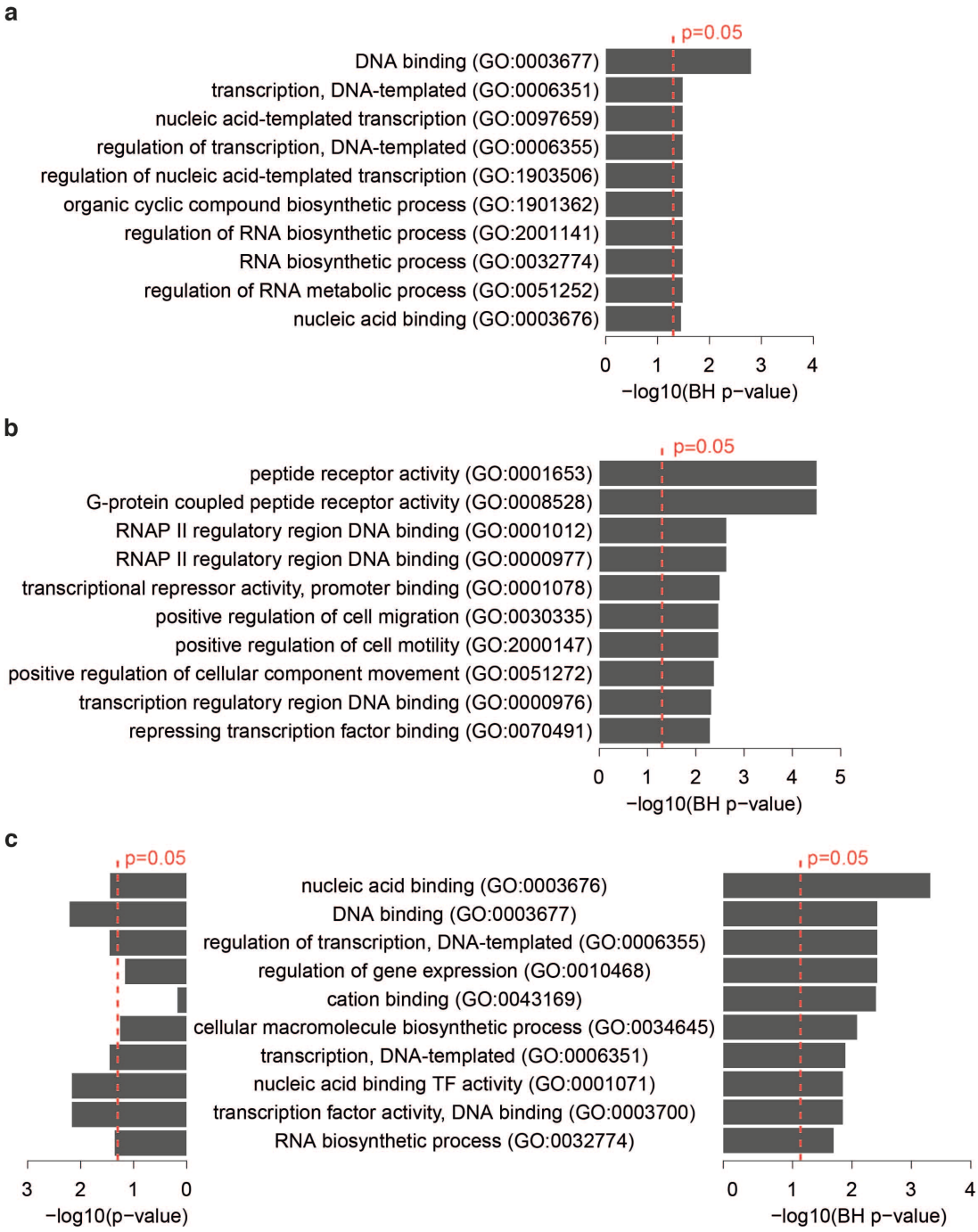


Fig 3.3 | GO enrichment analysis. Over-representations of GO terms among (**a**) all case genes, (**b**) case genes overlapped by singleton CNVs (i.e. overlapping one gene) and (**c**) GTS CNVR co-expressed genes (discussed in **section 3.4.4**) genes. CNVR, CNV region. Significance was assessed using right-tailed hypergeometric tests and a correction for multiple testing was applied using the BH method. The backgrounds consist of (**a, b**, **c – right panel**) all control (CT) genes and (**c – left panel**) CT CNVR co-expressed genes. The figures show the top 10 enriched GO terms. Of note, *RNAP*, *TF* and *DNA binding* refer to RNA polymerase, transcription factor and sequence-specific DNA binding (except for GO:0001012), respectively. The full description of GO:0001078 is *transcriptional repressor activity, RNA polymerase II core promoter proximal region sequence-specific binding*.

3.4.4 Subset of genes more likely to be implicated in GTS aetiology

The study of the ensemble of case genes from this cohort generated limited evidence for the association of particular CNVs, genes or biological pathways with the disease. Recognising this could result from our inability to differentiate between neutral and causal CNVs and genes, we next sought to identify a subset of case genes more likely to be involved in the aetiology of GTS.

Our analysis relies on the assumption that (1) the disruption of the same biological pathways in patients from this cohort underlies their shared disease phenotype (Oti and Brunner, 2007). Under the complementary assumption that (2) that genes expressed at the same time in the same tissues are more likely to work in the same functional processes (Lee *et al.*, 2004), we evaluated the similarity of expression patterns of case genes. For this, we used a gene co-expression network constructed using GTEx data, wherein pairs of genes were assigned an overall correlation coefficient, reflecting the similarity of their expression patterns across various bodily tissues (**section 2.4.1**). We performed the analysis using informative network links only, defined as those with co-expression coefficients $r > 0.6$, confirming our results at $r > 0.5$ and $r > 0.7$ (**Table 2.3**). Case genes display more similar expression patterns than 1,000 randomised gene sets, constructed by randomly sampling the same number of genes from the genome ($p = 0.002$) (**Fig 3.4a**) or from CT genes ($p = 0.001$) (**Fig 3.4b**). To minimize confounding effects (Honti *et*

al., 2014), randomised sets were matched for coding DNA sequence lengths and network connectivities.

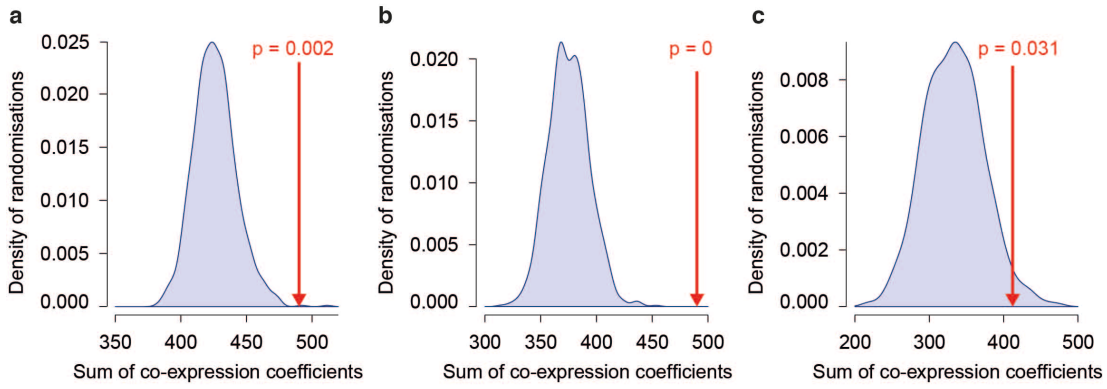


Fig 3.4 | Case genes are co-expressed across bodily tissues. Similarity of body-wide expression patterns, within the GTEx co-expression network, of case genes (red arrow) compared to (**a**) random genome genes or (**b, c**) control (CT) genes (blue density distribution of sum of co-expression coefficients). P-values were derived using a permutation-based approach (**section 2.5.2**) (**a, b**) Randomised genes sets were matched for coding DNA sequence length and network connectivity (Honti *et al.*, 2014). (**c**) Randomised CT genes were matched for genomic adjacency (i.e. they mapped to the same number of CNV regions (CNVRs) as case genes and were matched for number of genes per CNVR).

By merging all overlapping CNVs, we constructed 423 and 1,173 genomic regions (CNV Regions; CNVRs) disrupted in GTS and CT individuals, respectively (herein termed *GTS* or *CT CNVRs*). Of these, 147 case CNVRs and 423 control CNVRs overlap genes. GTS CNVR genes are more strongly co-expressed than those overlapped by CT CNVRs, matched for gene density ($p = 0.031$) (**Fig 3.4c**), hence confirming that functional clustering in the genome did not bias our findings. Although these results were replicated using a minimum expression correlation coefficient threshold between genes of $r > 0.5$ ($p = 0.033$), significance was not reached using $r > 0.7$ ($p = 0.19$) (**Table S3.3**). This could be due to reduced statistical power associated with a

1.26-fold decrease in number of genes present in GTEx networks with r > 0.6 and r > 0.7 (i.e. 205 vs 163 genes, respectively). Given the neurodevelopmental nature of the disorder, we further validated these results using an independent gene co-expression network depicting the similarity of spatiotemporal brain expression patterns, created using Brainspan data ($p \leq 0.005$) (**Table S3.3**).

Of the 299 genes overlapped by GTS CNVRs, 160 are co-expressed with one another within GTEx r > 0.6 (**Fig 3.5**), suggesting this subset is more likely to converge functionally. We termed these 160 genes – overlapped by 93 / 147 genic CNVRs, *GTS CNVR co-expressed genes* as compared to the set of 139 functionally more divergent *GTS CNVR non-co-expressed genes*.

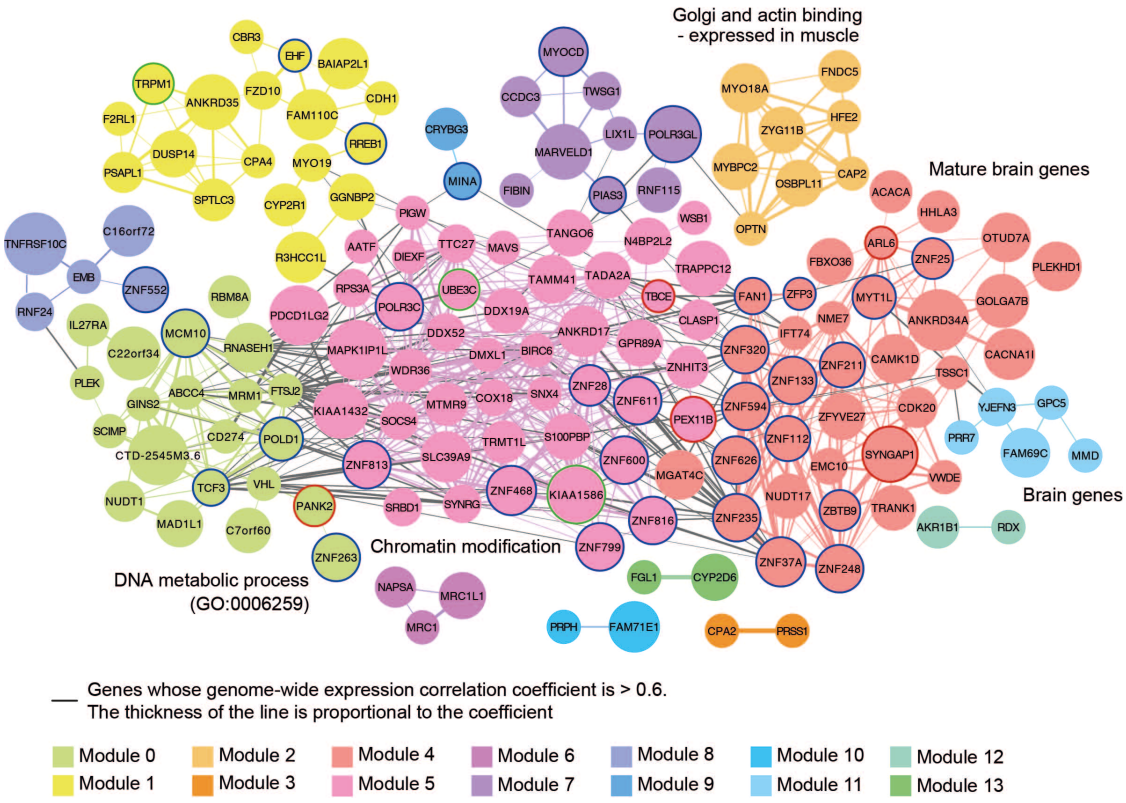


Fig 3.5 | GTS CNVR co-expression network. The co-expression network of 160 / 299 genes overlapped by case CNV regions (CNVRs), displaying similar body-wide expression patterns within the GTEx co-expression network with r > 0.6. The co-expression coefficients of gene pairs is proportional to edge thickness. Given its modular structure (Q = 0.45), the network was separated in modules of strongly co-expressed genes. Module memberships are depicted by the node's (gene's) colour and descriptions are shown, when available. Nodes with blue, red and green borders represent genes encoding transcription factors, high confidence ID and ASD genes, respectively. Of note, *MYT1L* is depicted as encoding a transcription factor but is also considered to be both an ID and ASD candidate gene.

The subset of GTS CNVR genes co-expressed in the GTEx co-expression network gathers the majority of brain-expressed ($p = 7.47 \times 10^{-12}$, OR = 5.11) and brain-specific ($p = 0.051$, OR = 6.28) genes observed in the cohort (**Fig 3.6a**). Correspondingly, this set is strongly expressed throughout the pre- and post-natal brain and has higher probabilities of haploinsufficiency (HIS; Steinberg *et al.*, 2015) compared to GTS CNVR non-co-expressed, CT and genome genes of similar lengths (BH $p < 0.055$) (**Fig 3.7; Table S3.4**). Of

note, GTS CNVR non-co-expressed genes are strongly expressed in the testis, small intestine and kidney compared to random genome genes, matched for CDS (BH $p < 0.03$; not shown, see **section 2.7.2** for methodology), confirming they do not constitute a set of generally lowly expressed genes. When comparing GTS CNVR co-expressed genes to an equivalent set of 848 / 1,249 CT CNVR genes co-expressed within GTEx $r > 0.6$, most enrichments disappear. Nevertheless, nominally higher expression levels persist in the early/mid prenatal brain ($p = 0.016$) and specifically in the striatum ($p = 0.034$) (**Fig 3.7**). These observations are consistent with the neurodevelopmental nature of the disorder and the possible contribution of the basal ganglia and striatum to its aetiology (Saka and Graybiel, 2003; Albin and Mink, 2006; Baym *et al.*, 2008). GO enrichment analysis revealed GTS CNVR co-expressed genes recapitulate the previously observed functions of transcriptional regulation compared to all CT genes or CT CNVR co-expressed genes (**Fig 3.3c**). Correspondingly, chromatin remodelling and histone modifying genes are more strongly co-expressed within GTEx $r > 0.6$ with this set than with GTS non-co-expressed genes, all CT genes and CT CNVR co-expressed genes ($p < 0.002$) (**Fig 3.6b**).

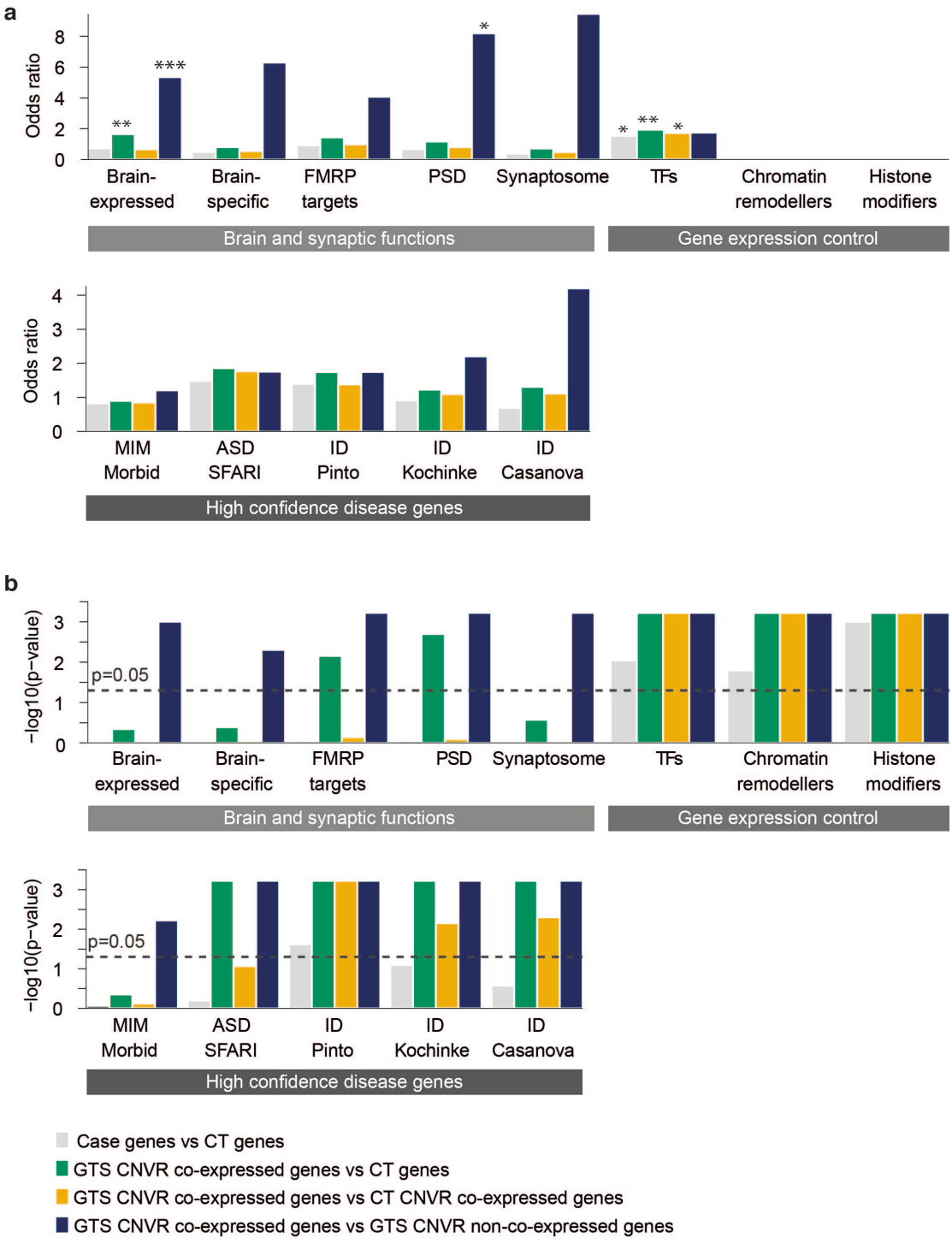


Fig 3.6 | GTS CNVR co-expressed genes are involved in transcriptional regulation. (**a**) Overlap and (**b**) similarity of expression patterns between case genes or GTS CNVR co-expressed genes and various gene sets (**Table 2.1**). (**a**) Overlaps were assessed using right-tailed Fisher tests. (**b**) The GTEx co-expression network with r > 0.6 was used to evaluate similarities of expression patterns using a permutation-based approach (**section 2.5.3**). Of note, in (**b**) all case genes were used as background instead of GTS CNVR non-co-expressed genes (dark blue). CT, control; CNVR, CNV region.

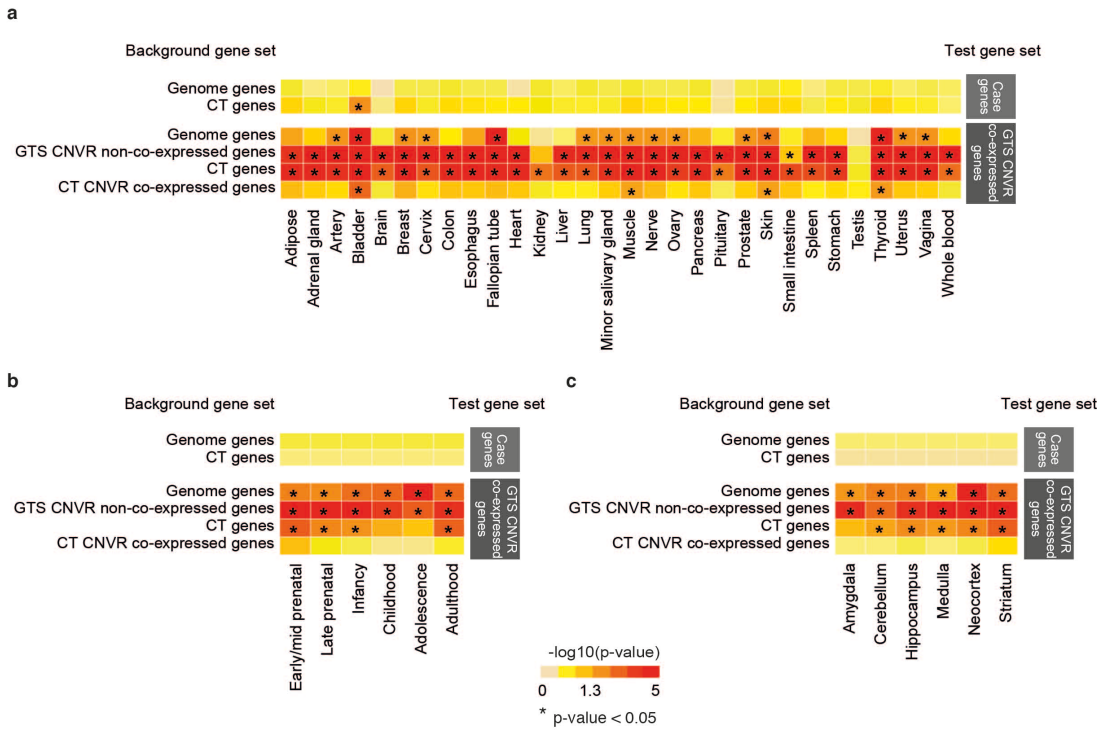


Fig 3.7 | Gene expression dynamics of case and GTS CNVR co-expressed genes. Expression patterns of case genes and GTS CNVR co-expressed genes (shown on the right of each plot) compared to several background genes shown on the y-axes in (**a**) 30 body tissues (GTEx dataset) and in the brain at (**b**) six developmental points and in (**c**) six regions (Brainspan datasets). P-values were derived using a permutation-based approach (**section 2.7.2**) when comparisons were made using genome genes (matched for coding DNA sequence length) or otherwise right-tailed Mann-Whitney U tests. BH corrections for multiple testing were applied to control for the numbers of tissues, stages or regions tested. CT, control; CNVR, CNV region.

The co-expression network formed by GTS CNVR co-expressed genes is modular (network modularity Q = 0.45) and could thus readily be separated using the Louvain algorithm into smaller, highly interconnected sub-networks, termed modules. Accordingly, genes from the same modules display an excess of intra- compared to inter- module links within GTEx r > 0.6 (**Fig 3.5**). GTS CNVR co-expressed genes were partitioned into 14 modules of strongly co-expressed genes, among which 7 included ≥ 5 genes (modules 0, 1, 4, 5, 7, 8 and 11) (**Fig 3.5**). Module 4 comprises genes specifically expressed in the brain and reproductive systems as compared to genes from the genome,

remaining GTS CNVR co-expressed genes, all CT genes and CT CNVR co-expressed genes (**Figs 3.8a, S3.1a, S3.2a and S3.3a**). In fact, module 4 includes brain-expressed ($p = 0.0098$, OR = 5.11) and brain-specific ($p = 0.042$, OR = 3.61) genes (**Table S3.5**) and correspondingly implicates neuronal functions. Indeed, nominally enriched GO terms include *regulation of neuron projection development* (GO:0010975; $p = 0.0047$), *regulation of neuron differentiation* (GO:0045664; $p = 0.010$) and *regulation of cell projection organization* (GO:0031344; $p = 0.011$) (**Table S3.6**). This association is strengthened by the observation that module 4 genes are more strongly co-expressed with FMRP target, postsynaptic density and synaptosome genes than remaining GTS and CT CNVR co-expressed genes ($p < 0.01$) (**Figs 3.8d and S3.4**). The involvement of such brain genes in the aetiology of GTS is possible given the well-documented connection between synaptic function and neurodevelopmental disorders such as ASD (Parikshak *et al*., 2015; De Rubeis *et al.*, 2014), schizophrenia (Parikshak *et al*., 2015; Kavanagh *et al*., 2014), ADHD (Purper-Ouakil *et al.*, 2011) and GTS (Fernandez *et al.*, 2012). Conversely, module 5 comprises genes with widespread bodily expression and whose expression in the brain is specific to the early/mid prenatal brain and to the cerebellum as compared to genome genes, remaining GTS CNVR co-expressed genes, all CT genes and CT CNVR co-expressed genes (**Figs 3.8, S3.1, S3.2 and S3.3**). Moreover, these genes are more strongly co-expressed with chromatin remodelling and histone-modifying genes than remaining GTS and CT CNVR co-expressed genes ($p < 0.01$) (**Figs 3.8d and S3.4**). This suggests module 5 genes have roles in regulation of gene expression, such that their disruption compromises

neurodevelopment given the importance of the successful coordination of various neuronal proliferation, differentiation and migration events. Altogether, this analysis implicates disruption of neuronal functions and gene regulation via chromatin modification in GTS, further supporting an etiological overlap between GTS and NDDs, such as ASD and ID, where such biological themes are often misregulated.

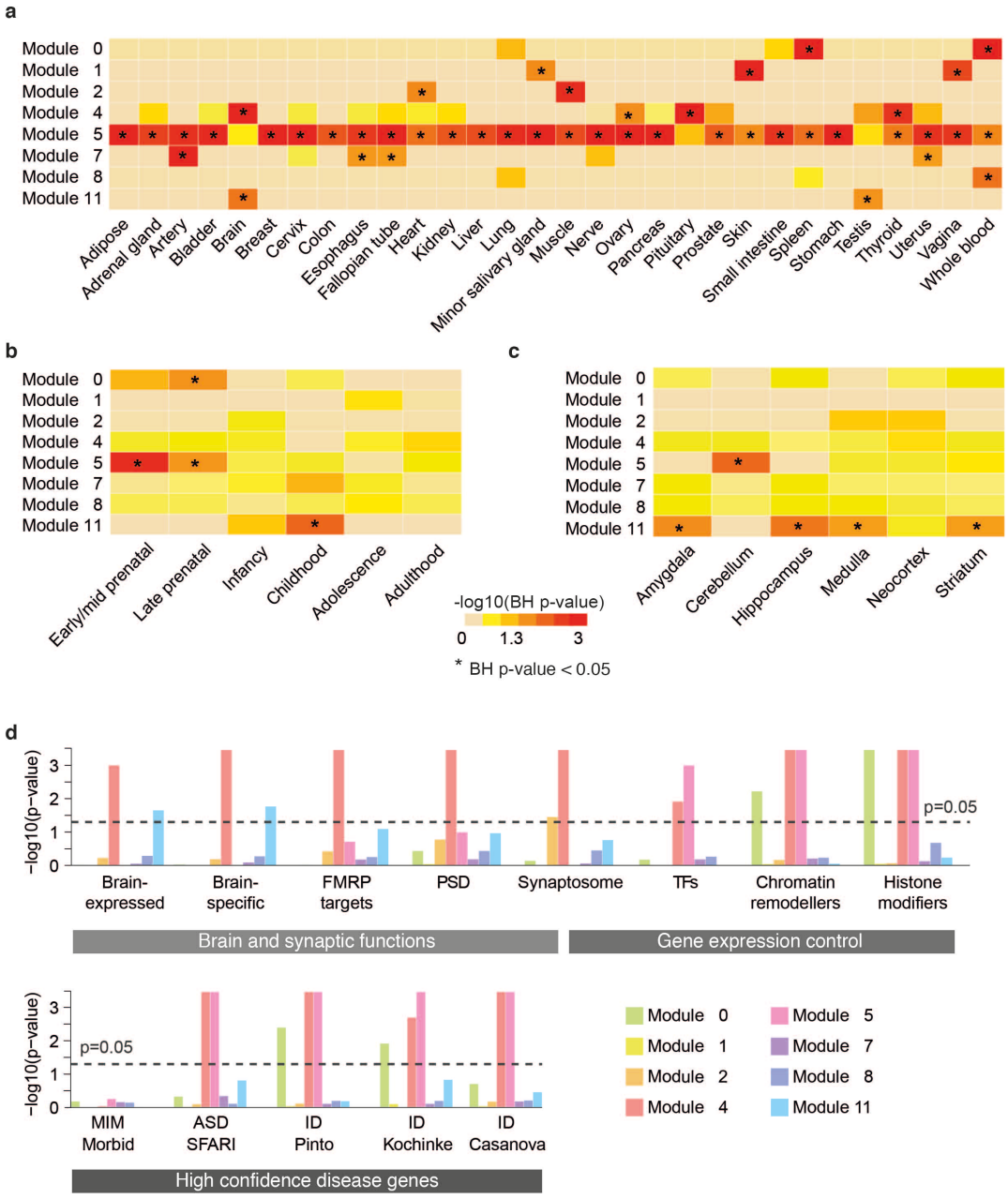


Fig 3.8 | Module characteristics. (**a**, **b**, **c**) Expression patterns of module genes compared to the remaining GTS CNVR (CNV region) co-expressed genes in (**a**) 30 body tissues (GTEx dataset) and in the brain at (**b**) six developmental points and in (**c**) six regions (Brainspan datasets). P-values were derived using right-tailed Mann-Whitney U tests and a Benjamini-Hochberg (BH) correction for multiple testing was applied to control for the numbers of tissues, stages or regions tested. (**d**) Using the GTEx co-expression network with r > 0.6, we evaluated whether module genes were more closely expressed with various gene sets (**Table 2.1**) than permuted gene modules (**section 2.5.3**). PSD, post-synaptic density; TFs, transcription factors; MIM Morbid, OMIM genes; ASD, autism spectrum disorders; ID, intellectual disability.

3.5 Discussion

In this genome-wide CNV analysis of 240 Danish GTS patients and 997 Danish controls, we detected no increased burden of pathogenic neurodevelopmental CNVs or genes and no strong evidence of significant GTS-associated regions. Nevertheless, several genes warrant further attention. Firstly, CNVs overlapping the *DLG2* gene were observed in four GTS cases, reaching genome-wide significance (**Table 3.5**). *DLG2* encodes a scaffolding protein involved in the formation and maintenance of postsynaptic density structures by recruiting components such as the N-Methyl-D-Aspartate (NMDA) receptor subunits. Given the role of the encoded protein in synaptogenesis, an association between *DLG2* and NDDs is therefore not unsurprising and accordingly disruptions in this gene have previously been observed in ASD (Egger *et al.*, 2014) and schizophrenia (Kirov *et al.*, 2012; Fromer *et al.*, 2014). Secondly, the *MYT1L* gene, encoding a protein involved in neuronal lineage specification and previously implicated in ID (MIM:616521), was nominally more frequently disrupted by CNV duplications in patients, suggesting it may be associated with various NDDs. Lastly, we provide supporting evidence for *PANK2* as a potential GTS susceptibility gene as it maps to *dn* CNV deletions observed in two patients from GTS and ADHD studies (Fernandez *et al.*, 2012; Williams *et al.*, 2010). As these genes have not previously been linked to GTS, replication in independent cohorts as well as functional studies are necessary in order to evaluate, validate and clarify these associations.

This study supports the notion of shared genetic risk between GTS and other NDDs by uncovering several neurodevelopmental genes disrupted by patient CNVs (n.b. significance was not reached) (**Table 3.6**). We observed five ID-implicated genes, namely *SYNGAP1* (MIM:612621; present in 1 control as well) encoding a postsynaptic density protein, *TBCE* encoding a protein involved in neuronal morphogenesis, *MYT1L* encoding a protein involved in neuronal differentiation, as well as *PEX11B* and *GALC*. Additionally, we identified three ASD genes from the SFARI Gene database, namely *UBE3C* and *KIAA1586,* encoding ligases involved in ubiquitination and post-translational protein modification, respectively and *TRPM1,* encoding a cation channel mediating synaptic transmission in retinal cells. Finally, two schizophrenia-associated genes were also present – *DISC1* (MIM:181500, MIM604906) and *CHRNA7* (MIM:118511) – although they also occurred in controls.

Given that rare CNVs are present in the general population, it is likely that only a subset of the CNVs found in GTS patients are pathogenic and that many of the genes mapping to multigenic CNVs are not associated with the disease phenotype. Selecting causal CNVs and genes is a difficult task and often achieved using algorithms, which prioritise CNV regions more common in cases than controls (Hehir-Kwa *et al.*, 2010). Additionally some algorithms select causal CNVs based on genes involved in functions misregulated in disease, causing abnormal phenotypes in animal models or whose function is predicted to be compromised based on genomic features (Hehir-Kwa *et al.*, 2010; Doelken *et al.*, 2013; Luo *et al.*, 2014; Foong *et al.*, 2015). Here, based

on the dual assumption that (1) genes involved in the same disorders are functionally related (Oti and Brunner, 2007) and (2) that genes expressed simultaneously in the same tissues are more likely to function together (Lee *et al.*, 2004), we identified a subset of co-expressed GTS CNVR genes, which we deemed more likely to be involved in disease pathogenesis (**Fig 3.5**).

Table 3.6 | OMIM genes observed in GTS patients.

Gene	Disease association	Protein function	# loss / gain	
			GTS	CT
CHRNA7	Schizophrenia (MIM:118511)	Ligand-gated ion channels that mediate fast signal transmission at synapses	0 / 1	0 / 1
DISC1	Schizophrenia (MIM:604906; MIM:181500)	Neurite outgrowth and cortical development	0 / 1	0 / 5
GALC	Krabb disease (MIM:245200)	Lysosomal enzyme involved in the catabolism myelin lipids	1 / 0	0 / 0
KIAA1586	ASD (SFARI Gene)	E3 SUMO-protein ligase involved in protein modification	0 / 1	0 / 0
MYT1L	ID (MIM:616521)	transcription factor associated with neuronal differentiation	0 / 3	0 / 0
PEX11B	Peroxisome biogenesis disorder (MIM:614920)	Involved in peroxisomal proliferation	0 / 1	0 / 0
SYNGAP1	ID (MIM:612621)	Ras GTPase activating protein, member of NMDA receptor complex, involved in post-synaptic signalling	0 / 1	0 / 1
TBCE	Progressive encephalopathy (MIM:617207); Hypoparathyroidism-retardation-dysmorphism syndrome (MIM:241410); Kenny-Caffey syndrome (MIM:244460)	Tubulin-folding protein involved in the maintenance of the neuronal microtubule network	0 / 1	0 / 0
TRPM1	ASD (SFARI Gene)	Cation channel essential expressed in retinal bipolar cells	0 / 1	0 / 1
UBE3C	ASD (SFARI Gene)	E3 ubiquitin-protein ligase	0 / 1	0 / 0

Based on GO enrichment analysis, GTS CNVR co-expressed genes are involved regulation of gene expression. Gene expression control occurs directly by managing the recruitment of RNA polymerases via Transcription Factor (TF) binding. Additionally, it is indirectly modulated via DNA methylation, histone modification and nucleosome remodelling, which alter the spatial organisation of DNA and nucleosomes and hence its accessibility.

Proteins with such regulatory functions have been implicated in NDDs (Ronan *et al.*, 2013), such as ASD (Bourgeron, 2015), ID (Vissers *et al.*, 2016) and schizophrenia (McCarthy *et al.*, 2014). Correspondingly, we found that an excess of GTS CNVR co-expressed genes encode TFs compared to CT genes (**Fig 3.6a**). Of the 33 genes encoding TFs identified in cases, two are foetally-expressed and encode proteins specifying neuronal lineages with possible links to disease. These are the ID-implicated *MYT1L* gene (MIM:616521) and the *TCF3* gene encoding a TF, which binds and regulates the expression of the *AUTS2* gene, linked to autism and other NDDs (Oksenberg *et al.*, 2014). Furthermore, although no disease-implicated DNA modifiers were found in this cohort, GTS CNVR co-expressed genes were more closely expressed with chromatin remodelling and histone modifying genes than CT genes (**Fig 3.6b**), suggesting these functions were more likely to be affected in cases based on the *guilt by association* principle.

The study of GTS CNVR co-expressed genes supports the notion of shared risk with ASD and ID as this set included most of the disease-implicated genes observed in this cohort (disease-associated genes are more common among GTS CNVR co-expressed compared to non-co-expressed genes with ORs > 1.2) (**Fig 3.6a**). Given the low incidence of neuropsychiatric genes in this cohort, we reinforced this view by showing GTS CNVR co-expressed genes are more closely expressed with disease risk genes than CT genes and GTS CNVR non-co-expressed genes (**Fig 3.6b**).

Further inspection of GTS CNVR co-expressed genes revealed they are predicted to be haploinsufficient and expressed in the brain and particularly so during early/mid prenatal development (**Fig 3.7; Table S3.4**). Furthermore, in line with the current views of neuroanatomical localisation of GTS and the implication of striato-cortical circuits (Stern *et al.*, 2000; Peterson *et al.*, 2003; Saka and Graybiel, 2003; Albin and Mink, 2006; Baym *et al.*, 2008; Paschou *et al.*, 2014), we found that GTS CNVR co-expressed genes are most strongly expressed in the striatum. Finally, we recognised similar patterns when comparing these genes to a set of neutral co-expressed genes identified using control CNVRs.

Interestingly during case-control comparisons, we did not find an over-representation of genes with neuronal functions, such as synaptic transmission and plasticity. Nevertheless, in-depth analysis of GTS CNVR co-expressed genes revealed they include a subset (i.e. module 4), which is strongly co-expressed with genes targeted by FMRP and localised to postsynaptic densities and synaptosomes (**Fig 3.8**). This suggests that, similarly to other NDDs, such as ASD (Vortsman *et al.*, 2017), both synaptic functions and processes of gene expression control are affected in GTS. We speculate that the lack of evident neuronal signal results from the large number of genes mapping to controls CNVs, which are brain-expressed and associated to synaptic functions (GO enrichment analysis; not shown). As CNVs from the general population have been shown to target particular gene types more frequently, including those involved in nervous system development and functions (de Smith *et al.*, 2007; Raychaudhuri *et al.*, 2010),

such results are not unexpected. In fact, they underscore the importance of characterising genes with regard to their potential implication in disease, as brain-related developmental and physiological functions alone are not sufficient in predicting pathogenicity.

Finally, we would like to note that this study presents several limitations. Firstly, the size of the patient cohort is relatively small and the resolution of CNV detection was compromised by the reliance on differing genotyping arrays between cases and controls. Despite this, we were able to control for platform differences and construct comparable case and control CNV sets. Secondly, although our study evaluates rare CNVs, which are more likely to be causal compared to common variants (Manolio *et al.*, 2009), we do not have information as to the inheritance of such CNVs and are unable to identify those arising *de novo*. This is a major limitation as *dn* CNVs are highly penetrant and of particular interest to neuropsychiatric and developmental conditions (Sanders *et al.*, 2011; Sebat *et al.*, 2007; Marshall *et al.*, 2008; Stefansson *et al.*, 2008; Malhotra *et al.*, 2011; Sanders *et al.*, 2015; Koolen *et al.*, 2009). Despite these caveats, this study demonstrates the utility of undertaking CNV analyses in GTS. We also underscore the need for additional larger studies with the aim of validating current findings as well as clarifying and extending our understanding of the pathophysiology behind the onset of GTS.

3.6 Conclusion

In this study, we performed a genome-wide CNV study of a Danish cohort of GTS patients. Focusing on rare CNVs, we identified a subset of genes displaying similar expression patterns, more likely to be involved GTS pathogenesis. These co-expressed GTS genes are brain-expressed, particularly in the developing brain, predicted to be haploinsufficient and involved in transcriptional control. They, therefore, implicate the disruption of neurodevelopmental processes via misregulation of gene expression in the aetiology of GTS. We, additionally, reinforce the perception of shared genetic risk among NDDs, as GTS CNVR co-expressed genes are related to known and observed genes in ASD, ID and ADHD. In particular, the *DLG2* gene, which encodes a postsynaptic density protein and has previously observed in ASD and schizophrenia, is more frequently observed in GTS patients from our cohort compared to controls. Similarly, the *MYT1L* gene is nominally more commonly disrupted by CNV duplications in patients and has previously been implicated in ID, thus suggesting a possible comorbidity. Additionally, we propose the *PANK2* gene as a possible GTS susceptibility gene, based on its disruption by *dn* CNV deletions in two unrelated patients, presenting with GTS and ADHD.

CHAPTER 4

Genotype and phenotype convergence:

a CNV study

4.1 Introduction

Neurodevelopmental disorders are a group of heterogeneous conditions, resulting from mental impairments and/or physical abnormalities, often involving a life-long need of supportive services. The most prevalent NeuroDevelopmental Disorders (NDDs), which affect up to 3% of births, include Developmental Delay (DD) (i.e. 1-3% of children) (Shevell *et al.*, 2003), Intellectual Disability (ID) (i.e. 0.5-1.5% of the population) (Maulik *et al.*, 2011; McKenzie *et al.*, 2016) and Autism Spectrum Disorders (ASD) (i.e. 1.46% of the population) (Christensen *et al.*, 2016). Specifically, DD involves global delays in development while ID is associated with substantial impairments in cognitive functioning (Intellectual Quotient, IQ < 70) and adaptive behaviours, and ASD is characterised by deficits in social interaction and communication and restricted and repetitive behaviours. Structural anomalies (i.e. Congenital Abnormalities; CAs) are also common and affect between 0.15% and 2.4% of births (Dolk *et al.*, 2010; Calzolar *et al.*, 2014). The qualitative impairments of the diagnostic features of NDDs and CAs are highly variable and often accompanied by additional comorbidities, such that ID occurs in 30% to 61% of ASD patients, for example (Van Naarden Braun *et al.*, 2015; Christensen *et al.*, 2016).

Environmental causes, such as parental age, maternal smoking status, alcohol use and infections, have been proposed for NDDs but twin and family studies frequently implicate an underlying genetic aetiology. While Genome Wide Association Studies (GWAS) have uncovered few common disease

variants (i.e. present in > 1% of the population), epidemiological studies support a causal role for rare variants (i.e. present in < 1% of the population). In fact, rare and in particular *de novo* (*dn*) Copy Number Variants (CNVs; deletions or duplications of DNA segments > 1 kb) and Single Nucleotide Variants (SNVs; DNA base substitutions) have been associated with many diseases, including ID and ASD (Merner *et al.*, 2015).

The development of high throughput genetic screens, such as chromosomal microarray, enabled CNV screens to become routinely performed. However, despite their relatively high diagnostic yield (i.e. 20% in ID, CAs or ASD; Miller *et al.*, 2010a), it often remains difficult to pinpoint the causal gene(s). More recently, the advent of next generation sequencing facilitated an examination of human genetic coding variation using whole-exome screens. Nevertheless, screening of large cohorts has confidently and specifically associated a relatively small set of genes with neurodevelopmental conditions. Indeed, reflecting the high phenotypic variability observed in such disorders, there is an equally compelling genetic heterogeneity attributable to a combination of (i) small effect sizes of individual mutations, (ii) gene additive, epistatic and other non-linear effects as well as (iii) gene pleiotropy and incomplete penetrance (see **chapter 1**).

Over the past decade, large -omics projects have helped build maps of human genetic variation in disease. This, in turn, led to the development of comprehensive genomics approaches, such as pathway analyses, which enable the simultaneous examination of hundreds of genetic variants. Under

the assumption that variants from individuals with the same disorder affect genes participating in a shared biological pathway, these approaches seek to uncover molecular processes recurrently disrupted in disease. Such methods have been successful in characterising the biological context of various psychiatric disorders. In ASD, for example, studies have implicated genes (i) with converging transcriptomic brain signatures, (ii) whose protein products physically interact and (iii) carry out functions of chromatin remodelling and synaptic plasticity (Ben-David and Shifman, 2013; Noh *et al.*, 2013; Steinberg and Webber, 2013; Cristino *et al.*, 2014; De Rubeis *et al.*, 2014; Iossifov *et al.*, 2014; Krumm *et al.*, 2014; Pinto *et al.*, 2014; Toma *et al.*, 2014; Chang *et al.*, 2015; Hormozdiari *et al.*, 2015; Mahfouz *et al.*, 2015; Sanders *et al.*, 2015; Krishnan *et al.*, 2016; Li *et al.*, 2016; Shohat *et al.*, 2017; Stessman *et al.,* 2017; Yuen *et al.*, 2017). Similar biological themes have been uncovered in other NDDs, such as ID, pointing to a shared pathophysiology (Guilmatre *et al.*, 2009; Kaufman *et al.*, 2010; Melom and Littleton, 2011; Srour and Shevell, 2014).

Statistical power is central to the identification of disease pathophysiologies. The size and phenotypic resolution of the examined patient cohort limit the ability to detect positive and significant signals, suggesting phenotypically heterogeneous cohorts are often statistically underpowered. This limitation is of particular importance to NDDs, given the high variability of core impairments and prevalence of comorbidities. For example, ID and ASD tend to co-occur with each other as well as with other autoimmune and psychiatric disorders, including Attention Deficit and Hyperactivity Disorder (ADHD),

Obsessive Compulsive Disorder (OCD), epilepsy, Gilles de la Tourette syndrome and anxiety. Furthermore, they are often accompanied by additional medical conditions, such as CAs, gastro-intestinal problems, sleep disturbances, motor dysfunction and neurological problems (e.g. cerebral palsy and seizures) (Simonoff *et al.*, 2008; Kohane *et al.*, 2012; Matson and Cervantes, 2013; Cooper *et al.*, 2015). Consequently, the study of phenotypically more homogeneous patient sub-groups should gain considerable power for detecting aetiologies underlying the corresponding abnormalities.

Analysis of phenotypically homogenous cohorts has led to the discovery of disease-associated common loci in ischaemic stroke (Traylor *et al.*, 2012) and migraines (Anttila *et al.*, 2013) but not ASD (Chaste *et al.*, 2015), possibly owing to the modest effect of common compared to rare variants. Beyond the study of common variation, an increased burden of rare and large CNVs was observed in DD patients presenting with craniofacial or cardiovascular defects compared to autism (Cooper *et al.*, 2011). In ID, CNV inheritance status was reported to be independent from symptomatic presentations (Qiao *et al.*, 2014). However, ID genes associated with particular phenotypes were shown to affect similar molecular processes, hence supporting the notion of phenotype and genotype convergence (Kochinke *et al.*, 2016).

Complementary to such top-down, phenotype to genotype analyses is the bottom-up approach aiming to uncover disorders and/or traits prevalent in genetically homogeneous patient groups, in which the same molecular

processes are disrupted. Accordingly, phenotypic coherence was observed among DD and ID patients, whose CNVs overlap (1) genomically-derived functional gene clusters (Andrews *et al.*, 2015a) and (2) genes with similar functional annotations or brain expression patterns (Andrews *et al.*, 2015b), respectively. Conversely, a recent study showed that while different types of CNVs (i.e. inherited, *dn*, deletions, duplications or affecting brain-expressed genes) may modulate ASD severity or sub-phenotypes, these relationships are not obvious (Merikangas *et al.*, 2015).

The lack of clear findings in the study by Merikangas *et al.*, 2015 could be a consequence of the lack of specificity of the evaluated patient groups. Indeed, comparing the phenotypic effects between CNVs of opposing directions (i.e. deletions and duplications), inheritance (i.e. *dn* and inherited) and expression patterns (i.e. brain-expressed or not) may be an inappropriate strategy to distinguish between NDD symptoms. Conversely, the study of single CNVs is statistically underpowered, given their low incidence and high phenotypic variability. Instead, in this study, we aimed to pool patients based on the similarity of their molecular disturbances.

Shared functional characteristics can be inferred for (1) genes with similar expression patterns or (2) genes whose protein products physically interact. Additionally, (3) relatedness can be deduced for genes with similar annotations in datasets such as Gene Ontology (GO) (Gene Ontology Consortium, 2000) and the Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000; Kanehisa *et al.*, 2017). Similarly, (4)

genes whose orthologues produce similar phenotypes when disrupted in animal models are more likely to converge at the functional level. Finally, (5) by combining the aforementioned types of annotations, gene pairs can be assigned functional similarity scores (Honti *et al*., 2014).

In addition to exome- or genome-wide genetic data, the investigation of genotype and phenotype relationships requires deeply and systematically phenotyped cohorts to enable comprehensive and reliable comparisons of clinical abnormalities. The Simons Foundation Autism Research Initiative (SFARI), for example, recruited over 2,600 autism simplex families (Simons Simplex Collection; SSC), which have been evaluated for genetic, clinical and neurobiological data (Fischbach and Lord, 2010). Patients were deeply and rigorously phenotyped to provide information on a variety of measures relating to core ASD symptoms, cognitive abilities, life milestones, family description, patient and family medical history, perinatal information and patient immunisations, diet, etc. Importantly, as phenotyping was systematic and missing data indicated, phenotypic comparisons can be confidently performed between all patients. Furthermore, the need for automated phenotypic comparisons led to the development of the Human Phenotype Ontology (HPO), which provides a global and structured set of phenotypic abnormalities observed in human disease (i.e. over 10,000 phenotypic terms) and the semantic relationships between them (Robinson *et al.*, 2008; Köhler *et al.*, 2017).

4.2 Chapter aims

The fundamental distinction between *genotype* and *phenotype* formulated by Johannsen in 1911 (Johannsen, 1911) has motivated research aiming to elucidate their influence over each other. Here, we sought to investigate this relationship in context of disease. We examined the phenotypes of patients harbouring *de novo* and inherited CNVs. For this, we used (i) a cohort of individuals with ID, DD or CAs recruited by the Department of Human Genetics at the Radboud University, Nijmegen, Netherlands (herein termed *Nijmegen*) (Vulto-van Silfhout *et al*. 2013a; Andrews *et al.*, 2015b) and (ii) the SSC cohort of individuals with ASD (Fischbach and Lord, 2010). Specifically, we (1) grouped genes affected in these cohorts based on their expression patterns and functional annotations and assigned patients carrying variants in these genes to each gene group. We, then, (2) evaluated the similarity of phenotypes displayed by patients from the same gene groups and (3) investigated the driving factors behind the observed phenotypic coherence. An overview of workflow of our study is displayed in **Fig 4.1**.

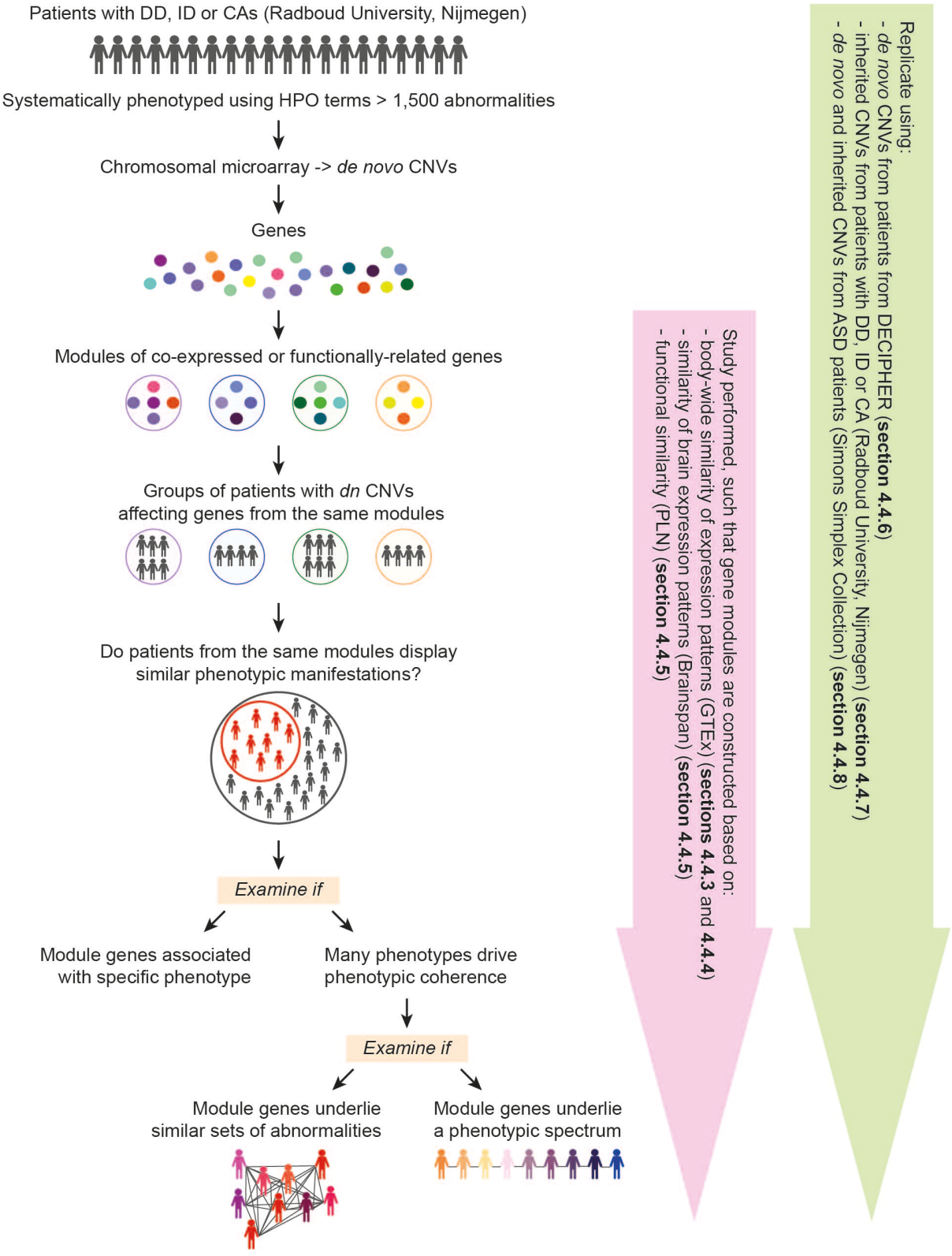


Fig 4.1 | Workflow to examine genotype and phenotype relationships in NDDs. DD, developmental delay; ID, intellectual disability; CAs, congenital abnormalities; HPO, Human Phenotype Ontology; ASD, autism spectrum disorders.

4.3 Specific materials and methods

4.3.1 Nijmegen cohort: patients with ID, DD and/or CAs

The cohort consists of 4,297 patients diagnosed with ID, DD and/or CAs and recruited by the Department of Human Genetics at the Radboud University, Nijmegen (herein termed *Nijmegen*) (Vulto-van Silfhout *et al*. 2013a; Andrews *et al.*, 2015b). Individuals were deeply and consistently phenotyped using terms from the HPO, such that lack of annotations illustrates absence of (rather than unrecorded) phenotype. Overall, the cohort is annotated with 1,350 distinct HPO terms, which were remapped to fit the newest HPO release. HPO terms are organised in a hierarchy, where daughter terms (located at the bottom of the HPO ontology tree) describe specific clinical abnormalities and are assigned to broader parent terms (daughter terms detain an *is-a* relationship with parent terms). Therefore, by parsing through the HPO ontology, we annotated each individual with his or her assigned abnormalities as well as their corresponding more general parent terms. Ultimately, we observed a total of 1,881 distinct (daughter and parent) HPO phenotypes within the 4,297 patients from the cohort (n.b. less abnormalities are observed within the set of patients used in this study; see below).

Patient DNA samples, taken via peripheral blood, were analysed using the Affymetrix 250k Nspl Single Nucleotide Polymorphisms (SNP) array platform (262,264 SNPs, 200 kb resolution). CNVs were called where five consecutive aberrant SNP losses or seven gains were detected (Vulto-van-Silfhout *et al.*,

2013a; Andrews *et al.*, 2015b). In total, 1,061 rare CNVs were called for the subset of patients for who parental DNA was available. 426 CNVs were confirmed *de novo* (*dn*) and 635 are inherited (inh). We excluded CNVs > 5 mb, for which several genes are likely to be causal, as well as patients with missing phenotypic annotations. The final dataset consists of 222 and 481 patients carrying 245 *dn* and 531 inh CNVs, which overlap 3,140 and 2,016 distinct genes, respectively. Analyses were performed separately for patients carrying *dn* and inh CNVs.

4.3.2 DECIPHER cohort: patients with chromosomal aberrations

We downloaded clinical and genetic (i.e. CNV) data for individuals from the DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources (DECIPHER) (Bragin *et al.*, 2014; Firth *et al.*, 2009). In total 27,261 CNVs were identified in 19,845 patients, annotated with 2,714 distinct HPO terms (3,234 terms following imputation of more general parent phenotypes). Among all identified CNVs, 3,578 are *dn* CNVs spanning < 5 mb and occur in 3,245 individuals. The DECIPHER cohort, used as a replication cohort in this study, refers to the subset of 1,604 patients carrying 1,682 *dn* CNVs < 5 mb, which overlap 2,602 genes, also observed among Nijmegen patients with *dn* CNVs.

4.3.3 SSC cohort of ASD patients: phenotypic data

We downloaded phenotypic data for 2,644 simplex ASD families (i.e. with one affected child), recruited by the SSC, from SFARI Base. Briefly, probands were assessed using multiple instruments to evaluate diagnostic and core autism features, adaptive behaviour, motor skills and cognitive abilities. Detailed clinical patient descriptions are also available and include information regarding comorbid features (e.g. neuropsychiatric disorders, sleep disturbances and gastrointestinal problems), life milestones (e.g. age at first words), perinatal information (e.g. gestational age and perinatal complications), physical abnormalities (e.g. dysmorphism), medications, immunisations, allergies and many others. A detailed family history of disease is also available.

As this study is exploratory with no prior knowledge regarding the relevance of various phenotypes, we constructed a comprehensive binary phenotypic dataset, detailing the occurrence and lack of various clinical abnormalities. Aiming to cover the full autistic spectrum (**Fig S4.1**), we incorporated information regarding a large number (n = 63) of non-core ASD comorbidities, clinical problems and environmental and familial predispositions (**Table 4.1**). These include food and environment allergies, ten autoimmune disorders, five types of gastro-intestinal problems, eleven autoimmune disorders present in the mother, five neurologic problems, nine psychiatric disorders, seven types of physical or congenital defects, the presence of different types of regression, whether the individual is low functioning and a few others. Of the 2,644

assessed patients, 136 lack genetic data and another ten individuals were excluded due to high rate of missing data, leaving 2,498 for analysis (herein the SSC cohort refers to the subset of 2,498 analysed individuals) (**section S3**). Missing data was imputed using the multivariate imputation by chained equations approach, implemented in R (van Buuren and Groothuis-Oudshoorn, 2011). The incidence of phenotypic variables pre- and post-imputation is displayed in **Table 4.1**. Detailed descriptions of the phenotypes evaluated and included in the dataset are available in **section S3**.

Table 4.1 | Prevalence of phenotypes in the SSC cohort.

Phenotypic group	Phenotype	Incidence in % (number of patient with/ number of patients assessed)	
		Before imputation	After imputation
Abnormal gestational age	Pre-term (< 37 weeks)	8.48 (169/1994)	9.17 (229/2498)
	Post-term (> 41 weeks)	5.27 (105/1994)	5.93 (148/2498)
Allergies	Food (any)	15.90 (365/2296)	16.53 (413/2498)
	Environment (any)	27.87 (615/2207)	24.46 (686/2498)
Autoimmune disorders	Diabetes mellitus I & II	0.080 (2/2286)	0.080 (2/2498)
	Systemic lupus erythematosus	0.087 (2/2297)	0.48 (12/2498)
	Hypothyroidism	0.53 (12/2278)	7.0 (175/2498)
	Hashimotos tyroiditis	0.087 (2/2300)	0.087 (2/2498)
	Rheumatoid arthritis	0.088 (2/2274)	0.088 (2/2498)
	Multiple sclerosis	0.043 (1/2299)	0.043 (1/2498)
	Bowel disorder	2.0 (654/2435)	7.89 (197/2498)
	Celiac disease	0.48 (11/2303)	6.73 (168/2498)
	Psoriasis	1.04 (24/2298)	7.25 (181/2498)
	Asthma	10.1 (237/2276)	15.29 (382/2498)
Birth defects	Cleft lip palate	0.35 (8/2298)	0.76 (19/2498)
	Congenital heart defect	0.88 (20/2276)	1.96 (49/2498)
	Dysmorphism	14.48 (73/585)	15.93 (398/2498)
	Kidney defect	0.39 (9/2300)	0.76 (19/2498)
	Microcephaly	2.76 (15/544)	4.32 (108/2498)
	Open spine	0.043 (1/2301)	0.043 (1/2498)
	Other	1.77 (40/2256)	2.93 (76/2498)
	Polydactyly	1.04 (24/2301)	1.40 (35/2498)
Epilepsy or seizures	Epilepsy	2.06 (50/2423)	2.32 (52/2498)
	Seizures	5.31 (122/2296)	11.13 (278/2498)
Gastro-intestinal problems	Constipation	26.86 (654/2435)	27.30 (682/2498)
	Severe abdominal pain	5.36 (130/2426)	5.68 (142/2498)
	Diarrhea	11.75 (287/2443)	12.01 (300/2498)
	Vomiting	5.57 (136/2443)	5.68 (142/2498)
	Gastroesophageal reflux disease	10.42 (254/2438)	10.65 (266/2498)
IQ < 60	Non-verbal IQ < 60	12.07 (277/2295)	13.01 (325/2498)
Language delay	First words delay (> 36 weeks)	25.07 (614/2449)	25.10 (627/2498)
	First phrases delay (> 42 weeks)	47.4 (1160/2446)	47.4 (1184/2498)
Mother autoimmune disorders	Diabetes mellitus I & II	1.76 (40/2268)	4.08 (102/2498)
	Systemic lupus erythematosus	0.65 (15/2297)	1.88 (47/2498)
	Hyperthyroidism	1.74 (40/2294)	3.84 (96/2498)
	Hashimotos tyroiditis	1.74 (40/2303)	3.92 (98/2498)
	Rheumatoid arthritis	1.80 (41/2280)	4.92 (123/2498)
	Multiple sclerosis	0.22 (5/2299)	0.22 (5/2299)
	Asthma	10.99 (250/2275)	16.65 (416/2498)
	Bowel disorder	3.22 (74/2298)	9.37 (234/2498)
	Celiac disease	0.52 (12/2303)	6.93 (173/2498)
	Psoriasis	2.69 (62/2303)	8.85 (221/2498)
	Adrenal insufficiency	0.17 (4/2302)	0.17 (4/2498)
Neurologic problems	Uncoordination/ clumsiness	20.20 (489/2421)	20.46 (511/2498)
	Movement abnormalities	3.28 (80/2436)	3.20 (80/2498)
	Cerebral palsy	0.29 (7/2441)	0.44 (11/2498)
Psychiatric disorders	ADHD	14.30 (329/2300)	18.93 (473/2498)
	Schizophrenia	0	NA
	Bipolar disorder	0.52 (12/2289)	0.52 (12/2289)
	Tourette's	1.35 (31/2297)	7.81 (195/2498)
	Chronic tics	4.88 (119/2438)	4.96 (124/2498)
	OCD	3.49 (80/2293)	9.49 (237/2498)
	Anxiety disorder	3.40 (91/2283)	10.17 (254/2498)
	Behaviour disorder	2.12 (49/2308)	8.09 (202/2498)
	Depressive disorder	0.88 (20/2266)	7.25 (181/2498)
Regression	Language (loss of ≥ 3 words)	17.82 (442/2480)	17.85 (446/2498)
	Social: Constructive play	6.41 (160/2495)	6.41 (160/2498)
	Social: Social engagement	23.37 (583/2495)	23.38 (584/2498)
	Motor: Hand movement	3.45 (86/2495)	3.44 (86/2498)
	Motor: Motor skills	4.37 (109/2495)	4.36 (109/2498)
	Motor: Self help	4.0 (100/2495)	4.0 (100/2498)
Severe ASD	ADOS CSS score of 10/10	14.51 (351/2419)	14.51 (351/2419)
Other	Developmental coordination	80.2 (1924/2399)	78.7 (1966/2498)
	Sub-optimal pregnancy	32.82 (820/2498)	32.82 (820/2498)

(Continues on the right)

The incidence (in number and percentage of individuals) of phenotypes included in the analysis pre- and post- imputation of missing values using multivariate imputations by chained equations.

4.3.4 SSC cohort of ASD patients: CNV data

All 14,823 rare CNVs (i.e. with a frequency of < 1% in the Database of Genomic Variants; MacDonald *et al.*, 2014) identified in SSC patients that passed quality control were downloaded from Sanders *et al.*, 2015. We restricted the analysis to the 2,498 patients used in this study and to CNVs < 5 mb. This left 130 and 2,202 patients displaying 138 *dn* and 10,799 inh (n.b. 2,829 / 10,799 are genic CNVs) CNVs, which affect 988 and 3,813 distinct genes, respectively.

4.3.5 Construction of patient modules

Aiming to partition the CNV-affected genes from our study cohorts by their expression dynamics or functional similarities, we exploited the previously constructed GTEx and Brainspan gene co-expression networks and the Phenotypic Linkage Network (PLN) (**sections 2.4 and S2**). As not all links from these networks are relevant (e.g. a co-expression coefficient of r = 0.4 is likely to be uninformative), we sought to focus on only the most informative links.

We considered the n = [100, 90, 80 .. 10, 1] % strongest PLN links and co-expression coefficients r > [0.4, 0.5 .. 0.9] and evaluated the modularity (i.e. strength of network division; high modularity reaches towards 1 and suggests there is an excess of intra- compared to inter-module links) of sub-networks comprising genes mapping to *dn* or inh CNVs from Nijmegen or SSC patients.

The n and r thresholds were chosen to optimise both modularity and the number of genes in the network, such that GTEx networks with r > 0.6, Brainspan networks with r > 0.8 and the top 25% PLN links were considered for further analysis (**Fig 4.2**).

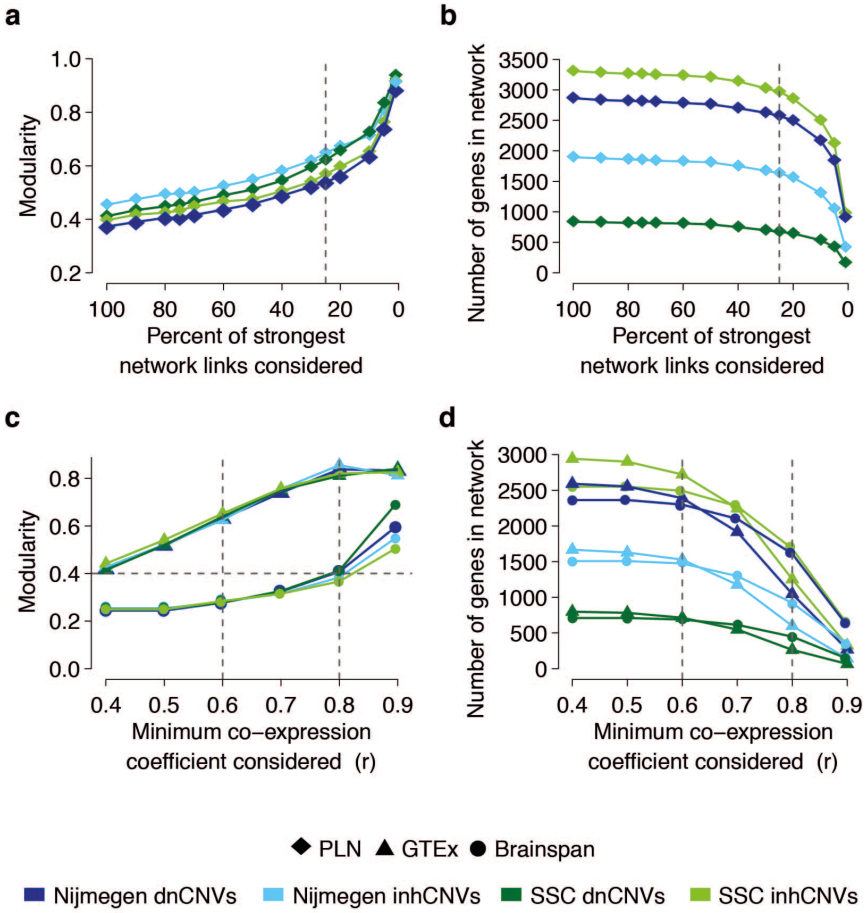


Fig 4.2 | Choice of gene network thresholds. We considered GTEx (triangle) and Brainspan (circle) co-expression sub-networks and the PLN (Phenotypic Linkage Network; diamond) sub-network consisting of genes mapping to CNVs from Nijmegen (blue) and SSC (green) patients. We restricted PLN sub-networks to the top k = [100, 90, 80 .. 10, 1] % strongest links and co-expression sub-networks to links with r > [0.4, 0.5 .. 0.9]. We examined (**b, d**) the number of genes present in each network and (**a, c**) the modularity associated with their partitioning using the Louvain algorithm. An optimal network division generates a high modularity while maximising the number of genes present in the network. We thus considered (**a, b**) the top 25% strongest PLN links and (**c, d**) GTEx and Brainspan co-expression networks with r > 0.6 and r > 0.8, respectively, for further analysis.

Sub-networks consisting of genes overlapped by *dn* or inh CNVs from Nijmegen or SSC patients were partitioned using the Louvain algorithm (**section 2.5.4**) (Blondel *et al.*, 2008). We, then, formed the corresponding patient groups by assigning patients to all modules they contribute to with at least one gene (individuals can have multiple module memberships). We restricted the analysis to modules comprising ≥ 10 patients.

Of note, the Louvain algorithm consistently uncovered the same gene modules when considering all n and r values. In fact, gene modules, identified using various thresholds, overlap at approximately 1:1 ratios (one-sided right-tailed Fisher tests with 5% false discovery rates; not shown). This confirms modules used in the main analysis are not confounded by the number of top PLN links (n) or minimum co-expression coefficients (r) considered.

4.3.6 Global genotype and phenotype convergence

To examine whether individuals with similar functional disruptions are globally more phenotypically similar (also termed phenotypically coherent) than expected by chance, we examined the distribution of pairwise patient phenotypic similarity scores (**section 2.6**). We compared similarity scores between patients with shared module memberships and those between individuals, which never occur within the same module. Significance was derived using one-sided right-tailed Mann-Whitney U tests.

4.3.7 Specificity of phenotypic coherence

To examine the specificity of phenotypic coherence observed for DECIPHER patients assigned to the same modules, we compared their cumulative phenotypic similarity score to that of 1,000 randomised sets of patients. Randomised sets were formed by permuting 1/3 of individuals between the evaluated module and the background (i.e. DECIPHER cohort). An empirical p-value was derived, reflecting the number of randomisations that achieve a higher cumulative phenotypic similarity as compared to the set of interest.

4.3.8 Identifying phenotypes driving phenotypic coherence

Considering all phenotypes present in the cohort examined, we assessed the excess or depletion of each abnormality among patients from the same modules compared to remaining patients from the cohort, using two-sided Fisher tests. We, then, examined the phenotypic coherence of module patients (**section 2.7.3**), following the sequential and combinatorial removal of phenotypes with increasing nominal p-values (i.e. we examined the effect on phenotypic similarity after removing the most strongly associated phenotypes to that module). Specifically, for the Nijmegen cohort, we sequentially removed the top 1 to the top 500 phenotypes most strongly associated with each module. Phenotypic similarity of pairwise patient combinations across the subset of remaining phenotypes was iteratively re-calculated (**section 2.6**) and phenotypic coherence was then evaluated (**section 2.7.3**). For the SSC cohort, phenotypic similarity was examined following the removal of the top 1

to the top 60 phenotypes with smallest p-values, as only 63 phenotypes were examined.

We selected the maximum number of top ranked abnormalities (with smallest p-values) that can be removed before the phenotypic similarity signal is lost (i.e. $p > 0.05$). The corresponding phenotypes were deemed to underlie phenotypic coherence and termed *essential phenotypes* (or *essential terms*).

For each phenotypically coherent patient module, we formed 100 randomised sets of phenotypes, matched for the number of essential phenotypes. We then re-evaluated the phenotypic coherence of module patients when excluding randomised essential terms from the analysis. We derived an empirical p-value by estimating the proportion of randomisations during which module patients are phenotypically similar across all remaining phenotypes (i.e. excluding randomised essential terms) (**section 2.7.3**). This empirical p-value reflects the likelihood of essential phenotypes to combinatorially underlie the observed phenotypic coherence.

4.3.9 Investigating phenotypic relationships between individuals

One property of acyclic graph trees – but not of strongly clustered networks – is that their integrity is compromised following removal of growing numbers of nodes (**Fig 4.3**). To examine whether patients from phenotypically coherent modules are organised in tree-like topologies, we evaluated how the similarity

signal evolves if increasing numbers of patients are removed from the module.

Based on the assumption that individuals from different groups are not phenotypically similar, we swapped increasing numbers of patients between the evaluated modules and the background. Patients were swapped instead of being removed to ensure results are not confounded by module size. We successively swapped k = [1, 2, 3 .. (group sample size - 5)] and k = [1, 10, 20 .. (group sample size - 10)] patients between modules of Nijmegen and SSC patients and the background, respectively. Given the large number of patients belonging to the SSC cohort (2,498 patients vs < 500 in the Nijmegen cohorts), the number of swapped patients was incrementally increased by 10 to optimise computational power. During swaps, we iteratively examined the phenotypic similarity of patients from the new modules (**section 2.7.3**). We note that for each k, we performed n = 100 swap permutations. We, then, assessed the relationship between the significance of the phenotypic similarity of new modules and the number of swapped individuals using linear regression in R. Significant models with negative slopes suggest patients are organised in tree-like structures, as increasing the number of randomly swapped individuals begins to breakdown the phenotypic similarity network (**Fig 4.3**).

As a negative control, we applied this pipeline to phenotypically discordant patient groups (modules whose phenotypic coherence is associated with p > 0.05), to ensure that no relationship is observed (linear regression with p >

0.05) or the relationship is positive (linear regression with $p < 0.05$ and slope > 0).

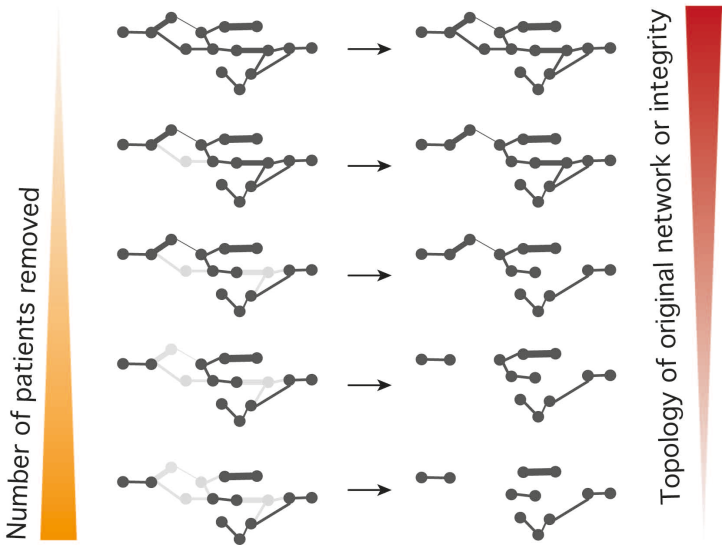


Fig 4.3 | Integrity of acyclic-like path trees. We consider a network where edges depict relationships between nodes, such that absence of edge represents a negligible relationship. (**left panel**) As increasing numbers of nodes are excluded from the network (light grey), (**right panel**) its integrity of is compromised and the initial network topology is lost.

4.4 Results

4.4.1 Summary of data

The Nijmegen cohort consists of 4,297 patients diagnosed with DD, ID and/or CAs. These individuals were systematically phenotyped by clinicians using phenotypic terms from the HPO database, such that absence of specific terms translates into absence of clinical abnormality. In addition to their annotated terms, patients were assigned all corresponding parent terms from the HPO ontology to enable direct phenotypic comparisons.

CNV analysis of this cohort confirmed 426 *de novo* (*dn*) and 635 inherited (inh) CNVs. Due to the difficulty of ascribing specific functionalities to large and gene-dense CNVs, we restricted the analysis to CNVs < 5 mb. The cohort was split based on CNV inheritance status such that the 222 and 481 patients carrying 245 *dn* and 531 inh CNVs, respectively, were analysed separately (**Fig 4.4a-b**).

Dn and inh CNVs span a median of 1.40 mb and 400 kb, respectively. They affect a total of 3,598 and 2,405 distinct genes, with up to 167 and 80 genes per CNV, respectively (**Fig 4.4c-d**). Among *dn* CNVs, there are 147 deletions (i.e. losses) and 98 duplications (i.e. gains), similar in size and gene density ($p = 0.86$ and $p = 0.74$, respectively; two-sided Mann-Whitney U tests). Inh CNVs comprise 195 losses and 336 gains, revealing a depletion of deletions compared to *dn* CNVs ($p < 1 \times 10\text{-}15$, Odds Ratio, $OR = 0.19$; two-sided

Fisher test), possibly owing the higher likelihood of pathogenicity of losses, reducing their probability of transmission (Lee *et al.*, 2007). Inh gains are longer than losses and overlap more genes (p = 3.77 x 10-4 and p = 8.13 x 10-12, respectively; right-tailed Mann-Whitney U tests).

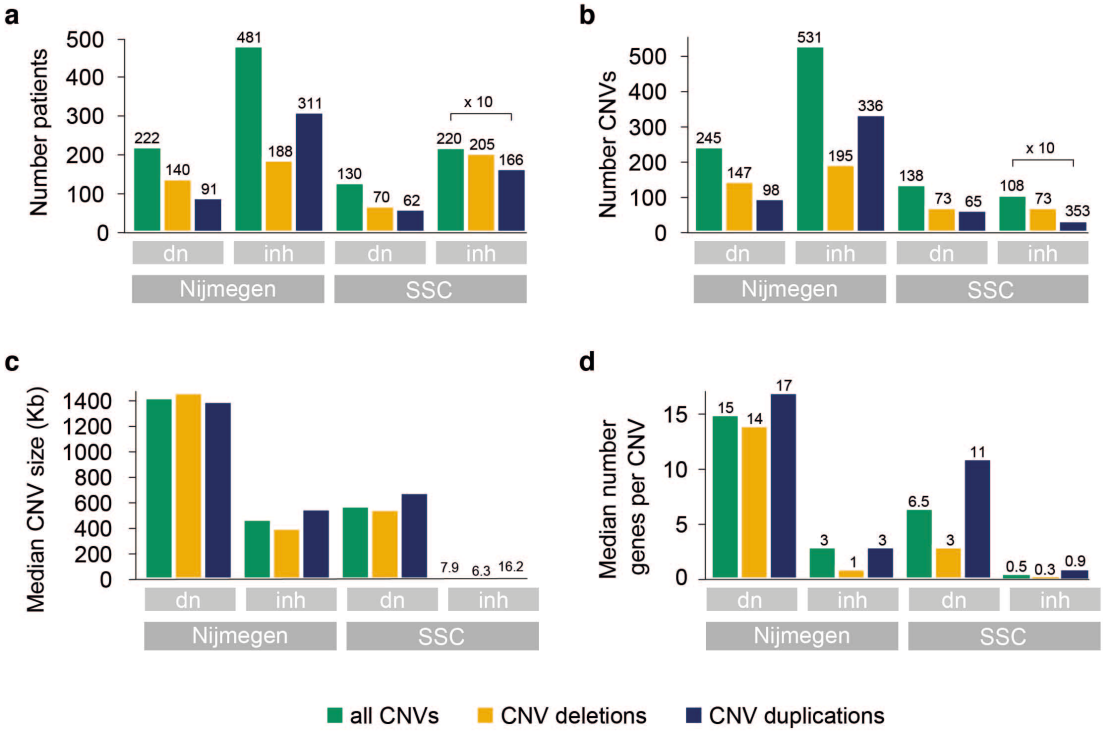


Fig 4.4 | CNV burden of the Nijmegen and SSC cohorts. The barplots depict (**a**) the number of Nijmegen or SSC patients carrying CNVs, (**b**) the total number of CNVs called in each dataset, (**c**) the median CNV size and (**d**) the median number of genes per CNV. Individuals carrying *de novo* (*dn*) or inherited (inh) CNVs were analysed separately. CNV burden and characteristics are shown for all CNVs (green) as well as for deletions (yellow) and duplications (dark blue), separately.

Phenotypic examination of patients with *dn* CNVs revealed 172 (77%), 90 (41%), 22 (10%) and 39 (18%) have varying degrees of ID, DD, CAs and ASD, respectively. Similarly, 333 (69%), 163 (34%), 39 (8%) and 78 (16%) patients with inh CNVs have the corresponding conditions (**Fig S4.2a-b**). Overall, Nijmegen patients are annotated with between 3 and 182 HPO terms

each, spanning a total of 1,033 distinct abnormalities (**Fig 4.5**). Patients with inh CNVs are annotated with fewer terms ($p = 1.93 \times 10^{-4}$; left-tailed Mann-Whitney U test) and a smaller proportion has ID ($p = 0.014$, OR = 0.65; left-tailed Fisher test), supporting the view that *dn* CNVs cause more or more severe phenotypes.

The concept of phenotypic similarity (or coherence), describing the agreement of clinical features between individuals, is central to this study. A phenotypic similarity score was derived for all pairwise combinations of patients using the *Goodall3* similarity index (**section 2.6**) (Boriah *et al.*, 2008). This score reflects the number of traits two individuals share or lack, weighted by their frequency. Hence, individuals are phenotypically more similar if they present with or without the same rare or common phenotypes, respectively.

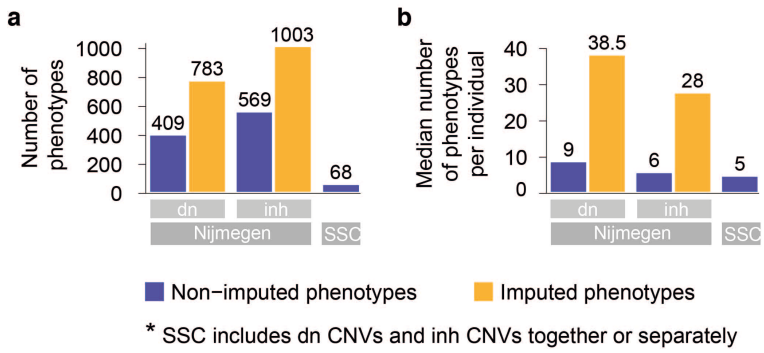


Fig 4.5 | Phenotypic burden of the Nijmegen and SSC cohorts. (**a**) The number of distinct HPO (Human Phenotype Ontology) terms and phenotypes observed in Nijmegen and SSC patients, respectively. (**b**) The median number of annotations per patient. Non-imputed phenotypes (blue) refer to clinically annotated abnormalities whereas imputed phenotypes include all parent terms of the HPO terms displayed by Nijmegen patients (SSC patients are clinically characterised without a phenotypic ontology). The same SSC statistics are observed across patients with *de novo* (*dn*) or inherited (inh) CNVs.

4.4.2 Coherence of neurodevelopmental CNV carriers

We constructed groups of Nijmegen patients carrying *dn* CNVs previously implicated in NDDs, such as the 16p11.2 and 22q11.21 deletions and duplications, and evaluated the similarity of their phenotypes (i.e. HPO terms) (**Fig 4.6a**). Of the twelve neurodevelopmental CNVs assessed, phenotypic coherence was only observed for individuals harbouring 7q11.23 *dn* microduplications. As expected, the seven carriers have some degree of ID, although they lack the typical speech and motor skill impairments and other comorbidities associated with the syndrome. Given the incomplete penetrance and pleiotropy of neurodevelopmental CNVs, the evaluated patient groups, which are small in size ($3 \leq n \leq 18$ patients), are presumably unable to recapitulate the full phenotypic landscape of the evaluated syndromes. Hence, small groups inevitably lack sufficient statistical power for detection. This prompted us to move beyond the study of single CNVs and into the examination of functionally related CNV genes, based on the observation that disrupting functionally related genes is likely to affect the same molecular pathways and yield similar phenotypes (Oti and Brunner, 2007).

4.4.3 Identification of modules of co-expressed genes

We sought to evaluate whether functionally related genetic disturbances translate at the phenotypic level. For this, we formed modules of genes similarly expressed across human tissues, characterised them functionally and compared the phenotypes of the corresponding CNV carriers.

Importantly, this analysis relied on the dual observation that (1) shared patient symptoms arise from the disruption of common biological pathways (Oti and Brunner, 2007) and that (2) genes expressed in similar patterns are more likely to work in the same functional processes (Lee *et al.*, 2004).

The GTEx gene co-expression network depicts the body-wide similarity of gene expression patterns. The GTEx sub-network, consisting of genes mapping to *dn* CNV genes from Nijmegen patients, comprises 79,653 links between 2,390 genes. This network is modular (modularity of network Q = 0.63) and thus readily divisible into modules of strongly co-expressed genes. Clustering using the Louvain algorithm revealed 27 modules comprising up to 203 genes each. We, then, constructed non-exclusive patient groups by assigning individuals to all modules their CNV genes map to. Patients belong to a median of 6 modules per individual although to increase statistical power, we only assessed the 11 modules with ≥ 20 genes and ≥ 10 individuals.

We functionally characterised the uncovered gene modules using Gene Ontology (GO) annotations and we evaluated the mouse abnormalities yielded following the disruption of their unique (1:1) orthologue using the Mouse Genome Informatics database (MGI; Blake *et al.*, 2017) (**Figs S4.3 and S4.4**):

- Module 0 genes (n = 130) are associated with nervous system development, based on the over-representation of GO terms relating to neurogenesis, neuronal differentiation and axonogenesis.

Correspondingly, mouse KnockOuts (KOs) display nervous system and behavioural phenotypes, such as synaptic transmission dysfunction and abnormalities of cerebral morphology.

- Module 1 genes (n = 100) are involved in muscle development and contraction, with enriched GO terms relating to mitochondrial energy metabolism and the electron chain transport. Muscle and in particular skeletal muscle morphology defects are observed in mouse models.
- Module 2 genes (n = 74) are nominally enriched for spermatogenesis-related GO terms and reproductive system MGI phenotypes.
- The function of module 3 genes (n = 110) is unclear (no enrichments at $p < 0.05$).
- Module 4 genes (n = 154) are associated with various metabolic processes and RNA processing and cause embryonic lethality, defects or abnormal development in mouse models.
- Module 5 genes (n = 98) are enriched for immune system processes, such as programmed cell death. Accordingly, mouse models display immune and hematopoietic system phenotypes, such as abnormal immune cell physiology and morphology.
- Module 9 genes (n = 117) are involved in extracellular matrix organisation and morphogenesis, with morphological and vascular abnormalities arising in mouse models.
- Module 10 genes (n = 88) are vaguely associated with lipid metabolism and cell junction organisation.

- Module 12 genes (n = 29) participate in various metabolic processes and localise to extracellular vesicles and exosomes. Their disruption causes defects of the liver and biliary systems in mouse models, including abnormalities in lipid and cholesterol levels and homeostasis.
- Module 13 genes (n = 94) are involved in chromatin organisation, mitosis, cell cycle as well as regulation of immune system processes. Immune and hematopoietic system abnormalities, such as immune cell development, proliferation and morphology are observed in mouse KOs.

Based on these characteristics, modules underlie disparate biological processes, many of which have been implicated in the aetiology of NDDs (**sections 1.3 and S1**). As modules are associated with broad functions and cause extensive phenotypes in mouse models, predicting the corresponding human abnormalities is challenging. Moreover, assessments of patient phenotypes revealed no HPO terms are over- or under-represented in patients from particular modules, following correction for multiple testing (**Fig S4.5a**).

4.4.4 Coherence of patients with similar CNV disruptions

Overall, patients with disruptions in co-expressed genes, as given by common GTEx module memberships, are phenotypically more similar than individuals with disparate disruptions (i.e. with different module memberships) ($p = 0.0050$; right-tailed Mann-Whitney U test) (**Fig S4.6a**). This supports the

notion of genotype to phenotype convergence. Specifically, nine of the ten analysed modules are associated with a stronger overall phenotypic similarity than expected by chance (**Fig 4.6a**).

Interestingly, no phenotypic coherence was observed among patients mapping to module 0. Module 0 genes are brain-specific (i.e. module 0 comprises 93% of brain-specific genes (n = 112) from the cohort; $p < 1 \times 10^{-10}$, OR = 139.7; right-tailed Fisher test) and encode proteins with nervous system functions. As the corresponding individuals exhibit abnormalities of higher mental function, their phenotypic inconsistencies are likely to arise from the variability of observed conditions and accompanying comorbidities. This suggests that while disruptions of brain-specific functions globally predispose towards neuropsychiatric illnesses, they may lack specificity.

To elucidate the features driving the clinical resemblance of patients from modules 1, 2, 3, 4, 5, 9, 10, 12 and 13, we iteratively re-examined the observed phenotypic similarity following the sequential exclusion of the top 1 to 500 HPO terms most strongly associated with each module (based on increasing p-values). For each module, we identified the collection of phenotypes required to achieve significant phenotypic similarity, such that their removal dissipates the similarity signal (herein termed *essential terms or phenotypes*) (**Fig 4.6b 4.6c, right panel**). We postulate these terms specifically underlie phenotypic similarity in a combinatorial manner. Conversely, removing randomised sets of essential terms, constructed by randomly choosing the same number of phenotypes, does not affect

significance (**Fig 4.5c, left panel**). These results indicate that the disruption of genes from particular modules predisposes towards the corresponding essential phenotypes.

The number of essential HPO terms ranges between 25 and 480, with a median of 180 terms per module (**Fig 4.6c, right panel**), suggesting that combinations of presence, and in particular absence, of large numbers of phenotypes (**Fig S4.7a**) influences the observed clinical resemblance. As HPO terms are not individually over- or under-represented, it is unlikely that individuals form tightly interconnected networks of phenotypically similar patients, displaying or lacking a common set of abnormalities. Instead, we speculated they are organised as acyclic path trees, such that similarity diffuses in a linear manner although individuals are highly similar to only few others (**Fig 4.3**). Accordingly, the phenotypic path connecting patients from opposite ends, who presumably display somewhat inconsistent abnormalities, should be well established, hence supporting a spectrum view of NDDs. Exploiting the property of acyclic path graphs to lose their integrity following the removal of growing numbers of nodes (**Fig 4.3**), we observed a negative relationship between the strength of the phenotypic coherence signal and the number of module patients swapped for background individuals ($p < 0.037$ and slope < 0, linear regression) (**Fig 4.6d-e**). Specifically, this means that as individuals within the module were randomly swapped for those outside, the network of phenotypically similar patients begins to breakdown. Conversely, for phenotypically discordant modules, the relationship between phenotypic similarity and number of swapped patients is either (1) insignificant, meaning

there is no relationship between the patient composition within the new modules and their phenotypic similarity; or (2) positive (**Fig 4.6f-g**). The latter (2) suggests that randomly switching individuals increases the likelihood of observing phenotypic coherence. Taken together, these observations suggest that while *dn* CNVs affecting co-expressed genes globally influence more similar disease presentations, no particular abnormality ensues. Instead, converging functional disruptions underlie a phenotypic spectrum, which will be shaped through the assessment of growing numbers of patients.

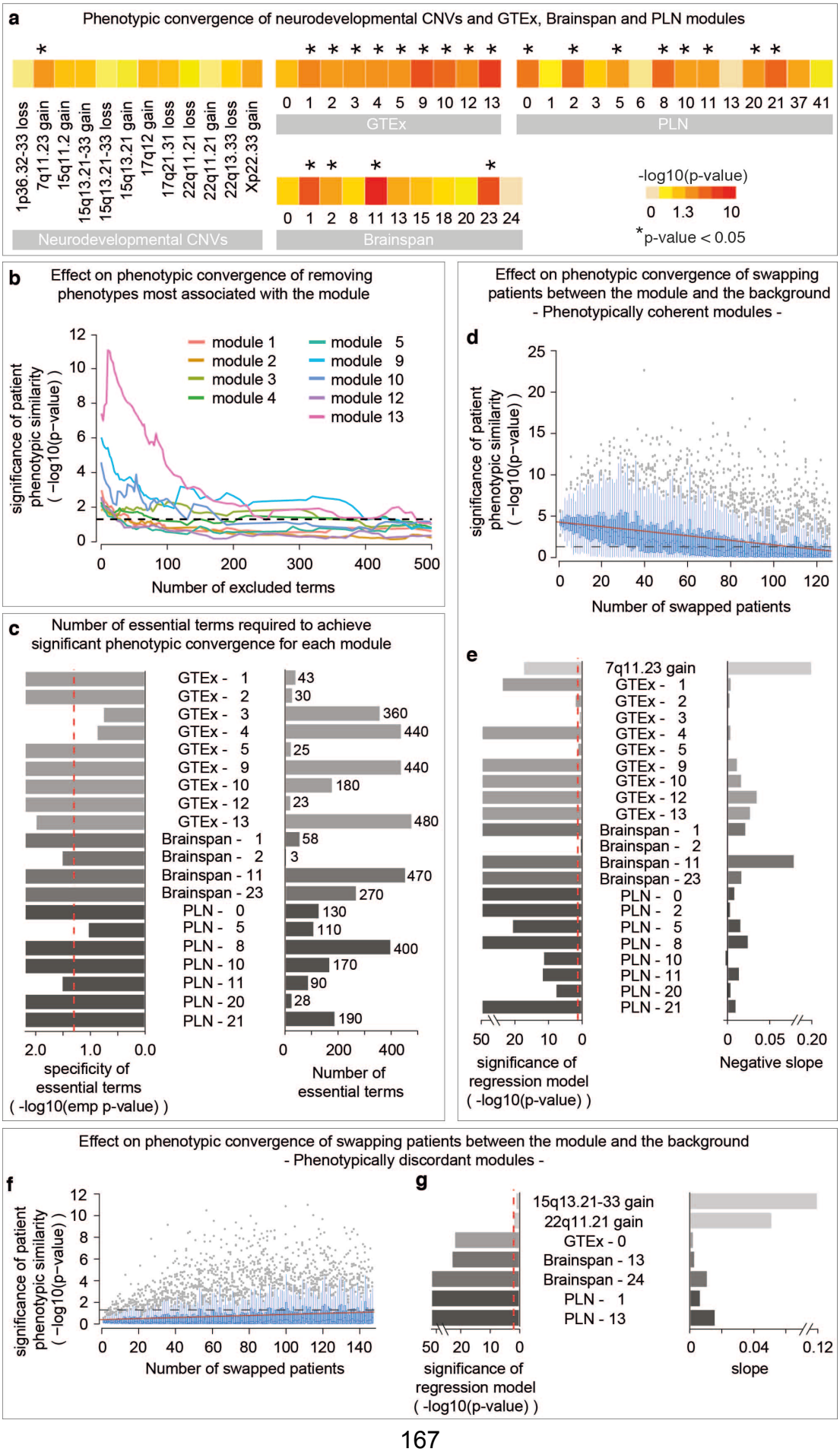

a
Phenotypic convergence of neurodevelopmental CNVs and GTEx, Brainspan and PLN modules
1p36.32-33 loss
7q11.23 gain
15q11.2 gain
15q13.21-33 gain
15q13.21-33 loss
15q13.21 gain
17q12 gain
17q21.31 loss
22q11.21 loss
22q11.21 gain
22q13.33 loss
Xp22.33 gain
Neurodevelopmental CNVs
0 1 2 3 4 5 9 10 12 13
GTEx
0 1 2 3 5 6 8 10 11 13 20 21 37 41
PLN
0 1 2 8 11 13 15 18 20 23 24
Brainspan
-log10(p-value)
0 1.3 10
*p-value < 0.05
b
Effect on phenotypic convergence of removing phenotypes most associated with the module
significance of patient phenotypic similarity (-log10(p-value))
module 1
module 2
module 3
module 4
module 5
module 9
module 10
module 12
module 13
Number of excluded terms
c
Number of essential terms required to achieve significant phenotypic convergence for each module
GTEx - 1 43
GTEx - 2 30
GTEx - 3 360
GTEx - 4 440
GTEx - 5 25
GTEx - 9 440
GTEx - 10 180
GTEx - 12 23
GTEx - 13 480
Brainspan - 1 58
Brainspan - 2 3
Brainspan - 11 470
Brainspan - 23 270
PLN - 0 130
PLN - 5 110
PLN - 8 400
PLN - 10 170
PLN - 11 90
PLN - 20 28
PLN - 21 190
specificity of essential terms (-log10(emp p-value))
Number of essential terms
d
Effect on phenotypic convergence of swapping patients between the module and the background
- Phenotypically coherent modules -
significance of patient phenotypic similarity (-log10(p-value))
Number of swapped patients
e
7q11.23 gain
GTEx - 1
GTEx - 2
GTEx - 3
GTEx - 4
GTEx - 5
GTEx - 9
GTEx - 10
GTEx - 12
GTEx - 13
Brainspan - 1
Brainspan - 2
Brainspan - 11
Brainspan - 23
PLN - 0
PLN - 2
PLN - 5
PLN - 8
PLN - 10
PLN - 11
PLN - 20
PLN - 21
significance of regression model (-log10(p-value))
Negative slope
Effect on phenotypic convergence of swapping patients between the module and the background
- Phenotypically discordant modules -
f
significance of patient phenotypic similarity (-log10(p-value))
Number of swapped patients
g
15q13.21-33 gain
22q11.21 gain
GTEx - 0
Brainspan - 13
Brainspan - 24
PLN - 1
PLN - 13
significance of regression model (-log10(p-value))
slope

Fig 4.6 | Genotype and phenotype convergence in Nijmegen patients with *de novo* CNVs. (**a**) Phenotypic coherence of patients with similar functional disruptions, given by common CNVs or module memberships. Of note, while GTEx modules are discussed in **sections 4.4.3** and **4.4.4**, Brainspan and PLN modules are reviewed in **section 4.4.5**. Significance was derived by comparing pairwise phenotypic similarities of group patients and remaining cohort patients using right-tailed Mann-Whitney U tests. (**b**) Phenotypic coherence of patients from GTEx modules following the removal of increasing numbers of HPO terms from the analysis (p-values are derived from right-tailed Mann-Whitney U tests; y-axis). Excluded phenotypes consist of those most strongly associated with patients from the evaluated modules; they are chosen based on increasing p-values from two-sided Fisher tests. Essential terms refer to those whose exclusion suppresses the similarity signal ($p > 0.05$; black dotted line). (**c, right panel**) The number of essential terms for each module. (**c, left panel**) To examine the specificity of essential terms, phenotypic coherence was re-evaluated for each module, following the exclusion of 100 permuted sets of essential terms (formed by randomly sampling the same number of HPO terms present in the cohort). An empirical p-value, shown here, was derived to reflect the proportion of permutations involving significant phenotypic similarity. (**d to g**) Phenotypic coherence following the swap of increasing numbers of patients between modules and the background. During each swap, we ran 100 permutations. Linear regression models were used to describe the relationship between phenotypic coherence and the number of swapped individuals. (**d, f**) Examples of the effect on phenotypic coherence of swapping increasing numbers of patients between the background and (**d**) a phenotypically coherent module (i.e. GTEx module 13) or (**f**) a phenotypically discordant module (negative control; GTEx module 0). The red line depicts the associated regression model. (**e, g**) The (**left panel**) significance and (**right panel**) slope of regression models fitted to (**e**) phenotypically coherent and (**g**) discordant modules.

4.4.5 Replicating coherence using auxiliary genomics datasets

To verify that the observed genotype to phenotype convergence is robust to the type of functional associations used to construct modules, we identified gene modules using two additional genomics resources and repeated the analysis.

Firstly, in view of the neurological phenotypes of patients from the cohort, we employed a spatiotemporal brain gene co-expression network constructed using Brainspan data (Sunkin *et al.*, 2013). Secondly, we used an integrated network, termed Phenotypic Linkage Network (PLN), in which gene links

depict likelihoods of functional interaction based on multiple sources of evidence (Honti *et al.*, 2014). The Brainspan network and PLN comprise 2,574 and 1,609 genes from *dn* Nijmegen CNVs, respectively. Brainspan and PLN sub-networks were separated into 28 and 49 modules including up to 496 and 403 genes each, respectively. The analysis was restricted to 11 and 14 modules, which were assigned ≥ 20 genes and ≥ 10 individuals, respectively (i.e. between 12-153 and 21-173 patients per module).

Although some modules have similar functional annotations to those identified using the GTEx network, most do not overlap at 1:1 ratios (**Fig S4.8a**). Specifically, analogous modules linked to (1) transcriptional regulation and RNA processing and (2) synaptic functions were identified, while remaining modules vary in terms of functional annotations.

In line with previous observations, phenotypic similarity is globally stronger among patients from the same modules as compared to individuals from different modules ($p = 1.50 \times 10\text{-}6$ for Brainspan and $p = 3.26 \times 10\text{-}4$ for PLN modules; right-tailed Mann-Whitney U tests) (**Fig S4.6b-c**). Nevertheless, the signal is weaker than for GTEx modules, as evidenced by a lower proportion of phenotypically similar modules (i.e. 4/11 Brainspan and 8/14 PLN modules) (**Fig 4.6a**). The greater performance of GTEx could be due to the use of body-wide expression data as Brainspan brain-only data may be inadequate to discriminate between neurodevelopmental abnormalities, despite being advantageous to recognise them. Conversely, owing to the study biases of PLN annotations (i.e. GO, MGI and KEGG), module construction may have

centred on known disease genes, hence impeding the identification of functionally distinct modules. Correspondingly, many Brainspan and PLN modules involve major neurodevelopmental themes, namely regulation of gene expression through chromatin modification or transcriptional control and synaptic transmission.

Modules made from Brainspan and PLN networks and comprising phenotypically similar individuals recapitulate the observations made for GTEx modules. Specifically, (1) no specific phenotypes are over- or under-represented (**Fig S4.5b-c**). Instead, (2) large numbers of abnormalities influence phenotypic coherence. The removal of 3-470 and 28-400 most strongly associated HPO terms (i.e. essential terms) from Brainspan and PLN modules, respectively, is required to suppress the similarity signals (**Fig 4.6b-c**) (PLN module 2 was excluded from the study as removal of the top 500 HPO terms did not suppress the signal). Moreover, (3) we observed a strong negative relationship between phenotypic similarity and the integrity of the patient composition of phenotypically coherent modules. Hence, these results support a spectrum view of NDDs and highlight the importance of statistical power in shaping it (**Fig 4.6d-g**).

Taken together, these results advise that although the identification of unequivocal phenotype and genotype associations may be infeasible or challenging, the evaluation of increasing numbers of patients will shape the associated genetic and phenotypic landscapes. In turn, this will facilitate the

inference of aetiopathologies based on the similarity of patients' holistic clinical abnormalities and vice-versa.

4.4.6 Replicating coherence using patients from DECIPHER

We, next, sought to validate the identified phenotypic associations beyond our discovery cohort. For this, we turned to an independent set of patients and asked whether individuals carrying *dn* CNVs affecting genes, forming the previously identified phenotypically coherent modules, display analogous clinical abnormalities. We considered 1,604 patients carrying *dn* CNVs from the DECIPHER database of individuals with rare diseases. While patients were not as systematically phenotyped as those from the Nijmegen cohort, clinical data was provided in the form of HPO annotations, therefore facilitating comparisons.

We mapped DECIPHER patients to phenotypically coherent Nijmegen modules based on their CNV disruptions and evaluated the similarity of their abnormalities compared to that of remaining DECIPHER patients. To ensure our analyses are comparable, phenotypic similarity was measured across the subset of HPO terms displayed by Nijmegen patients. Nevertheless, results were replicated when considering the full clinical spectrum of DECIPHER patients (i.e. 365 additional HPO terms) (not shown).

For a majority of evaluated modules (7/9 GTEx, 3/4 Brainspan and 6/8 PLN modules), phenotypic coherence persists across DECIPHER patients,

supporting the relevance of our previous observations (**Fig 4.7a**). This effect is specific as permuting a third of module patients eliminates the similarity signal (**Fig 4.7b**). Additionally, we note the considerable gain in statistical power of detection associated with the interrogation of a larger cohort, as suggested by the highly significant p-values. Using our phenotypic pipeline (**section 4.3.8**), we identified sets of essential phenotypes, driving the phenotypic similarity of DECIPHER patients from these modules (i.e. between 33 and 480 essential terms per module; phenotypic coherence persists for PLN module 5 following the removal of the top 500 most strongly associated abnormalities). An excess of these essential phenotypes are over-represented among essential HPO terms identified using the corresponding Nijmegen patient modules (**Fig S4.9**), establishing the accuracy of our approach. These results provide corroborative evidence that the disruption of genes, participating in common biological process, as given by shared module memberships, underlie the development of more similar phenotypic spectra, encompassing particular sets of abnormalities.

Reasoning that CNVs from Nijmegen patients capture only a portion of genetic variance in NDDs, we expanded modules by considering all direct interactors of the protein products of the corresponding module genes, based on the String Protein-Protein Interaction (PPI) network. However, no phenotypic coherence was observed among DECIPHER patients contributing with CNV gene(s) to the newly expanded modules (**Fig 4.7c**). This could be a consequence of the dramatic increase in number of patients mapping to each

module (i.e. > 1,000 individuals per module, representing over a 2-fold increase in module sizes) and the substantial gene and patient overlap between groups. We speculate this large overlap is in part due to ascertainment biases present in PPI networks, such that common sets of well known and hub genes, with widespread expression patterns and functions, are consistently merged with modules, inherently weakening specificity.

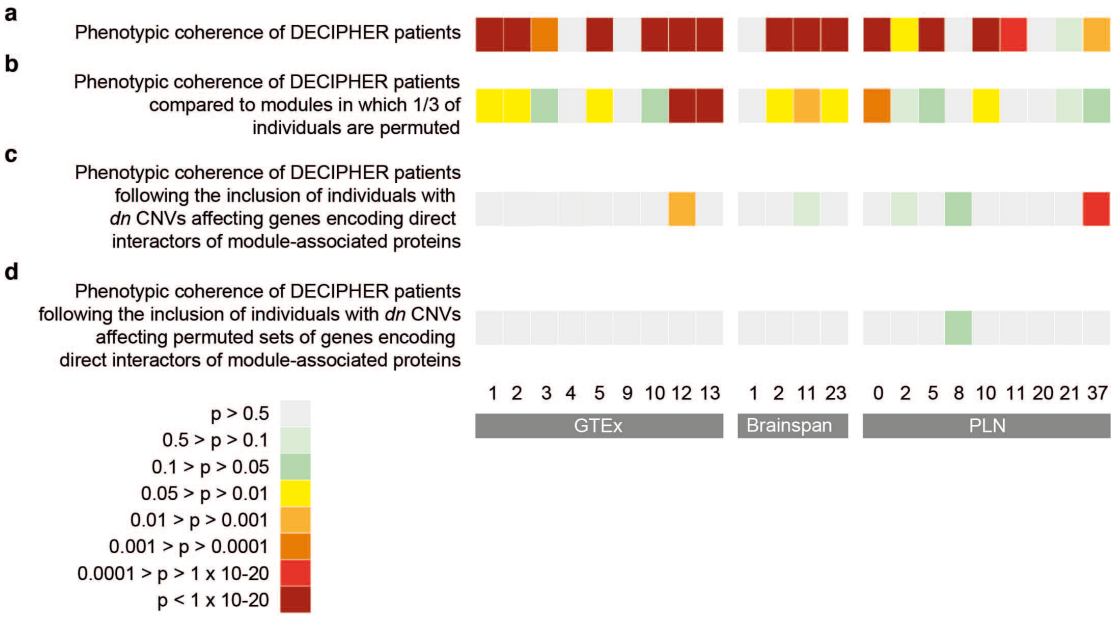


Fig 4.7 | Replication of phenotypic coherence using DECIPHER patients. (**a, b**) Phenotypic similarity of DECIPHER patients carrying *de novo* (*dn*) CNVs affecting genes from phenotypically coherent Nijmegen modules compared to (**a**) remaining DECIPHER patients and (**b**) 1,000 randomised patients sets, constructed by permuting a third of individuals. (**c, d**) Phenotypic similarity across DECIPHER patients mapping to expanded Nijmegen modules, constructed by including (**c**) genes, which encode proteins interacting with the products of module genes based on protein-protein interaction data (PPIs) from String or (**d**) the same number of randomly sampled genes. (**a, c, d**) Comparisons were performed using right-tailed Mann-Whitney U tests. (**b**) Significance was derived by evaluating the proportion of permutations, during which module patients are more similar than randomised modules. Displayed here are the associated empirical p-values. Of note, String PPIs were restricted to genes from Nijmegen modules.

4.4.7 Coherence of patients with inherited CNVs

Aiming to replicate the observed genotype to phenotype convergence, we repeated the analysis using the 481 Nijmegen patients carrying a total of 531 inherited (inh) CNVs. Following the same methodology, we formed (1) groups of neurodevelopmental CNV carriers and (2) modules of patients whose CNV genes shared an excess of connections within the GTEx and Brainspan co-expression networks and the PLN. The exclusion of modules with < 20 genes and < 10 individuals left 11 GTEx, 11 Brainspan and 14 PLN modules for analysis comprising between 26-178, 12-135 and 10-245 patients, respectively. An excess of patients have unique module memberships compared to patients with *dn* CNVs ($p < 1 \times 10^{-7}$, OR > 3.3; right-tailed Fisher tests), possibly owing to shorter and less genic CNVs.

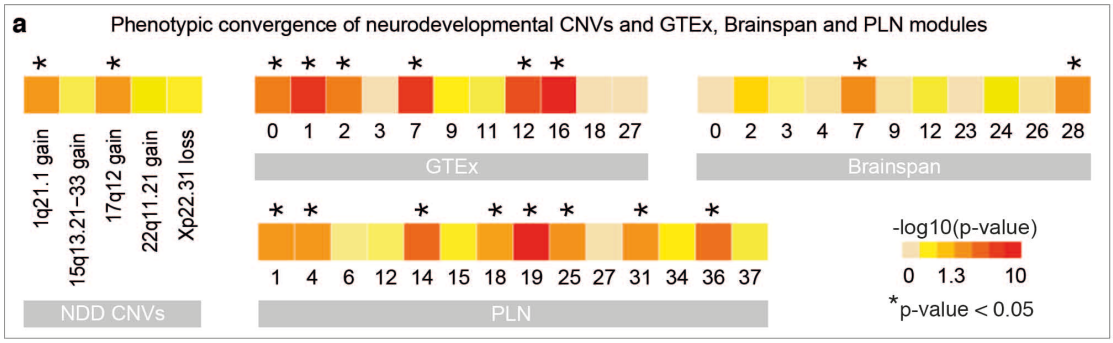
Although the low incidence of individual neurodevelopmental CNVs undoubtedly reduced statistical power, we observed strong phenotypic similarity among carriers of inh 17q12 and 1q21.1 duplications (**Fig 4.8a**). Individuals with the 1q21.1 duplication display both cognitive and eye abnormalities, although the incomplete CNV penetrance, high phenotypic variability and the lack of independent study cohorts hinders deeper investigation of this CNV. On the other hand, the phenotypic coherence observed for 17q12 duplications was suggestively reproduced among Nijmegen patients with *dn* CNVs and in the combined cohort ($p = 0.083$ and $p = 4.46 \times 10^{-5}$, respectively; right-tailed Mann-Whitney U tests). Of the nine carriers, six have cognitive impairments, among which three are diagnosed

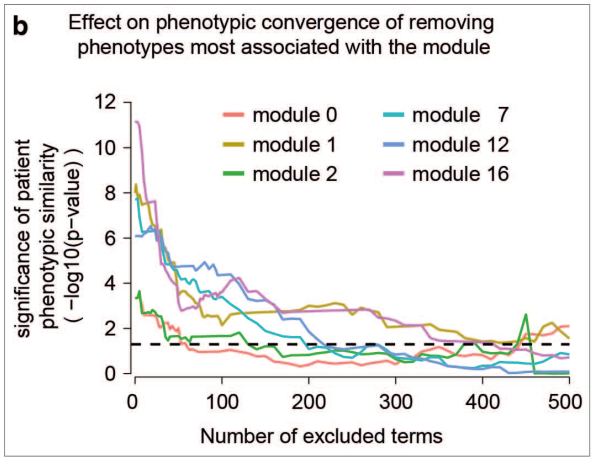
with ID and language/speech delays, consistent with the typical learning disabilities, speech and language delays present in the microduplication syndrome. Nonetheless, in line with the wide comorbidity spectrum, carriers also display additional abnormalities, which are not individually over-represented.

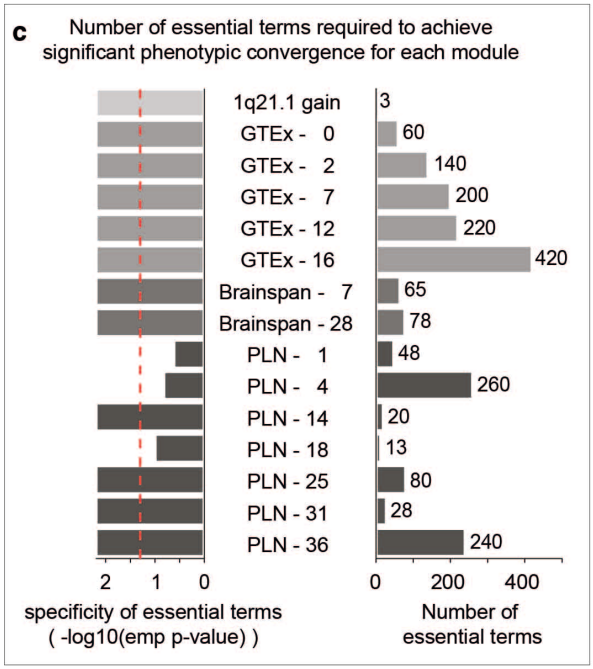
Phenotypic coherence is globally replicated using the three networks, such that patients from the same module present more similar clinical abnormalities than those from different modules ($p = 1.36 \times 10^{-9}$ for GTEx, $p = 0.022$ for Brainspan and $p = 2.24 \times 10^{-22}$ for PLN; right-tailed Mann-Whitney U tests) (**Fig S4.6d-f**). Nevertheless, there is an overall reduction in number of phenotypically coherent gene modules (6/11 GTEx, 2/11 Brainspan and 8/14 PLN modules) (**Fig 4.8a**). This decline could be due to an increase in number of neutral modules or modules only weakly associated with clinical abnormalities, owing to the lower likelihoods of pathogenicity of inh compared to *dn* CNVs.

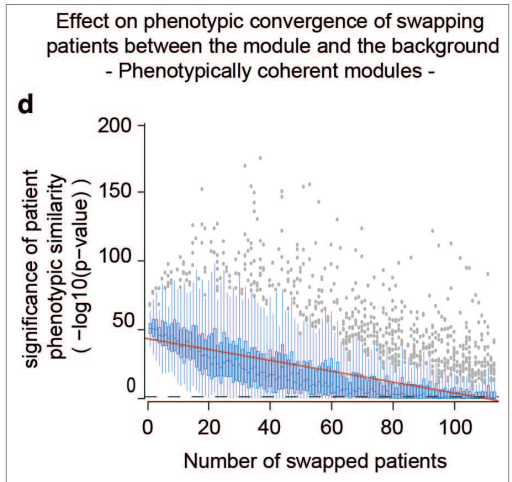
The uncovered modules (i.e. inh modules) recapitulate themes observed for those identified using genes mapping to *dn* CNVs from Nijmegen patients (i.e. *dn* modules). These themes include transcriptional regulation, nervous system development, synaptic transmission, cell adhesion and ubiquitination. In fact, many inh and *dn* modules match at approximately 1:1 ratios (**Fig S4.8d-f**), although the significance of phenotypic patient similarities is not consistently replicated for overlapping modules. This may be a consequence of the presence of additional neutral genes in modules with no influence over clinical

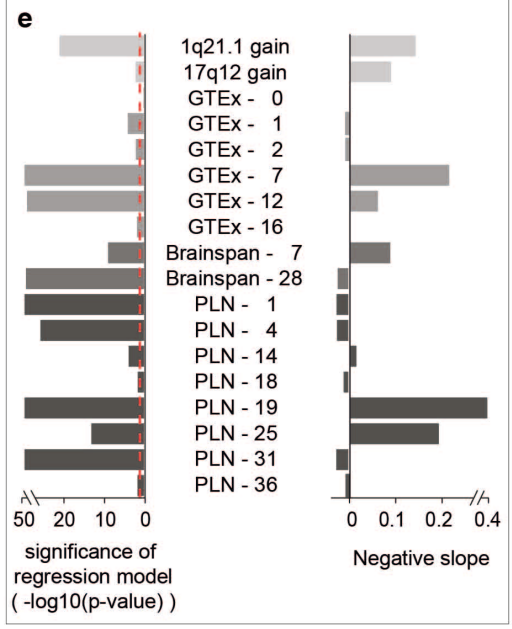
outcomes. Although alternate explanations for this lack of replicability can be proposed, they are not as potent. Indeed, (1) the lack of similarity cannot be attributed to reduced statistical power as *dn* and inh modules comprise similar numbers of patients ($p > 0.5$; two-sided student t tests). Additionally, (2) the observations made using *dn* modules are unlikely to be false positives as they were reproduced using the DECIPHER cohort. Lastly, (3) it is improbable that *dn* and inh CNVs carry differential impacts based on the observation that patients from overlapping modules are phenotypically more similar to one-another than to remaining patients ($p < 0.032$; right-tailed Mann-Whitney U tests; not shown).


a
Phenotypic convergence of neurodevelopmental CNVs and GTEx, Brainspan and PLN modules
1q21.1 gain
15q13.21−33 gain
17q12 gain
22q11.21 gain
Xp22.31 loss
NDD CNVs
0 1 2 3 7 9 11 12 16 18 27
GTEx
0 2 3 4 7 9 12 23 24 26 28
Brainspan
1 4 6 12 14 15 18 19 25 27 31 34 36 37
PLN
-log10(p-value)
0 1.3 10
*p-value < 0.05


b
Effect on phenotypic convergence of removing phenotypes most associated with the module
significance of patient phenotypic similarity (−log10(p−value))
module 0
module 1
module 2
module 7
module 12
module 16
Number of excluded terms


c
Number of essential terms required to achieve significant phenotypic convergence for each module
1q21.1 gain
GTEx - 0
GTEx - 2
GTEx - 7
GTEx - 12
GTEx - 16
Brainspan - 7
Brainspan - 28
PLN - 1
PLN - 4
PLN - 14
PLN - 18
PLN - 25
PLN - 31
PLN - 36
3
60
140
200
220
420
65
78
48
260
20
13
80
28
240
specificity of essential terms (-log10(emp p-value))
Number of essential terms


Effect on phenotypic convergence of swapping patients between the module and the background
- Phenotypically coherent modules -
d
significance of patient phenotypic similarity (−log10(p−value))
Number of swapped patients


e
1q21.1 gain
17q12 gain
GTEx - 0
GTEx - 1
GTEx - 2
GTEx - 7
GTEx - 12
GTEx - 16
Brainspan - 7
Brainspan - 28
PLN - 1
PLN - 4
PLN - 14
PLN - 18
PLN - 19
PLN - 25
PLN - 31
PLN - 36
significance of regression model (-log10(p-value))
Negative slope

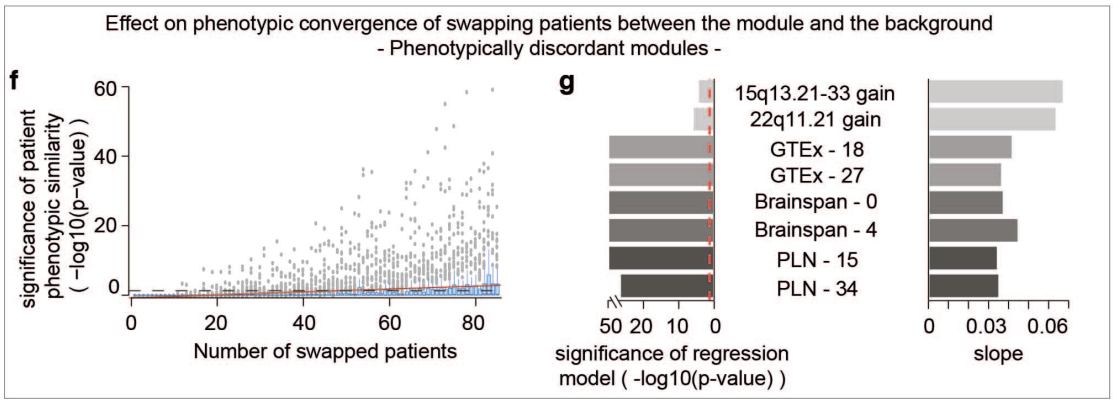

Effect on phenotypic convergence of swapping patients between the module and the background
- Phenotypically discordant modules -
f
significance of patient phenotypic similarity (−log10(p−value))
Number of swapped patients
g
15q13.21-33 gain
22q11.21 gain
GTEx - 18
GTEx - 27
Brainspan - 0
Brainspan - 4
PLN - 15
PLN - 34
significance of regression model (-log10(p-value))
slope

Fig 4.8 | Genotype and phenotype convergence in Nijmegen patients with inherited CNVs. (**a**) Phenotypic coherence of patients with similar functional disruptions, given by common CNVs or module memberships. Significance was derived by comparing pairwise phenotypic similarities of group patients and remaining patients using right-tailed Mann-Whitney U tests. (**b**) Phenotypic coherence of patients from GTEx modules following the removal of increasing numbers of HPO terms from the analysis (p-values are derived from right-tailed Mann-Whitney U tests; y-axis). Excluded phenotypes consist of those most strongly associated with patients from the evaluated modules; they are chosen based on increasing p-values from two-sided Fisher tests. Essential terms refer to those whose exclusion suppresses the similarity signal ($p > 0.05$; black dotted line). (**c, right panel**) The number of essential terms for each module. (**c, left panel**) To examine the specificity of essential terms, phenotypic coherence was re-evaluated for each module, following the exclusion of 100 permuted sets of essential terms (formed by randomly sampling the same number of HPO terms present in the cohort). An empirical p-value, shown here, was derived to reflect the proportion of permutations involving significant phenotypic similarity. (**d to g**) Phenotypic coherence following the swap of increasing numbers of patients between modules and the background. During each swap, we ran 100 permutations. Linear regression models were used to describe the relationship between phenotypic similarity and the number of swapped individuals. (**d, f**) Examples of the effect on phenotypic coherence of swapping increasing numbers of patients between the background and (**d**) a phenotypically coherent module (i.e. PLN module 19) or (**f**) a phenotypically discordant module (negative control; GTEx module 18). The red line depicts the associated regression model. (**e, g**) The (**left panel**) significance and (**right panel**) slope of regression models fitted to (**e**) phenotypically coherent and (**g**) discordant modules.

Examination of modules comprising phenotypically similar patients revealed the same trends, as those observed for patients with *dn* CNVs. Specifically, phenotypic coherence does not derive from particular clinical abnormalities (**Fig S4.5d-f**), but from the combination of a large number of HPO phenotypes. Between 13 and 420 essential terms were identified per module (phenotypic similarity persists following the removal of the top 500 HPO terms for GTEx module 1 and PLN module 19) (**Fig 4.8b-c**). Of note, strong phenotypic similarity is observed among patients from GTEx modules 0 and 2, Brainspan module 7 and PLN module 14, following the removal of large numbers (> 450) of HPO terms (**Fig S4.10a-d**). Nevertheless, signals, occurring beyond the exclusion of essential terms, may represent statistical

anomalies, rather than true biological signals. Indeed, the HPO terms remaining for analysis have an extremely low (n ≤ 3 patients) (or high) incidence in the cohort, and are not over- or under-represented among module patients (**Fig S4.7a-d**). Accordingly, the similarity signal is artificially inflated by the random segregation of very few individuals, presenting (or lacking) these phenotypes (unassociated with the module evaluated) within the module. It is, hence, important to acknowledge the limitations of the Goodall3 similarity index, which can artificially inflate patient similarities across extremely rare or common phenotypes, which are unlikely to be of biological relevance. Taken together, these results propose that while common molecular processes (recapitulated by modules) influence more similar clinical presentations than expected by chance, sample size is central to the statistical ability to detect phenotypic coherence (**Fig 4.8d-g**). Correspondingly, the analysis of the combined Nijmegen cohorts as well as the meta-analysis including all Nijmegen patients (with *dn*, inh and CNVs of unclear inheritance) revealed strong phenotypic coherence among patients with similar CNV disruptions ($p < 1 \times 10\text{-}10$; right-tailed Mann-Whitney U tests; not shown). Nevertheless, given the lower pathogenicity and higher incidence of inh CNVs, we advise assessing *dn* CNVs separately in the study of human disease.

4.4.8 Phenotypic coherence of ASD patients

Seeking to examine whether the phenotypic coherence of individuals with common functional disruptions extends across NDDs, we repeated the

analysis using the SSC cohort of individuals with ASD. Beyond core deficits in social interaction and communication and restricted and repetitive behaviours, which are themselves highly variable, ASD often presents with additional comorbidities and medical problems. Indeed, the disorder is often regarded as a constellation of phenotypes, some of which are more frequent or co-occur more often than others, making the SSC well suited to investigate genotype and phenotype relationships.

The SSC cohort includes 2,498 patients with phenotypic data concerning 63 clinical attributes, including psychiatric, developmental and neurological comorbidities, congenital abnormalities, immune and gastro-intestinal conditions and perinatal complications (**Table 4.1**) (the choice and construction of phenotypic variables are described in **section S3**). Patients possess up to 49 phenotypes, with a median of 5 phenotypes per individual (**Fig 4.5**). The SSC cohort presents several disadvantages worth mentioning. Firstly, consistent biases in clinical impressions are present across assessment centres (Fischbach and Lord, 2010). Secondly, the cohort is predominantly high functioning, possibly justifying the lower prevalence rates of comorbid disorders, as these are more common in low functioning patients (Cohen *et al*., 2014). Finally, while a systematic approach was undertaken to choose the phenotypes to assess, these remain subject to selection bias. Nevertheless, the deep and systematic phenotyping of SSC individuals enables accurate patient comparison, making this cohort well suited for this study.

Comparably to the Nijmegen cohorts, we formed datasets comprising 130 and 2,202 patients with 138 *dn* and 10,799 inh CNVs < 5 mb, respectively (**Fig 4.4a-b**). *Dn* and inh CNVs span a median of 580 kb and 8 kb, overlapping up to 68 and 44 genes, respectively. D*n* CNVs are longer and more gene-dense than inh CNVs ($p < 1 \times 10^{-16}$, right-tailed Mann-Whitney U tests) (**Fig 4.4c-d**). *Dn* CNVs include 73 losses and 65 gains while inh CNVs consist of 7,272 losses and 3,527 gains. In agreement with previous observations, inh gains are longer and more gene-dense than losses ($p < 1 \times 10^{-16}$, right-tailed Mann-Whitney U tests).

The study of *dn* CNVs was statistically underpowered given the low number of carriers (n = 130), the high overlap between modules and their small sizes (i.e. > 60% of modules include < 30 patients). In fact, our analysis revealed individuals occurring within the same modules are not phenotypically similar ($p > 0.9$ for GTEx, Brainspan and PLN; right-tailed Mann-Whitney U tests), prompting us to analyse the set of 2,202 SSC patients carrying inh CNVs.

The assessment of inh neurodevelopmental CNVs generated a significant phenotypic similarity signal for carriers of 16p11.2 duplications ($p = 0.011$) (**Fig 4.9a**). Although the phenotypic spectrum of this microduplication syndrome is large and lacks a distinguishing set of clinical features, the most common characteristics (i.e. microcephaly and low weight) are absent from the 13 SSC carriers. The most common abnormality observed is speech delay (i.e. *delay in use of first phrases*; 8/13 individuals) although this is not significant ($p = 0.41$, two-sided Fisher test). The validity of the observed

phenotypic coherence is further disputed as the similarity signal was not replicated using the whole SSC cohort (n = 2,498) or the subset comprising patients with *dn* CNVs ($p = 0.12$ and $p = 0.94$, right-tailed Mann-Whitney U tests). Moreover, reciprocal deletions associated with more severe phenotypes are not associated with phenotypic similarity (**Fig 4.9a**).

By partitioning the GTEx, Brainspan and PLN sub-networks composed of genes affected by inh CNV from SSC patients, we identified gene modules underlying distinct biological pathways, recapitulating those described for the Nijmegen cohorts (**Fig S8d-f**). Following the exclusion of statistically underpowered modules with < 20 genes and < 10 individuals, we evaluated 11 GTEx, 10 Brainspan and 15 PLN modules, comprising between 17-517, 12-381 and 13-528 SSC patients, respectively.

Patients occurring within the same modules are not globally more phenotypically similar than those from different modules ($p > 0.9$; right-tailed Mann-Whitney U tests) (**Fig S4.6g-i**). We, nevertheless, identified 1 GTEx, 4 Brainspan, 5 PLN phenotypically coherent modules (**Fig 4.9a**). As found with the Nijmegen NDD data, no clinical features are over-represented among patients from these modules (**Fig S4.5c**). Instead, their phenotypic similarity stems from the combination of presences and absences of large numbers of phenotypes (**Fig 4.9b-c**). The observed phenotypic similarity signals are strongly reduced when swapping increasing numbers of individuals from phenotypically coherent modules for background individuals ($p < 0.005$ and slope < 0 – except for Brainspan module 8 and PLN modules 2 and 4) (**Fig**

4.9d-e). Conversely, this pronounced negative relationship is not observed for phenotypically dissimilar modules (**Fig 4.9f-g**). Of note, the phenotypic coherence of patients from Brainspan module 9, following the removal of 58 phenotypes, represents a statistical anomaly, resulting from the previously discussed limitations associated with the Goodall3 similarity index (**Fig S4.7e**) (**section 4.4.7**).

Several limitations could have contributed to the overall decline in phenotypic coherence of the SSC cohort, including a sub-optimal choice of phenotypes to evaluate and the study of inh CNVs, which have a lower diagnostic yield and pathogenicity relative to *dn* CNVs. Another possibility is that *dn* rare exonic variants, whose role in ASD aetiology is well recognised (O'Roak *et al.*, 2011; O'Roak *et al.*, 2012; Neale *et al.*, 2012; Sanders *et al.*, 2012; Iossifov *et al.*, 2012; Iossifov *et al.*, 2014), underlie a substantial proportion of the phenotypic variability observed.

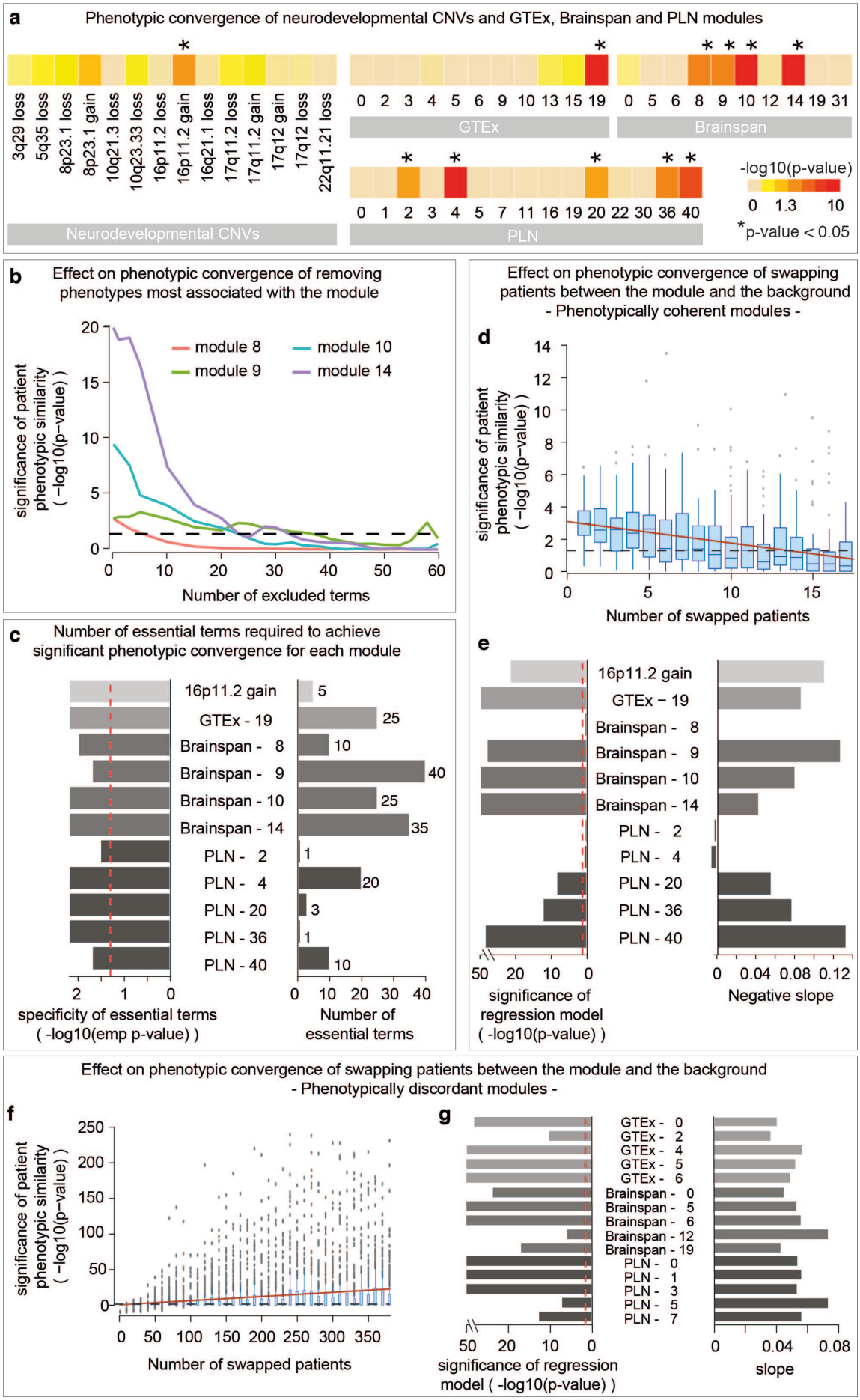

a
Phenotypic convergence of neurodevelopmental CNVs and GTEx, Brainspan and PLN modules
3q29 loss
5q35 loss
8p23.1 loss
8p23.1 gain
10q21.3 loss
10q23.33 loss
16p11.2 loss
16p11.2 gain
16q21.1 loss
17q11.2 loss
17q11.2 gain
17q12 gain
17q12 loss
22q11.21 loss
Neurodevelopmental CNVs
0 2 3 4 5 6 9 10 13 15 19
GTEx
0 5 6 8 9 10 12 14 19 31
Brainspan
0 1 2 3 4 5 7 11 16 19 20 22 30 36 40
PLN
-log10(p-value)
0 1.3 10
*p-value < 0.05
b
Effect on phenotypic convergence of removing phenotypes most associated with the module
significance of patient phenotypic similarity (-log10(p-value))
module 8
module 10
module 9
module 14
Number of excluded terms
c
Number of essential terms required to achieve significant phenotypic convergence for each module
16p11.2 gain
GTEx - 19
Brainspan - 8
Brainspan - 9
Brainspan - 10
Brainspan - 14
PLN - 2
PLN - 4
PLN - 20
PLN - 36
PLN - 40
5
25
10
40
25
35
1
20
3
1
10
specificity of essential terms (-log10(emp p-value))
Number of essential terms
d
Effect on phenotypic convergence of swapping patients between the module and the background - Phenotypically coherent modules -
significance of patient phenotypic similarity (-log10(p-value))
Number of swapped patients
e
16p11.2 gain
GTEx – 19
Brainspan - 8
Brainspan - 9
Brainspan - 10
Brainspan - 14
PLN - 2
PLN - 4
PLN - 20
PLN - 36
PLN - 40
significance of regression model (-log10(p-value))
Negative slope
Effect on phenotypic convergence of swapping patients between the module and the background - Phenotypically discordant modules -
f
significance of patient phenotypic similarity (-log10(p-value))
Number of swapped patients
g
GTEx - 0
GTEx - 2
GTEx - 4
GTEx - 5
GTEx - 6
Brainspan - 0
Brainspan - 5
Brainspan - 6
Brainspan - 12
Brainspan - 19
PLN - 0
PLN - 1
PLN - 3
PLN - 5
PLN - 7
significance of regression model (-log10(p-value))
slope

Fig 4.9 | Genotype and phenotype convergence in SSC patients with inherited CNVs. (**a**) Phenotypic coherence of patients with similar functional disruptions, given by common CNVs or module memberships. Significance was derived by comparing pairwise phenotypic similarities of group patients and remaining patients using right-tailed Mann-Whitney U tests. (**b**) Phenotypic coherence of patients from Brainspan modules following the removal of increasing numbers of phenotypes from the analysis (p-values are derived from right-tailed Mann-Whitney U tests; y-axis). Excluded phenotypes consist of those most strongly associated with patients from the evaluated modules; they are chosen based on increasing p-values from two-sided Fisher tests. Essential phenotypes refer to those whose exclusion suppresses the similarity signal ($p > 0.05$; black dotted line). (**c, right panel**) The number of essential terms for each module. (**c, left panel**) To examine the specificity of essential terms, phenotypic coherence was re-evaluated for each module, following the exclusion of 100 permuted sets of essential terms (formed by randomly sampling the same number of phenotypes present in the cohort). An empirical p-value, shown here, was derived to reflect the proportion of permutations involving significant phenotypic similarity. (**d to g**) Phenotypic coherence following the swap of increasing numbers of patients between modules and the background. During each swap, we ran 100 permutations. Linear regression models were used to describe the relationship between phenotypic similarity and the number of swapped individuals. (**d, f**) Examples of the effect on phenotypic coherence of swapping increasing numbers of patients between the background and (**d**) a phenotypically coherent module (i.e. Brainspan module 9) or (**f**) a phenotypically discordant module (negative control; GTEx module 0). The red line depicts the associated regression model. (**e, g**) The (**left panel**) significance and (**right panel**) slope of regression models fitted to (**e**) phenotypically coherent and (**g**) discordant modules (n.b. only five examples of discordant modules are shown for each network).

4.5 Discussion

In this study, we performed a systematic analysis of genotype and phenotype relationships in disease. Specifically, we exploited extensively phenotyped and genotyped cohorts of patients with NDDs, including DD, ID and ASD, to investigate the paradigm that shared genetic aetiologies underlie the development of distinguishable clinical subtypes of disease. By grouping individuals based on their CNV hits, we showed that patients with disruptions in functionally related genes are overall more likely to display similar aberrant phenotypes. However, this phenotypic coherence is associated with patients sharing or lacking combinations of many phenotypes, rather that exhibiting specific clinical abnormalities.

Our approach identified multiple functional gene clusters underlying phenotypic coherence. The observation that this coherence derives from the assortment of large numbers of clinical abnormalities points to strong pleiotropic effects and variable expressivity of the associated processes (**Figs 4.6, 4.8 and 4.9**). Indeed, no single abnormality is uniquely associated with a functional cluster (**Fig S4.5**). Instead, patients with the corresponding disruptions present with different combinations of a subset of phenotypes, globally making them more homogeneous than expected by chance. By replicating the phenotypic coherence of the uncovered pathways using the DECIPHER cohort, we were able to support the validity of this hypothesis (**Fig 4.7**). Furthermore, the discovery of phenotypically coherent gene groupings using inherited CNVs and in ASD supports the generality of this viewpoint

across NDDs and modes of CNV inheritance. Taken together, these findings propose that common molecular processes underlie similar phenotypic spectra. In fact, in order to shape the corresponding clinical landscapes, large numbers of patients (carrying the associated genetic lesions) must be assessed. The clinical variability of globally coherent patients is likely to be influenced by additional common or rare variants and environmental factors, which modulate gene pleiotropic effects and the associated variable expressivity. As a result, the interrogation of large cohorts will be necessary to develop a comprehensive overview of phenotypic variation in NDDs.

While we identified phenotypically coherent modules of genes mapping to both *dn* and inh variants, an overall decrease in signal was observed for inh CNVs (**Figs 4.6, 4.8 and 4.9**). This should not be unsurprising given their expected lower pathogenicity compared to *dn* variation. Hence, our genetic clustering approach must either distinguish between causal and neutral genes or these genes will corrupt gene groupings (modules). As a result, many modules are likely to comprise phenotypically inconsistent or a mixture of concordant and discordant patients, respectively. Regardless, this points to a strong reduction in statistical power, suggesting much larger samples are necessary to examine the clinical outcomes of inherited variation. Inevitably, the lack of positive findings could be a consequence of lack of biological signal, such that other factors including rare *dn* single base substitutions or environmental components underlie much of the observed phenotypic variance. Particularly, such scenarios could clarify the weak genotype to phenotype convergence observed in ASD (**Fig 4.9a**).

The notion of gene relatedness, central to the construction of gene and thus patient modules, is subject to interpretation and current opinions of biological relevance. Functional associations can be inferred from various types of genomics datasets, which vary in accuracy, specificity and completeness. To circumvent the introduction of ascertainment biases, we used unbiased body-wide or brain expression data to derive gene couplings, based on the assumption that co-expressed genes are more likely to participate in the same molecular processes (Lee *et al.*, 2004). Furthermore, to maximise coverage and minimise false positives, we also used an integrated gene network, which weights the confidence of interaction, based on multiple lines of evidence (Honti *et al.*, 2014). Functional clustering of genes mapping to patient CNVs using these resources revealed multiple affected biological pathways (**Figs S4.3, S4.4 and S4.8**). Notably, all three networks consistently uncovered (1) modules comprising synaptic genes expressed in the postnatal brain and (2) modules of foetally-expressed brain genes with roles in cell proliferation and control of gene expression, by regulating transcription or modifying chromatin states (**Fig S4.11**). Our results, therefore, implicate these two themes as the primary biological functions disrupted across NDDs, in line with previous ASD studies (Steinberg and Webber, 2013; Vorstman *et al.*, 2017) and our Gilles de la Tourette syndrome analysis (**chapter 3**). Beyond these processes, different functional genomics approaches uncovered some additional similar modules although many did not match (**Fig S4.8a-c**). Analogous modules were consistently identified in ID/DD and ASD using the same networks, supporting the notion of aetiological overlap across NDDs (**Figs S4.8d-f**). Nevertheless, there was a large variability in performance of each network to

detect associations underlying phenotypic coherence. This indicates different types of genomic resources should be systematically exploited as they may be of varying relevance in uncovering phenotypically homogeneous patient groups using a genotype-first approach.

The genotype-first approach has successfully been applied to many NDD genes in order to describe genotype and phenotype associations (Stessman *et al.*, 2016). Nevertheless, while the consideration of individual variants provides a powerful tool to elucidate the clinical subtypes displayed by patients with shared genetic lesions, it involves several limitations. Firstly, it requires the identification of adequately large numbers of individuals to circumvent biases associated with the phenotypic resolution captured by the sample. As a result, only those genes recurrently disrupted in NDDs will be interrogated. Secondly, it is inherently subject to clinicians' views of what constitutes a clinically relevant abnormality and relies on the manual inspection of each patient, therefore hindering the investigation of large numbers of variants and individuals. Here, we developed a standardised and automated framework to systematically group patients based on their functional disruptions and quantify the homogeneity of their phenotypes. Importantly, we demonstrated the considerable gain in statistical power to detect genotype to phenotype convergence using large clinical patient cohorts. Our methodology not only facilitates the interpretation of genome-wide structural variation but also enables the construction of maps of phenotypic variation in disease. The identified pathways, in turn, could be used to analyse patients with unknown disorders or classify individuals based

upon their clinical presentations and infer the likely affected molecular processes. Notably, our findings advise patients with developmental abnormalities (regardless of the DSM-5 (American Psychiatric Association, 2013) category they fall into) can be evaluated together (the Nijmegen cohort comprises patients with ID, ASD and learning disabilities). This holistic examination, nevertheless, necessitates systematic phenotyping using patient-specific structured ontologies, such as the HPO or the Research Domain Criteria (RDoC) (Cuthbert and Insel, 2013), specifically developed for mental disorders.

Our study presents several limitations worth mentioning. Firstly, the study samples evaluated are relatively small. Indeed, the Nijmegen datasets comprise less than 500 patients, making our results inherently reliant on the phenotypic resolution of the cohort. While the successful replication of our results using DECIPHER patients supports the validity of genotype and phenotype convergence in disease, the phenotypic inconsistencies of certain modules could inevitably be associated with insufficient statistical power and lack of systematic phenotypic in DECIPHER rather than lack of biological meaning.

Secondly, Nijmegen patients possess less than 20% of existing HPO terms, suggesting the cohorts capture only a small proportion of global human phenotypic variation. Indeed, this study does not extend beyond the examination of neurodevelopmental disorders. Furthermore, the utilised ontologies and observed phenotypes do not encompass the full spectrum of

the disorders evaluated, which could be enriched using additional neurobiological, brain imaging and quantitative behavioural data. These limitations are of particular relevance to SSC cohort, which does not provide information on a set of pre-determined clinically relevant abnormalities. Instead, the extensive phenotyping of patients and their families stipulates the manual choice of variables to examine, which is inherently biased. While our phenotypic selection process was subject to a standardised approach aiming to maximise the number of ASD-associated comorbidities and problems considered (**section S3**), we must acknowledge some shortcomings. Firstly, many correlates, such as sleep abnormalities, were not assessed. Secondly, distinct aspects of ASD may have contributed unequally to our analysis, as a result of the collection of features evaluated and the disparate number of variables relating to them. We, hence, highlight the need for systematic comparable phenotyping of patient cohorts using clinically defined sets of biologically relevant and standardised ontologies (e.g. HPO or RDoC). Finally, we note that the phenotyping process itself is confounded by age, time of assessment and environmental components, such that the inclusion of longitudinal data and information regarding therapies and medical treatments could provide more comprehensive overviews of patients' clinical profiles.

Thirdly, our analysis focuses on rare structural variants and neglects the contribution of non-coding and common variation. Their exclusion is arguably justified by the difficulty of interpreting variants arising outside of DNA coding sequences and the inconsistent results generated by common variants in NDDs, respectively. Nevertheless, recent estimates propose that common

variants underlie up to 50% of ASD risk, such that stronger genetic correlations are observed among patients displaying ASDs of similar severity (Veatch *et al.*, 2014; Gaugler *et al.*, 2014). This suggests common variation may influence the development of variably expressive forms of disease. Finally, our analysis additionally disregards the role of rare *dn* gene point mutations, which are strongly associated with NDDs, such that the study of specific gene variants has led to the identification of well-characterised clinical subtypes of disease (Stessman *et al.*, 2016). As inherited variants reveal lower rates of pathogenicity and only a small proportion of SSC probands carry *dn* CNVs, the overall failure to detect strong genotype to phenotype relationships in ASD may be a consequence of the proportion of phenotypic variability due to inherited structural variation. Indeed, the simplex design of SSC families selects against causal inherited variation, such that *dn* exonic variants are likely to exert substantial effects on disease risk.

Despite the discussed shortcomings, our findings provide supporting evidence for the paradigm that the perturbation of common biological pathways underlies the onset of analogous phenotypes. Furthermore, we highlight the importance of large-scale systematically genotyped and phenotyped cohorts in shaping disease landscapes, which can, in turn, be used to develop effective and personalised treatments targeted to converging patient groups.

4.6 Conclusion

This work provides a systematic investigation of genotype and phenotype relationships in NDDs using genotyped and rigorously phenotyped patients cohorts. Paradoxically, we observed phenotypic coherence among patients with similar functional disruptions, although no particular abnormality could be linked to the associated molecular processes. Instead, biological pathways underlay phenotypic spectra, such that the assessment of large numbers of individuals is necessary to shape the phenotypic landscape of the associated disruptions. Altogether, this work suggests disease aetiologies could be inferred based on the phenotypic similarity between patients, further paving the road towards prognostic predictions and personalised medicine.

CHAPTER 5

Genotype and phenotype convergence:

an exome study

5.1 Introduction

Autism Spectrum Disorder (ASD) is one of the most common NeuroDevelopmental Disorders (NDDs) and affects one child in 68 in the United States (DMNSY and CDC, 2014). The disorder is characterised by abnormalities in social interaction and communication, as well as restricted and repetitive behaviours.

ASD unveils a tremendously broad and heterogeneous phenotypic spectrum, wherein the qualitative impairments of diagnostic features are highly variable and often accompanied by comorbid disorders. These encompass Intellectual Disability (ID), epilepsy, Attention Deficit and Hyperactivity Disorder (ADHD), obsessive compulsive disorder (OCD) and anxiety disorders, which occur in between 10% to 80% of cases (Jeste, 2011; Mannion and Leader, 2013; Mannion *et al.*, 2014). Furthermore, ASD commonly manifests with additional medical problems, including gastro-intestinal and sleep disturbances, which affect up to 90% of patients (Jeste, 2011; Mannion and Leader, 2013). Epidemiological, family and twin studies implicate an underlying genetic aetiology, such that all forms of genetic variation (i.e. rare, common, inherited, *de novo*, locus, gene and structural variation) contribute to its architecture, uncovering an equally striking genetic heterogeneity.

Over the past decade, large-scale genetic screens of ASD patients have established a pivotal role for rare (i.e. present in less than 1% of the population), and in particular *de novo* (*dn*) germline variation. Genetic screens

led to the discovery of many high-risk gene mutations, termed single nucleotide variants (SNVs), and structural variants involving the deletion or duplication of DNA segments > 1 kb (termed Copy Number Variants; CNVs) (**Tables 1.1 and 1.2**). Functional evaluation of the large number of *dn* variants uncovered through these hypothesis-free screens consistently implicated processes of chromatin remodelling and synaptic functions in the aetiology of disease (Iossifov *et al.*, 2012; O'Roak *et al.*, 2012; Ben-David and Shifman, 2013; Steinberg and Webber, 2013; Cristino *et al.*, 2014; De Rubeis *et al.*, 2014; Iossifov *et al.*, 2014; Krumm *et al.*, 2014; Pinto *et al.*, 2014; Toma *et al.*, 2014; Chang *et al.*, 2015; Hormozdiari *et al.*, 2015; Mahfouz *et al.*, 2015; Sanders *et al.*, 2015; Krishnan *et al.*, 2016; Li *et al.*, 2016; Shohat *et al.*, 2017; Stessman *et al.,* 2017; Vorstman *et al.*, 2017; Yuen *et al.*, 2017). Nevertheless, genes mapping to distinct molecular processes, including immunity and inflammation, have also been identified (Garbett *et al.*, 2008; Voineagu *et al.*, 2011; Gupta *et al.*, 2014).

While the bulk of ASD research aims to elucidate the biological pathways globally affected in disease, the compelling phenotypic heterogeneity suggests disparate molecular processes could influence variably expressive forms of disease. This view builds on the postulation that ASD is an umbrella term for a large number of conditions with distinct genetic aetiologies (Percy, 1990). Consequentially, it advises that individuals with genetic disruptions that converge on the same biological processes might present with similar clinical profiles.

Studies of genotype and phenotype relationships in ASD have primarily focused on the examination of genetic burden. Accordingly, the presence of *dn* SNVs and CNVs has been linked to more severe phenotypes, accompanied by ID (or simply lower Intellectual Quotients; IQ) and seizures (Sanders *et al.*, 2011; Iossifov *et al.*, 2014; Merikangas *et al.*, 2014; Robinson *et al.*, 2014; Samocha *et al.*, 2014; Sanders *et al.*, 2015; Stessman *et al.*, 2017; Weiner *et al.*, 2017). Recently, Merikangas *et al.* investigated the phenotypic manifestations of several types of CNVs and revealed that CNVs affecting ID/ASD and brain genes influence language ability and adaptive function, respectively (Merikangas *et al.*, 2014). While these findings remain to be replicated, they suggest different types of functional disruptions may modulate the autistic phenotype. To investigate this hypothesis, we systematically examined the phenotypic coherence of *dn* and inherited CNVs affecting functionally similar and co-expressed genes across a wide range of ASD comorbidities and correlates (**chapter 4**). Nevertheless, the assessment of *dn* CNVs was statistically underpowered and that of inherited CNVs did not generate strong positive findings, possibly owing to lack of biological signal. Indeed, the study made use of the Simons Simplex Collection (SSC) (Fischbach and Lord, 2010), which is designed to be depleted of causal inherited variation, as a result of the lack of ASD among family members of each patient. Instead, the SSC provides a powerful cohort to investigate the phenotypic manifestations of *dn* SNVs, given the availability of Whole-Exome Sequencing (WES) and extensive clinical data.

While the role of rare and *dn* variants in ASD is well-established, the contribution of Common Variants (CV), such as base substitutions termed Single Nucleotide Polymorphisms (SNPs) is less substantiated (n.b. SNPs, themselves, are not necessarily causal and often represent markers of CV in trait-associated regions). Genome-Wide Association Studies (GWAS), aiming to identify SNPs more commonly observed in cases compared to controls, failed to detect or replicate loci predisposing to ASD or ASD subtypes (e.g. severe ASD, ASD with regression or ASD with severe restricted and repetitive behaviours) (Liu *et al.*, 2008; Ma *et al.*, 2009; Anney *et al.*, 2010; Anney *et al.*, 2012; Connolly *et al.*, 2013; Chaste *et al.*, 2015; Kuo *et al.*, 2015; Liu *et al.*, 2016a; Mitra *et al.*, 2016; Tao *et al.*, 2016; Cantor *et al.*, 2017). Furthermore, a recent study found no pronounced association between ASD and SNPs affecting 32 functional groups of genes associated with NDDs (Jansen *et al.*, 2017). Nevertheless, a signal was observed for genes regulated by the Fragile X Mental Retardation Protein (FMRP) (Steinberg and Webber, 2013; Jansen *et al.*, 2017), supporting a role for common variation in ASD. Moreover, polygenic risk for ASD, schizophrenia and educational attainment (burden of SNPs linked to these conditions through GWAS) has been associated ASD risk, independently of gender or ID comorbidity (Weiner *et al.*, 2017). In line with these findings, recent heritability studies revealed a substantial portion of the genetic variance in ASD derives from common variation (Anney *et al.*, 2012; Klei *et al.*, 2012; Gaugler *et al.*, 2014), suggesting the additive effects of large numbers of SNPs are likely to contribute to disease. This is further corroborated by the observation that

individuals presenting with ASD of comparable severities (i.e. mild or severe) are more genetically similar than expected by chance (Veatch *et al.*, 2014).

One intriguing hypothesis is that common variation acts as a modifier of rare *dn* penetrant mutations, hence influencing the onset of variably expressive forms of disease. This scenario is typically exemplified using the Bardet-Biedl syndrome (Badano *et al.*, 2006), wherein CVs exert negligible individual effects, but modulate phenotypic variability in presence of rare causal mutations. The potency of such an integrative model in ASD is supported by the observation that common duplication affecting the *SLC2A3* gene act as genetic modifiers of the abnormalities displayed by patients with the 22q11.2 microdeletion syndrome (i.e. Velo-Cardio-Facial syndrome) (Mlynarski *et al.*, 2015). Furthermore, family background, which is determined by common variation, contributes to the clinical variability of ASD patients carrying *dn* 16p11.2 duplications (Moreno-de-Luca *et al.*, 2015). The modifying potential of CVs over a broader ASD spectrum remains, however, to be determined.

5.2 Chapter aims

In this chapter, we sought to expand our previous analysis of genotype and phenotype relationships in disease (**chapter 4**), by examining the clinical outcomes of functionally converging *dn* SNVs in ASD. Aiming to disentangle the heterogeneity of ASD, we applied the same study pipeline as in **chapter 4**. Specifically, we (1) grouped genes affected by *dn* SNVs in SSC individuals based on their functional characteristics, then forming the associated patient groups, whom possess variants in functionally related or co-expressed genes. We, then, (2) examined the phenotypic coherence of each group and (3) asked whether phenotypically similar gene modules underlie phenotypic spectra, rather than specific abnormalities, in order to reproduce our previous results. Finally, (4) we evaluated the role of CVs in modulating disease expressivity, and assessed whether they act as genetic modifiers of penetrant rare *dn* mutations.

5.3 Specific materials and methods

5.3.1. Gene Lists

We collated sets of genes belonging to gene classes associated with NDDs. These include confident ASD genes, brain-specific genes, synaptosome genes, postsynaptic density genes, FMRP target genes, chromatin remodelling genes, histone modifying genes and transcription factors (**section 2.1; Table 2.1**). Additionally, we extracted 20 genes involved in disorders of the Ras/MAPK pathway (i.e. RASopathies) from Mitra *et al.*, 2017. Moreover, we collated a list of 51 genes participating in the mammalian Target Of Rapamycin (mTOR) signalling pathway from Laplante and Sabatini, 2012.

5.3.2 SSC cohort: phenotypic data

The phenotypic patient dataset constructed in **chapter 4** and encompassing 63 clinical attributes (**Table 4.1**) was used in downstream analyses (**sections 4.3.3 and S3**).

5.3.3 SSC cohort: rare *de novo* exonic variants

All rare *de novo* (*dn*) SNVs identified in 2,508 ASD patients from the SSC were downloaded from Iossifov *et al*., 2014. We restricted the analysis to Likely Gene-Disrupting (LGD) variants, namely (1) Loss of Function (LoF) mutations – nonsense, frame-shift and splice-site variants – and (2) MISsense

mutations predicted to be deleterious (delMIS) by the CADD scoring system (with C-scores ≥ 15) (Kircher *et al*., 2014). We removed 13 delMIS and seven LoF mutations, which arose in probands and their respective sibling(s), speculating these mutations are not causal due to the lack of disease phenotype in the sibling. Probands were associated with genes, if these were affected by the aforementioned mutations. The final dataset consisted 2,498 patients, of which 1,177 display 1,625 *dn* LGD SNVs. Specifically, 444 LoF and 1,181 delMIS mutations occur in 401 and 937 patients, respectively. On average, patients display 0.18 LoF and 0.47 delMIS mutations.

5.3.4 SSC cohort: common variants

5.3.4.1 Samples and genotyping

Genotyping of individuals from SSC families was performed as described in Sanders *et al.*, 2015. Three genotyping arrays were used and individuals from the same families were analysed on the same array. Specifically, 335, 1,193 and 1,077 families were assessed using the Illumina 1Mv1, 1Mv3 and Omni2.5 arrays, respectively. Genotyping data was downloaded as PLINK files (Purcell *et al.*, 2007) from SFARI Base. Quality Control (QC) was performed separately for individuals genotyped on each array in PLINK (Purcell *et al.*, 2007), following the protocol described in Anderson *et al.*, 2010. Due to possible confounding effects, we excluded sex chromosomes from the study (Clayton, 2009). We re-mapped genotypes from the hg18 to the hg19 genome assembly using the UCSC LiftOver tool (Kent *et al.*, 2002).

Of note, while all samples were subject to QC, we subsequently restricted the analysis to the subset of individuals present in our SSC study sample (n = 2,148).

5.3.4.2 Per individual quality control

To circumvent the introduction of false positives and noise, we estimated the same metrics as described in **section 3.2.4**, using PLINK (Purcell *et al.*, 2007). We removed (1) individuals with discordant sex information, as given by F-statistic values between 0.25 and 0.75; (2) individuals with genotyping success rates < 98% and abnormal heterozygosity rates, defined as ± 2 standard deviations from the mean; and (3) duplicated or related subjects with *Identity By Descent* (IBD) values ≥ 0.185. In total, 474 samples failed QC, leaving 9,746 individuals for downstream analyses, of which 2,486 were cases. The number of samples failing QC at each stage is presented in **Table 5.1**.

Table 5.1 | Per sample quality control.

	1Mv1		1Mv3		Omni2.5	
	Case	Control	Case	Control	Case	Control
Total	335	1,019	1,193	3,433	1,077	3,163
Discordant sex	0	9	2	32	0	8
Genotyping call rate < 98% or heterozygosity ± 2 sd	10	37	54	147	53	121
Duplicated and/or related individuals	0	0	0	0	0	0
Extremely divergent ancestry (PC 2 < 0.05)	38	96	147	420	154	453
Total remaining	287	882	991	2,854	870	2,582

sd = standard deviation

Number of SSC subjects removed at each QC step.

To control for population stratification, we performed a principal component analysis of SSC samples and examined the ancestry of SSC samples compared to HapMap3 reference samples (International HapMap Consortium, 2003; International HapMap Consortium, 2005). We observed a large number of individuals deviating from the principal European CEU HapMap cluster (**Fig 5.1**). Applying stringent value thresholds for Principal Component (PC) 2 leads to the exclusion of close to 50% of our study sample (**Fig 5.1**). We, therefore, applied a more permissive threshold and excluded 338 cases and 970 controls with extremely divergent ancestries, as defined by PC 2 < 0.05 (**Fig 5.1; Table 5.1**). While other studies employing the SSC (Sanders *et al.*, 2011; Chaste *et al.*, 2015; Sanders *et al.*, 2015; Mitra *et al.*, 2016) did not filter individuals based on ancestry, we deemed it a necessary QC step to circumvent the introduction of confounding effects in our genetic correlation analysis (**section 5.3.4**).

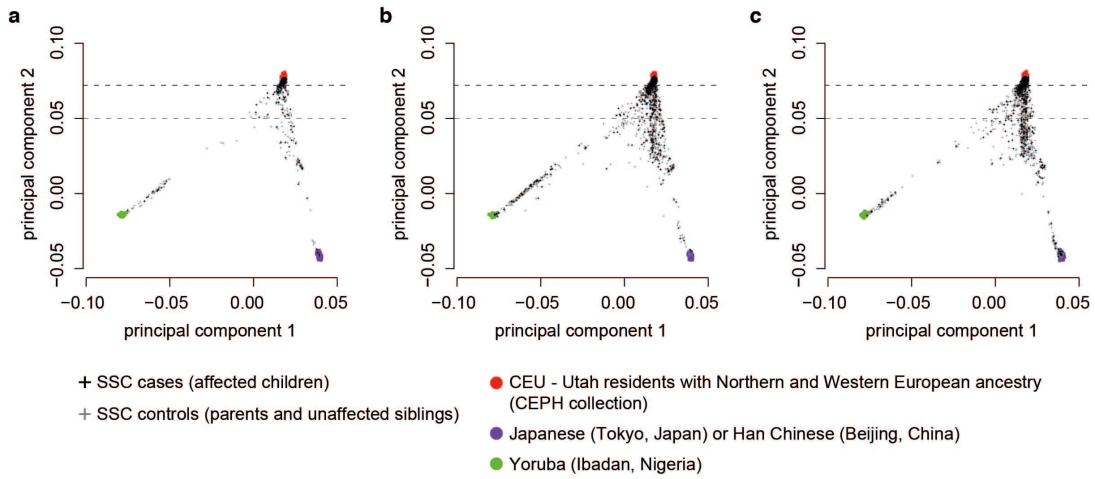


Fig 5.1 | Ancestry clustering. Principal component analysis of genotype data from SSC individuals examined on the (**a**) 1Mv1, (**b**) 1Mv3 and (**c**) Omni2.5 platforms. The reference panel consists of HapMap3 samples (International HapMap Consortium, 2003; International HapMap Consortium, 2005). Individuals lying outside of the primary European cluster are removed, as given by principal component 2 < 0.05.

We separated subjects into a pre-determined number of ancestry groups using a k-means algorithm, implemented in R. This algorithm assigns subjects to clusters in order to maximise the Euclidean distance between individuals from different groups, based on various scores (i.e. PCs). The first two PCs relate to ancestry and capture an appreciable proportion of the total variance (based on a scree plot; not shown). Using the two first PCs, we constructed four ancestry groups, as suggested by the Calinski-Harabasz and Duda-Hart criterions (Milligan and Cooper, 1985) and a scoring index developed by Solovieff *et al.*, 2010. We, then, assessed whether patients from modules (**section 5.3.7**) were enriched for particular ancestry groups compared to remaining patients, using one-sided right-tailed Fisher tests. A correction for multiple testing was applied using the Benjamini-Hochberg (BH) method to account for the number of modules tested for each array and ancestry group. This analysis was performed separately for each genotyping platform and was restricted to individuals genotyped on the associated array (only modules including at least 5 patients interrogated on the assessed array were considered). No modules comprised an excess of individuals of particular ancestries following BH correction, independently of the genotyping platform. Nominally, Brainspan module 2 is enriched for patients of Yoruban (Nigerian) ancestry when considering the Omni2.5 platform ($p = 0.040$, BH $p = 0.20$; right-tailed Fisher test) but not the 1Mv1 and 1Mv3 arrays. Of note, these individuals are removed from downstream analysis during filtering of extremely divergent individuals (**Table 5.1**). No association between functional gene modules and ancestry is observed when repeating the

analysis on the imputed dataset (p > 0.1; right-tailed Fisher tests) (**section 5.3.3.4**). We, thus, concluded our analysis is not confounded by ancestry.

5.3.4.3 Per SNP quality control

Following the protocol described by Anderson *et al.*, 2010, we removed markers able generate false positive data (see **section 3.2.5** for details concerning QC measures). Specifically, we removed SNPs with (1) minor allele frequencies < 0.01, (2) genotyping rates of < 98%, (3) significant deviations from the Hardy-Weinberg equilibrium at p < 1 x 10-5 and (4) markers mapping to sex chromosome. We, additionally, removed (5) SNPs with significantly different genotyping call rates between cases and controls (i.e. parents and unaffected siblings), at p < 1 x 10-5, as allele frequency differences are likely to be due to genotyping errors. In total 844,515 1Mv1, 926,140 1Mv3 and 1,561,861 Omni2.5 SNPs remained for analysis. The number of SNPs failing QC at each stage is presented in **Table 5.2**.

Table 5.2 | Per SNP quality control.

	1Mv1	1Mv3	Omni2.5
Total	1,072,814	1,199,033	2,440,283
Minor allele frequency < 0.01	172,049	186,808	831,935
Significantly different genotyping call rate between cases and control (p < 1 x 10-5)	1,075	1,681	4,003
Genotyping call rates < 98%	30,300	56,148	11,681
Significant deviation from Hardy-Weinberg equilibrium (p < 1 x 10-5)	5,276	12,634	8,516
SNPs on X chromsome	40,097	45,590	54,354
SNPs on Y chromosome	2,283	4,637	1,897
Total remaining	844,515	926,140	1,561,861

Number of genotyped makers removed at each QC stage.

5.3.4.4 Genotype imputations

Using the reference hg19 genome from UCSC (Kent *et al.*, 2002), we aligned all SNPs from the 1Mv1, 1Mv3 and Omni2.5 genotype datasets to the positive strand. We removed four SNPs for which the reference allele did not match that of the reference genome. PLINK files were converted to VCF format and genotypes were pre-phased using EAGLE2 (Loh *et al.*, 2016) and imputed using the Sanger Imputation Service Beta (McCarthy *et al.*, 2016). The 1000 Genome Phase III (v5) integrated variant set was used as the reference panel for imputations (1000 Genomes Project Consortium, 2015).

We removed (1) low quality SNPs with high uncertainty of imputation as given by the INFO metric < 0.3 and (2) multiallelic or duplicate SNPs. Additionally, we excluded SNPs with (3) with minor allele frequencies < 0.01, (3) excessive missing genotyping call rates (i.e. > 2%) and (4) significant deviations from the Hardy-Weinberg equilibrium (i.e. $p < 1 \times 10^{-5}$). We, then, merged the 1Mv1, 1Mv3 and Omni2.5 imputed genotype datasets, by considering the 8,201,975 overlapping SNPs.

5.3.4.4 Genetic correlation

Considering the 2,148 patients for which genotype data was available post-QC, we estimated the genetic similarity of pairwise combinations of patients as the proportion of alleles they share relative to the total number of markers

interrogated. Genetic similarity was calculated in PLINK (Purcell *et al.*, 2007) using the *--genome* command, which evaluates *Identity By State* (IBS) as:

$$Genetic\ similarity = \frac{IBS(2) + 0.5 \times IBS(1)}{IBS(0) + IBS(1) + IBS(2)}$$

, where *IBS (2)*, *(1)* and *(0)* refer to the number of loci, at which the two individuals share two, one or no alleles, respectively.

5.3.5 Expression quantitative trait loci

Genotype and RNA sequencing data across 17 tissues was downloaded from the Database of Genotypes and Phenotypes (dbGaP) (Tryka *et al.*, 2014) for 449 donors from the Genotype-Tissue Expression (GTEx) project (Lonsdale *et al*., 2013).

Genotype data includes 11,555,102 genotyped and imputed variants, filtered according to the GTEx release v6 documentation. Only variants with genotyping call rates > 95%, INFO scores > 0.4 and tissue-specific minor allele frequencies ≥ 1% were maintained.

RNA sequencing data was normalised following the protocol described in the GTEx release v6 documentation. We considered protein-coding genes with Reads Per Kilobase of transcript per Million mapped reads (RPKM) > 0.1 in at least 10 individuals. For each tissue, gene expression values were subject to quantile normalisation (i.e. expression values were quantile normalised to the

average empirical distribution observed across samples and gene expression values were inverse quantile normalized to a standard normal distribution across samples).

Covariates consisting of three genotyping PCs, gender and genotyping array were considered. Additionally, 15 factors, identified using the Probabilistic Estimation of Expression Residuals (PEER) method (Stegle *et al*., 2010) and computed on normalised tissue-specific gene expression levels, were included.

A comprehensive search for expression Quantitative Trait Loci (eQTLs) was conducted using the Matrix eQTL R package (Shabalin, 2012). Linear regression models, including the aforementioned covariates, were applied to all gene-SNP pairs and significance of association was derived using the t-statistic. False discovery rates were estimated using the BH procedure. Both local (i.e. *cis*) and distal (i.e. *trans*) eQTLs were detected. *Cis* eQTL windows were defined as ± 1 mb around genes' transcript start site, as described in the GTEx release v6 documentation. The analysis was performed using all brain tissues (we investigated brain eQTLs only). This analysis was performed by Viola Volpato, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, UK.

5.3.6 SNP and eQTL sets

Genetic correlations between SSC patients were estimated both across all variants from the merged imputed dataset and across 95 SNPs nominally linked to ASD through GWAS (Vardarajan *et al.*, 2013; Chaste *et al.*, 2015; Warrier *et al.*, 2015; Mitra *et al.*, 2016).

We considered gene modules and several gene sets associated with NDDs (**section 5.3.1**). For a given module or gene set, we measured the genetic correlations of SSC patients (**section 5.3.3.4**) across genotyped and imputed eQTLs acting in *cis* or *trans* to the respective genes. Additionally, we calculated genetic distances across SNPs located within a 5 kb (Mitra *et al.*, 2016; Mitra *et al.*, 2017) and 10 kb flanking region of each gene from the evaluated module or set. The modules and gene sets considered and numbers of associated variants are displayed in **Table 5.3**.

5.3.7 Genotype and phenotype analysis pipeline

The analysis performed in this chapter follows the same methodology and implements the same pipeline as in **chapter 4**. Specifically, (1) genes disrupted by *dn* LGD SNVs in SSC patients were partitioned into modules, based on their functional similarity or similarity of expression patterns, using the Louvain algorithm (**section 4.3.5**). For this, we used the 25% strongest links from our Phenotypic Linkage Network (PLN), and the GTEx and Brainspan co-expression networks with r > 0.6 and r > 0.8, respectively (**Fig**

S5.1 ; Table 2.3). Patients were assigned to modules, based on their gene disruptions. We, then, (2) evaluated whether patients from each module converge at the phenotypic level (**section 4.3.6**), (3) identified the abnormalities driving this similarity (**section 4.3.8**) and (4) examined their implications (**section 4.3.9**).

Table 5.3 | SNP and eQTL sets.

Gene set	Number of genes	SNPs in gene flanking region		Brain eQTLs	
		5 kb	10 kb	*Cis*	*Trans*
GTEx module 1	191	5489	6145	4114	40961
GTEx module 2	52	1656	1818	760	6308
GTEx module 3	190	6097	6712	3861	32782
GTEx module 11	88	2467	2785	1918	25306
Brainspan module 0	94	3943	4238	1630	13035
Brainspan module 1	199	4797	5444	4078	43410
Brainspan module 2	281	15954	16945	5297	26293
PLN module 3	126	2723	3107	2953	35543
PLN module 5	95	2321	2634	1828	32966
PLN module 9	23	821	889	541	9335
ASD genes (SFARI Gene score 1)	17	897	964	203	787
ASD genes (SFARI Gene score 1 or 2)	56	6188	6410	787	4594
ASD genes (SFARI Gene score 1, 2 or 3)	193	19544	20234	3242	25522
Brain-specific genes	381	20104	21396	4965	28435
Chromatin remodelling genes	374	8816	9830	6267	54554
FMRP target genes	860	46203	48821	13610	67393
Histone modifying genes	157	4060	4458	2484	18702
mTOR signalling genes	51	2318	2475	1048	7704
Postsynaptic density genes	1167	48436	51627	17639	121745
RASopathy genes	20	572	662	264	2933
Synaptosome genes	161	5764	6215	2766	39876
Transcription factors	3245	61993	69670	42585	200285

5.3.8 Genetic correlation comparisons

We considered the phenotypic similarity of SSC patients, estimated using the Goodall3 index (**section 2.6**) across 63 clinical variables (**Table 4.1**). For each module, we examined the relationship between phenotypic patient similarity and genetic similarity using linear regression models in R. Significant

models were defined as those associated with a significant p-value, adjusted for the number of modules tested using the BH method (BH $p \leq 0.05$), and a positive slope.

For each module, we constructed two types of negative controls:

Firstly (**a**), we sought to assess the likelihood of observing significant genotype-phenotype correlations by chance. For each module, we ran 1,000 permutations, during which we permuted patient phenotypic similarities within the module. During each permutation, we re-assessed the correlation between phenotypic and genetic similarity using linear regression in R. Empirical p-values were derived, reflecting the proportion of permutations associated with nominally significant regression models and positive slopes.

Secondly (**b**), we sought to evaluate whether the observed genotype-phenotype correlation is specific to the modules examined. Specifically, we assessed whether particular sets of CVs globally modulate the autistic phenotype, or whether they specifically modify the phenotypic spectrum associated with the module examined (i.e. specifically act as genetic modifier of genes comprised in the modules assessed). For each module, we constructed 1,000 randomised modules by randomly selecting the same number of SSC patients with genotype data (n = 2,148 individuals). During each randomisation, we re-examined the correlation between phenotypic and genetic similarities using linear regression in R. Empirical p-values were derived, reflecting the proportion of randomisations, during which randomised

modules were associated with nominally significant regression models with positive slopes.

A role for CVs in modulating the expressivity of disease is supported by (1) BH-adjusted significant regression models with positive slopes and (2) significant empirical p-values ($p \leq 0.05$), derived from permuting genetic similarities within modules (test **a**). Moreover, (3) significant empirical p-values ($p \leq 0.05$), derived from randomising module patients (test **b**), suggest CVs specifically act as genetic modifiers of rare penetrant mutations affecting module genes.

5.3.9 Centre of assessment bias

We sought to ensure the patient examination centre does not confound the observed genotype and phenotype coherence, given the biases in clinical impressions across sites (Fischbach and Lord, 2010). As data regarding assessment centres is not available within the SSC, we were unable to verify that modules do not comprise an excess of individuals inspected at the same clinic.

Reasoning that (1) patients were examined at sites closest to their homes and that (2) genetic and geographical distances are correlated, we compared the pairwise genetic similarities of module patients to that of 1,000 randomised modules, across SNPs from the imputed and merged dataset. We derived empirical p-values, reflecting the proportion of randomisations, during which

permuted patients were more similar than module patients. As module patients were not genetically clustered ($p > 0.24$), we concluded the assessment centre does not confound phenotypically coherent modules. This is further supported by the observation that genetic and phenotypic similarities are not globally correlated ($p = 0.38$; linear regression).

5.4 Results

Expanding our phenotypic analysis beyond the study of CNVs (**chapter 4**), we employed the SSC cohort to investigate the phenotypic manifestations of functionally converging, likely damaging *de novo* (*dn*) gene disruptions.

5.4.1 Summary of data

WES of SSC patients revealed they harbour on average 0.65 ± 0.82 Likely Gene-Disrupting variants (LGD; loss of function variants or missense variants predicted to be deleterious; **section 5.3.3**), 0.18 ± 0.43 Loss of Function variants (LoF; nonsense, splice-site or frameshift mutations) and 0.47 ± 0.69 MISsense variants likely to be deleterious (delMIS). LGD, LoF and delMIS variants affect a total of 1,436, 421 and 1,075 distinct genes, respectively (**Fig 5.2**). The incidence of evaluated phenotypes is similar between the whole cohort and the subset of 1,177 patients with ≥ 1 *dn* LGD SNV(s), with patients displaying a median of 5 phenotypes each ($p = 0.75$; two-sided Mann-Whitney U test). The most common abnormalities are displayed in **Fig S5.2**.

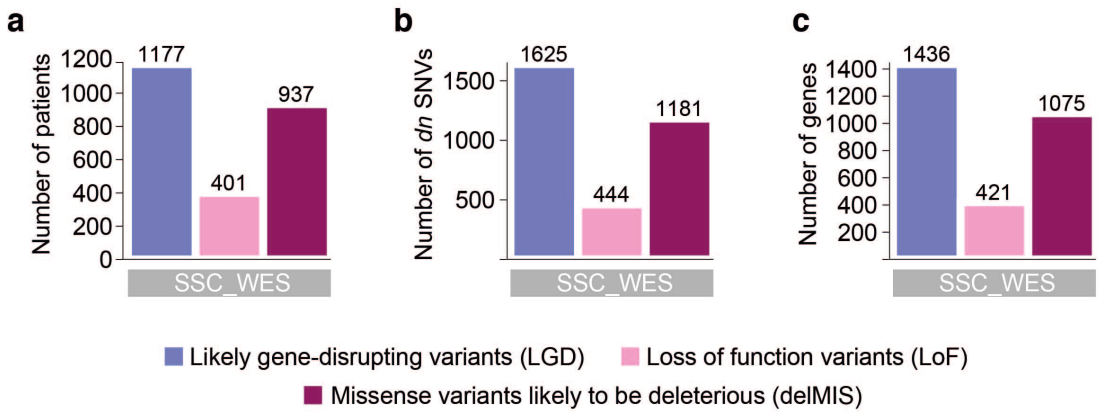


Fig 5.2 | Burden of *de novo* exonic events in the SSC. The barplots depict (**a**) the number of SSC patients carrying *de novo* (*dn*) mutations, (**b**) the total number of variants detected and (**c**) the number of distinct genes they affect. Mutational events are stratified by type, namely likely gene-disrupting variants (LGD; blue), loss of function variants (LoF; pink) and missense variants likely to be deleterious (delMIS; purple). SSC_WES, cohort of sequenced SSC patients (n = 2,498).

5.4.2 Phenotypic coherence of single gene variants

We, firstly, examined the clinical outcomes of individuals with disruptions in the ASD-associated genes: *CHD8* (i.e. seven patients with LoF variants) and *SCN2A* (i.e. six patients with delMIS variants). While *SCN2A* is a syndromic ASD/epilepsy gene, *CHD8* represents the strongest ASD candidate gene to date and has been linked to ASD with accompanying macrocephaly and gastro-intestinal problems. Nevertheless, gastro-intestinal problems and epilepsy are not over-represented among SSC patients with mutations in *CHD8* and *SCN2A*, respectively. In fact, mutational carriers are not enriched for any of the comorbidities examined ($p > 0.087$; right-tailed Fisher tests) (n.b. macrocephaly was not included in the analysis). This suggests the analysis is either underpowered due to low sample size or the association between *CHD8* or *SCN2A* and their corresponding phenotypes is weak.

Furthermore, affected patients are not phenotypically more similar than expected by chance (**Fig 5.3b**).

5.4.3 Modules of co-expressed and functionally similar genes

Using the Louvain algorithm, we partitioned genes disrupted by *dn* LGD SNVs in SSC patients into groups of genes with similar (1) body-wide or (2) spatiotemporal brain expression patterns, based on GTEx and Brainspan co-expression networks, respectively. Additionally, (3) we derived modules of functionally similar genes using the PLN. Specifically, we identified 12 GTEx, 26 Brainspan and 21 PLN modules. We constructed the associated patient modules and applied our phenotypic pipeline to 9 GTEx, 5 Brainspan and 11 PLN modules comprising between 17-235, 14-299 and 24-219 individuals, respectively. As most patients display a single variant (i.e. 78% of patients with *dn* LGD SNVs), they are depleted of multiple module memberships, hence facilitating direct associations between genotype and phenotype, if present.

By performing functional enrichment analyses, using annotations from the Gene Ontology (GO) and Mouse Genome Informatics (MGI) databases, we characterised gene modules at the functional level (**Tables S5.1 to S5.6**). GTEx modules encompass a wide range of functions:

- Module 0 (215 genes): nervous system physiology, including synaptic transmission, such that mouse KnockOut (KO) models display synaptic and behavioural abnormalities.
- Module 1 (191 genes): cell cycle and response to DNA damage leading to partial or complete lethality in mouse models.
- Module 4 (82 genes): muscle contraction causing muscle morphology and contractibility abnormalities in mouse models.
- Module 5 (193 genes): regulation of gene expression through chromatin modification, inducing lethality or growth retardation in mouse KOs.
- Module 6 (122 genes): cell adhesion and morphogenesis, leading to cardiovascular development abnormalities in mouse models.

Of note, we are unable to assign GTEx modules 2 (52 genes), 3 (190 genes), 10 (17 genes) and 11 (82 genes) with specific functionalities, given the lack of interpretable GO/MGI enrichments. Although functionally similar modules were identified using the PLN (**Fig S5.3d**), we observed a better functional delineation in PLN modules, possibly owing to the inclusion of annotation datasets, such GO, KEGG and Reactome, in PLN construction. Specifically, PLN modules involve:

- Module 13 (173 genes): gene expression regulation via chromatin modification
- Module 8 (135 genes): synaptic transmission
- Module 3 (126 genes): RNA processing
- Module 4 (57 genes): cell cycle control

- Module 5 (95 genes): various metabolic processes
- Module 7 (149 genes): ubiquitination and Wnt signalling
- Module 0 (206 genes): cell adhesion
- Modules 2 and 6: transmembrane and ion transport specific the renal/urinary (module 2; 59 genes) and musculature (module 6; 88 genes) systems
- Modules 9 (23 genes) and 11 (60 genes): Golgi vesicle transport and other signal transduction pathways.

Conversely, Brainspan modules relate to two major themes, namely (1) neurodevelopment, through regulation of gene expression via chromatin modification. Such perturbations produce lethality and other embryonic phenotypes in mouse models, such as growth retardation (modules 0 and 3; 94 and 232 genes, respectively). Secondly, (2) a module (module 2; 281 genes) comprising genes involved in mature nervous system processes including synaptic transmission and plasticity was identified. Accordingly, disruption of gene orthologues in mouse models leads to abnormalities in brain and neuronal morphology and functions. These modules presumably arise from the separation of genes into those expressed prenatally and involved in brain development and synaptogenesis and those expressed postnatally and linked to mature brain functions, such as synaptic transmission and plasticity (**Fig S5.4**). The functionality of Brainspan modules 1 (199 genes) and 5 (14 genes) is unclear, given the lack of interpretable GO/MGI enrichments.

The uncovered modules reveal similar functions to those identified using *dn* CNVs from individuals with developmental delay and/or ID (**chapter 4**) (**Fig S5.3a-c**). This suggests different types of mutations target common gene classes and supports the notion of aetiological overlap across NDDs.

5.4.4 Phenotypic coherence of biologically related perturbations

Considering all sequenced SSC patients (n = 2,498), we found that patients with similar molecular disruptions, defined as common GTEx, Brainspan or PLN module memberships, are phenotypically more similar than individuals from different modules ($p < 1 \times 10\text{-}10$ for GTEx, Brainspan and the PLN; right-tailed Mann-Whitney tests) (**Fig S5.5**). In fact, the three networks perform comparably with 55.6% of GTEx (i.e. five of nine modules), 60% of Brainspan (i.e. three of five) and 63.6% of PLN (i.e. seven of eleven modules) modules showing significant phenotypic similarity (**Fig 5.3c, left panel**).

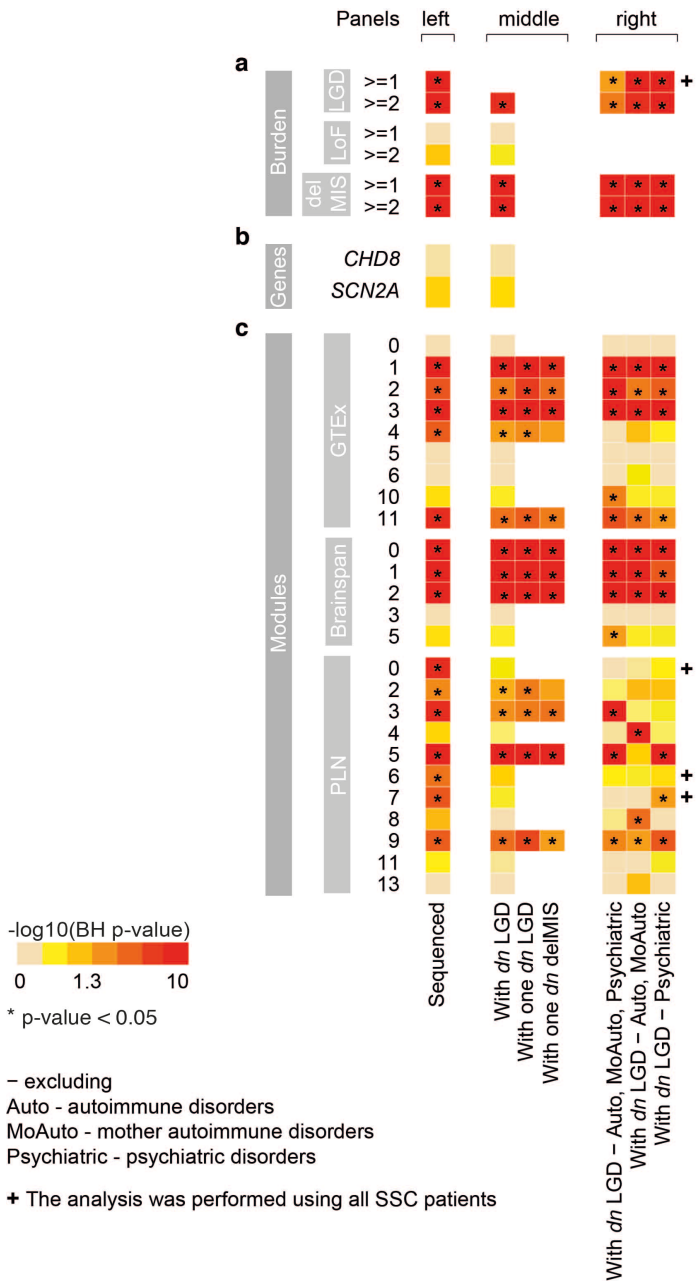


Fig 5.3 | Genotype and phenotype convergence. Phenotypic coherence of patients with (**a**) the same mutational burden, (**b**) the same gene mutations and (**c**) with similar functional disruptions, as given by common module memberships. Significance was derived by comparing pairwise phenotypic similarities of group patients and remaining patients using right-tailed Mann-Whitney U tests. The backgrounds used during comparisons consist of (**left panel**) all sequenced SSC patients (n = 2,498) and those carrying (**middle and right panels**) ≥ 1 LGD (n = 1,177) or (**middle panel**) (n = 812) one *dn* LGD or one *dn* delMIS (n = 599) SNVs. Phenotypic similarity was examined across (**left and middle panels**) all variables from the phenotypic dataset or (**right panel**) following the exclusion of phenotypes relating to patient autoimmune (Auto) and psychiatric (Psychiatric) disorders and to autoimmune disorders present in their respective mothers (MoAuto). LGD, likely gene-disrupting variants; delMIS, missense variants likely to be deleterious.

Patients with ≥ 1 *dn* LGD variant(s) are more phenotypically similar than those without (**Fig 5.3b, left panel**), suggesting our analysis could have been confounded by *dn* exonic burden. We, therefore, re-examined the phenotypic coherence of patient modules, when restricting the analysis to the subset of 1,177C SSC patients with ≥ 1 *dn* LGD disruption(s). Phenotypic similarity was replicated ($p < 1 \times 10\text{-}3$; one-sided Mann-Whitney tests), such that all but three PLN modules (i.e. modules 0, 6 and 7) retained their coherence signal (**Fig 5.3c, middle panel**). PLN modules 0, 6 and 7 were excluded from downstream analysis, due to confounding effects. Furthermore, as the similarity of individuals with ≥ 1 LGD SNVs is primarily due to those patients with ≥ 2 variants and, in particular, ≥ 2 delMIS variants (**Fig 5.3a, left panel**), we showed phenotypic coherence of patient modules persists when repeating the analysis using individuals with only one LGD or one delMIS variant (**Fig 5.3c, middle panel**). We, thus, conclude our results are not confounded by *dn* exonic burden and suggest functional gene convergence translates at the phenotypic level.

5.4.5 Autoimmunity and ASD

ASD severity does not drive the phenotypic similarity of patients with ≥ 1 *dn LGD disruption(s)*

While no study, to our knowledge, has systematically investigated the clinical profiles of patients with *dn* penetrant gene mutations, several observations suggest comorbidities are more frequent in these individuals. Firstly, LoF

variants have higher likelihoods of pathogenicity. Secondly, an excess of low functioning ASD and seizures is observed among patients with *dn* LoF variants (Robinson *et al.*, 2014; Samocha *et al.*, 2014; Weiner *et al.*, 2017). Lastly, low functioning individuals experience greater impairments compared to patients with higher IQs (Canne *et al.*, 2011; Cohen *et al.*, 2014; Robinson *et al.*, 2014). As a result, we hypothesised the phenotypic similarity of patients with ≥ 1 *dn* LGD variant(s) compared to patients without LGD disruptions was due to severer clinical profiles. However, while individuals with ≥ 1 *dn* LGD variant(s) are low functioning compared to patients without LGD disruptions ($p = 0.0097$, OR = 1.33; right-tailed Fisher test), they remain phenotypically similar following the exclusion of the low IQ variable from the analysis ($p = 2.32 \times 10\text{-}19$; right-tailed Mann-Whitney U test). Moreover, no other abnormalities are over-represented and no increase in comorbidity rate is observed among patients with ≥ 1 *dn* LGD variant(s) compared to individuals without LGD variants ($p = 0.72$; two sided Mann Whitney U test). We, thus, conclude other factors, beyond phenotypic severity, contribute to their apparent resemblance.

Absence of immune and psychiatric disorders drives the phenotypic similarity of patients with ≥ 1 *dn LGD disruption(s)*

Patients with ≥ 1 *dn* LGD variant(s) (and their mothers) are depleted of autoimmune and/or psychiatric disorders ($0.004 < p < 0.045$ and $0.7 < OR < 0.86$ for conditions separately and in combination; two sided Fisher tests). Hence, we hypothesised their phenotypic coherence derives from the shared

absence of these disorders. Agreeingly, we observed a strong reduction in the phenotypic similarity of individuals with ≥ 1 *dn* LGD variant(s) following the exclusion of patient and mother autoimmune and psychiatric phenotypes from the analysis (i.e. significance drops from p = 2.2 x 10-51 to p = 0.013; right-tailed Mann-Whitney U tests) (**Fig 5.3a, right panel**). Similarly, the phenotypic similarity signal observed for modules confounded by the presence of *dn* LGD variants (i.e. PLN modules 0, 6 and 7) disappears when patient and mother autoimmune phenotypes are excluded from the analysis (**Fig 5.3c, right panel**). These observations point to a negative correlation between the presence of *dn* LGD variants and that of psychiatric and autoimmune disorders (in patients and their mothers). We, hence, speculate patient and gestational immune disorders can lead to disease without an underlying causal genetic ASD mutation. This is supported by recent epidemiological research, which proposes autoimmune dysfunction represents a risk factor for ASD (Edmiston *et al.*, 2017). Conversely, in individuals without an environmental predisposition to ASD, genetic abnormalities, under the form of *dn* LGD variants, would cause disease. In fact, *dn* LGD from patients from PLN modules 0, 6 and 7 are associated with higher CADD deleteriousness scores compared to remaining *dn* LGD variants observed in the cohort (p = 4.06 x 10-6, p = 0.013 and p = 0.021, respectively; right-tailed Mann Whitnet U tests). This suggests the cprresponding patients are enriched for highly penetrant gene variants, which are presumably able to solely cause disease. Taken together, our findings advise ASD implicates both genetic and environmental (immune) aetiologies. Of note, in addition to patient and mother autoimmune disorders, psychiatric disorders also confound the genotype and

phenotype analyses. Nevertheless, this is likely to be a result of the co-occurrence of autoimmune and psychiatric conditions in patients ($p < 1 \times 10^{-10}$, OR > 3.1, right-tailed Fisher tests), in line with the view that autoimmune dysfunction predisposes to ASD and accompanying psychiatric conditions (Edmiston *et al.*, 2017).

Following the exclusion of autoimmune and psychiatric phenotypes from the analysis, we note patients with ≥ 1 or ≥ 2 *dn* LGD disruption(s) remain phenotypically similar compared to patients without or with one *dn* LGD disruption(s), respectively ($p = 0.013$ and $p = 1.14 \times 10^{-4}$, respectively; right-tailed Mann-Whitney U tests) (**Fig 5.3a, right panel**). This suggests additional factors, beyond the shared absence of autoimmune psychiatric phenotypes, underlie the phenotypic similarity of patients with ≥ 1 and ≥ 2 *dn* LGD disruption(s). This phenotypic coherence is likely to be related to the presence of low-functioning ASD, as removing the low IQ variable further diminishes abolishes the similarity signal when considering patients with ≥ 1 *dn* LGD disruption(s) (not shown). Moreover, albeit statistically insignificant, patients with ≥ 2 *dn* LGD variant(s) display more deleterious phenotypes than patients with ≥ 1 *dn* LGD variant(s) (median of 5 compared to 4 deleterious phenotypes, $p = 0.10$; right-tailed Mann-Whitney U test). This suggests stronger clinical impairments could further drive phenotypic similarity as a result of shared presences of phenotypes.

Phenotypic similarity of patients with dn ≥ 1 delMIS or LoF mutations

Considering delMIS and LoF mutations separately, we observed strong phenotypic similarity among patients with ≥ 1 *dn* delMIS but not LoF variant(s) (**Fig 5.3a**). This discrepancy is unlikely to be associated with autoimmune and psychiatric confounding effects (discussed above). Indeed, both individuals with ≥ 1 delMIS or ≥ 1 LoF variant(s) appear to be depleted of psychiatric and autoimmune (in the patient and their respective mothers) disorders compared to patients without variants ($p < 0.082$, $OR < 0.88$, separately and in combination; left-tailed Fisher tests). Furthermore, patients with ≥ 1 delMIS variants do not display more severe impairments. In fact, conforming to their greater penetrance, patients with ≥ 1 *dn* LoF variant(s) present with an excess of comorbidities compared to individuals carrying delMIS mutations, which is not confounded by low functioning ASD ($p = 0.0018$ and $p = 0.0028$ when considering and omitting the low IQ phenotypic variable, respectively; right-tailed Mann-Whitney U tests). Additionally, the inconsistency in phenotypic similarity is not caused by disparate molecular perturbations as *dn* delMIS and LoF SNVs affect common biological pathways (**Fig S5.6**). Therefore, we speculate the lack of phenotypic coherence among individuals with ≥ 1 *dn* LoF SNV(s) is likely to be due to insufficient statistical power. The number of patients with ≥ 1 *dn* LoF SNV(s) ($n = 401$) may be too low relative to the 934 phenotypically coherent individuals with ≥ 1 *dn* delMIS mutation(s) to detect a signal.

5.4.6 Phenotypically coherent modules underlie wide phenotypic spectra

Genotype to phenotype convergence

Restricting the analysis to patients with ≥ 1 *dn* LGD variants, we applied our phenotypic pipeline from **chapter 4** to gene modules comprising phenotypically similar patients. By showing patients from the same GTEx, Brainspan and PLN modules are phenotypically more similar than patients taken from different modules ($p < 1 \times 10\text{-}3$; right-tailed Mann-Whitney U tests) (**Fig S5.5**), we replicated our results from **chapter 4**, advising they extend across NDDs and forms of genetic variation.

Many phenotypes underlie phenotypic coherence of gene modules

No particular clinical feature describes module patients, given the lack of nominally enriched phenotypes (**Fig S5.7**). Instead, phenotypic coherence stems from the combination of many abnormalities, such that between 10 and 43, 23 and 43 and 15 and 25 phenotypes influence the clinical outcomes of patients from GTEx, Brainspan and PLN modules, respectively (**Fig 5.4a-b**). As discussed in **chapter 4**, the Goodall3 similarity index, used in this study, can present limitations relating to the up-weighting of extremely rare abnormalities when estimating phenotypic similarity. This is exemplified using patients from GTEx module 11 and PLN module 9, which are phenotypically similar across the seven and two (or less) phenotypes least associated with the corresponding modules (i.e. following the exclusion of ≥ 48 and 60

phenotypes most strongly associated with GTEx module 11 and PLN module 9, respectively, from the phenotypic similarity analysis) (**Figs 5.4a and S5.8**). Patients displaying the corresponding phenotypes are highly similar as these abnormalities have a low incidence (< 7 patients display them) (**Fig S5.8**). This is because the Goodall index strongly up-weights the shared presence of extremely rare phenotypes. Nevertheless, given that these phenotypes are not over-represented among module patients (p-values close to 1) (**Fig S5.8**), their phenotypic coherence is unlikely to represent a true biological signal. Instead, the phenotypic similarity signal may have been artificially inflated by the random segregation of the few individuals displaying such rare phenotypes within the same module.

Gene modules underlie more similar phenotypic spectra

Excepting PLN module 2 and GTEx module 4, phenotypic similarity diminishes when permuting an increasing number of patients between modules and background SSC patients ($p < 0.03$ and slope < 0, linear regression) (**Fig 5.4c-d**). This suggests the integrity of the patient phenotypic similarity network is critical in accurately quantifying clinical coherence. These observations suggest that SSC patients are organised in acyclic graph trees (**Fig 4.3**). Accordingly, the clinical analysis of large numbers of individuals with hits in common biological pathways (i.e. modules) is necessary to construct the phenotypic path between them. These observations enable us to extend our previous interpretation of pathway-driven phenotypic coherence (**chapter 4**). Accordingly, we postulate that both gene and structural variants, which

perturb common molecular processes, underlie the onset of similar phenotypic spectra.

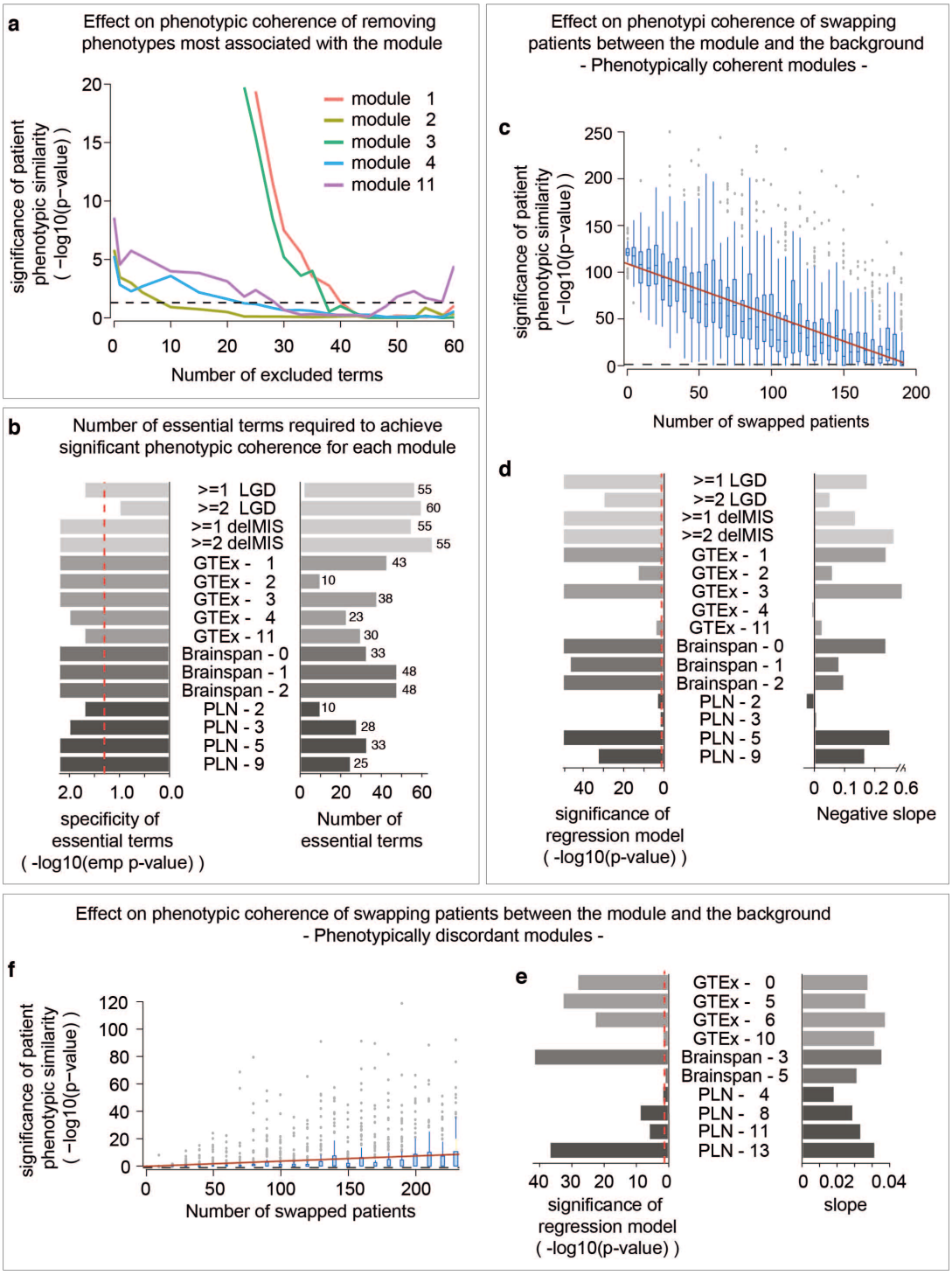


Fig 5.4 | Phenotypic coherence underlies phenotypic spectra. (**a**) Phenotypic coherence of patients from GTEx modules following the exclusion of increasing numbers of phenotypes from the analysis (p-values are derived from right-tailed

Mann-Whitney U tests; y-axis). Excluded phenotypes consist of those most strongly associated with patients from the evaluated modules; they are chosen based on increasing p-values from two-sided Fisher tests. Essential terms refer to the phenotypes whose exclusion decreases the phenotypic similarity signal ($p > 0.05$; black dotted line). Of note, data points associated with the removal of < 25 phenotypes for modules 1 and 3 are not shown as $p < 1 \times 10\text{-}20$. (**b, right panel**) The number of essential terms for each module. (**b, left panel**) To examine the specificity of essential terms, phenotypic coherence was re-evaluated for each module, following the exclusion of 100 permuted sets of essential terms (formed by randomly sampling the same number of phenotypes present in the cohort). An empirical p-value, shown here, was derived to reflect the proportion of permuted sets of essential phenotypes associated with significant phenotypic similarity. (**c to f**) Phenotypic coherence following the swap of increasing numbers of patients between modules and the background. During each swap, we ran 100 permutations. Linear regression models were used to describe the relationship between phenotypic similarity and the number of swapped individuals. (**c, f**) Examples of the effect on phenotypic coherence of swapping increasing numbers of patients between the background and (**c**) a phenotypically coherent module (GTEx module 3) or (**f**) a phenotypically discordant module (negative control; GTEx module 0). The red line depicts the associated regression model. (**d, e**) The (**left panel**) significance and (**right panel**) slope of models fitted.

5.4.7 Phenotypic coherence supports immune and non-immune ASD aetiologies

Inspection of the essential phenotypic terms, required to achieve significant phenotypic similarity for each phenotypically coherent module, reveals two trends. Firstly (1), essential phenotypes consist of psychiatric and autoimmune phenotypes, such that the depletion of these conditions in patients (and their mothers) underlies phenotypic similarity. Secondly (2), when a subset of all studied abnormalities constitutes essential phenotypes, it is the combination of their presences and absences that influences phenotypic coherence.

GTEx module 4 and PLN module 2 are associated with trend (1), such that patients from these modules (and their mothers) are depleted of autoimmune

and psychiatric phenotypes (**Fig S5.9**). As discussed in **section 5.4.5**, the observed phenotypic coherence stems from the shared absence of these phenotypes between module patients. Indeed, these individuals are not phenotypically similar when the corresponding phenotypes are excluded from the analysis (i.e. when examining phenotypic similarity across all non autoimmune and psychiatric phenotypes) (**Fig 5.3c, right panel**).

The remaining phenotypically similar modules are associated with trend (2). This coherence derives from the shared presence or absence of a particular set of phenotypes (not necessarily autoimmune or psychiatric abnormalities), such that module patients remain phenotypically similar even after removing psychiatric and autoimmune phenotypes from the analysis (**Fig 5.3c, right panel**). In this case, module patients are globally more similar across the corresponding essential phenotypes than expected by chance.

5.4.8 CVs as genetic modifiers of phenotypic variability

Aiming to evaluate whether CVs influence the clinical outcomes of disease, we examined the relationship between patient phenotypic and genetic similarities using linear regression.

Genetic similarity across genome-wide SNPs and SNPs from ASD GWAS

Generally, the phenotypic distance between patients is independent from the number of common alleles they share across the genome, or when

considering 95 nominally significant SNPs uncovered through ASD GWAS (Vardarajan *et al.*, 2013; Chaste *et al.*, 2015; Warrier *et al.*, 2015; Mitra *et al.*, 2016) (**Fig 5.5**). This remains true for patients with shared or distinct module memberships, as well as individuals from phenotypically coherent modules. This lack of positive findings is, however, foreseeable, given the substantial fraction of neutral variants among the eight million evaluated loci, and the uncertain relevance of unreplicated ASD SNPs (no ASD SNPs have been replicated in independent cohorts).

CVs affecting genes with NDD-associated biological functions influence phenotypic variability of disease

We, next, sought to examine the relationship between phenotypic similarity and genetic similarity across CVs affecting genes from NDD-associated functional classes (**Table 5.3; Fig 5.5**).

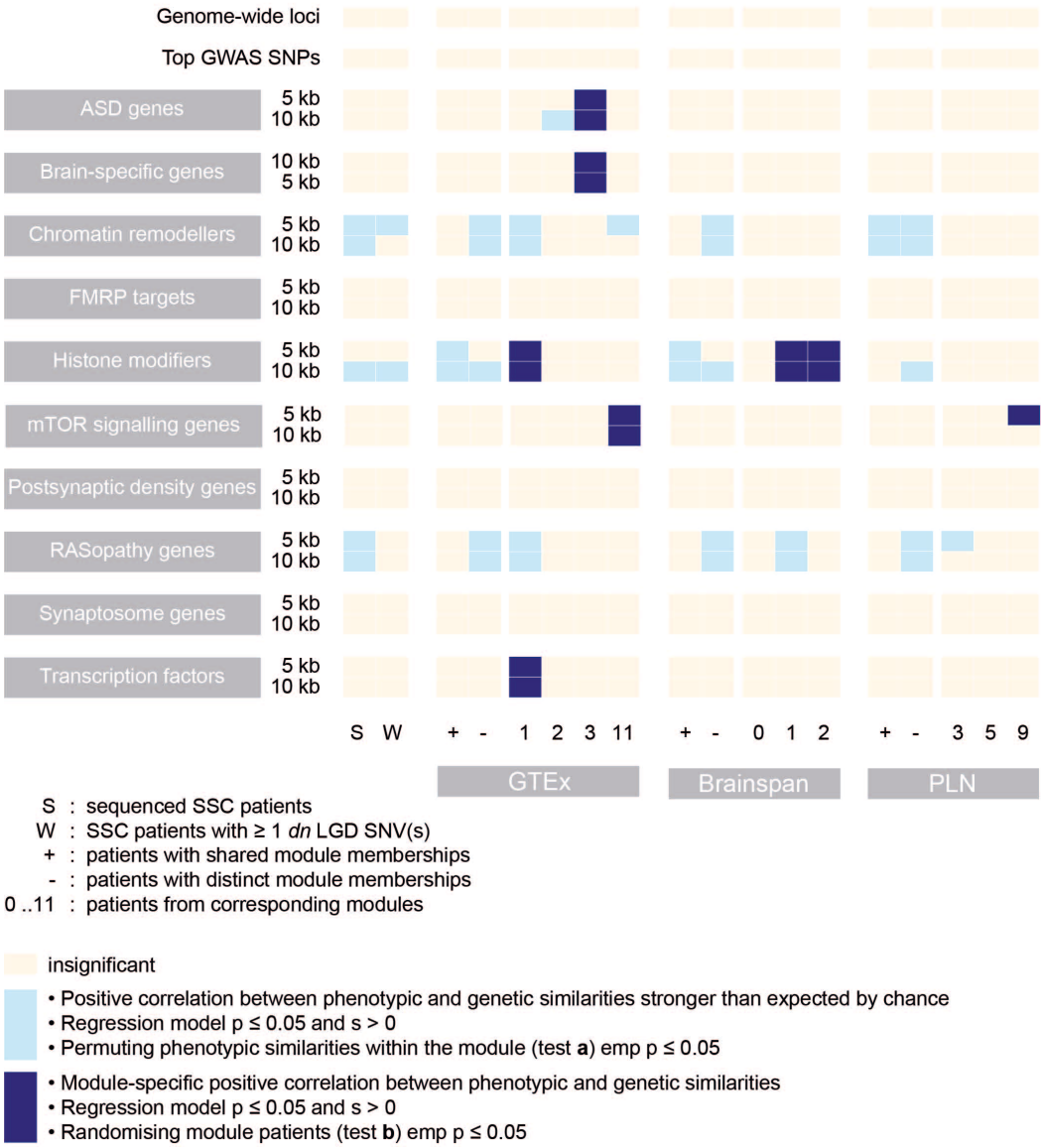


Fig 5.5 | Correlation between phenotypic and genetic similarities across genome-wide SNPs, ASD GWAS SNPs and SNPs within NDD-associated genes. The relationship between phenotypic similarity and genetic similarity was examined using linear regression across genome-wide SNPs, ASD GWAS SNPs and SNPs within 5 or 10 kb of genes from NDD-associated gene classes (**section 5.3.1; Table 5.3**). This relationship was tested across sequenced patients (S), patients with *de novo* (*dn*) likely gene-disrupting (LGD) SNVs (W), patients with shared module memberships (+), patients that never occur within the same module (-) and module patients (depicted using the module number). Beige depicts a lack of correlation between genetic and phenotypic similarities. Light blue depicts a positive correlation between phenotypic and genetic similarities, which is stronger than expected by chance (i.e. common variants (CVs) act as genetic modifiers of rare variants). Dark blue depicts a significant positive relationship between phenotypic and genetic similarities, which is specific to the module examined (i.e. CVs specifically modulate the phenotypes associated with the disruption of module genes). Significance levels were derived as described in **section 5.3.8** (light blue is associated with test **a** and dark blue with test **b**).

The genetic similarity of SSC patients, across SNPs within 5 or 10 kb of chromatin remodelling genes, is globally correlated with their phenotypic similarity (**Fig 5.5**). The signal is, however, not associated with patients with shared module memberships (or from particular gene modules), but derives from patients with distinct module memberships (**Fig 5.5**). This suggests that while CVs affecting chromatin remodelling genes may modulate clinical variability, this modulatory effect is likely to be hidden behind that of *dn* LGD variants. Accordingly, although the phenotypic distance between patients can be influenced by their genetic similarity, rare penetrant gene mutations exert a stronger influence on phenotypic manifestations of disease. Indeed, they are likely to be the primary determinants of phenotypic similarity in patients with functionally converging genetic disturbances. These observations are recapitulated when considering RASopathy genes (**Fig 5.5**), which similarly to chromatin remodeling genes are involved in regulation of cell growth, differentiation and survival.

We, next, examined whether CVs near genes participating in NDD-associated biological processes, other than chromatin remodelling, could also act influence phenotypic manifestations of disease. However, no such relationship was observed for CVs affecting brain genes (i.e. brain-specific genes, genes encoding proteins localised to postsynaptic densities and synaptosomes) (**Fig 5.5**). Nevertheless, disruption of brain genes is frequently observed in healthy individuals, in addition to NDD patients (**chapter 3**), suggesting CVs in such genes may contribute to healthy variation. Accordingly, CVs in disease-associated genes, such as essential genes and

those depleted of variation (e.g. chromatin remodeling genes), are less tolerable and could hence modulate disease phenotypes. This view is supported by the observation that the phenotypic similarity of SSC patients is positively correlated with their genetic similarity across SNPs within 5 or 10 kb of brain and synaptic genes, disrupted by *dn* LGD variants in ASD patients (i.e. Brainspan module 2; **Fig 5.6b**). In line with observations made for SNPs near chromatin remodelling genes, this signal derives from patients with distinct module memberships (**Fig 5.6b**). Hence, we postulate that CVs affecting ASD-associated synaptic genes may influence the phenotypic manifestation of disease, although at lower levels than rare penetrant variants. Further strengthening this observation, we observed the same pattern for PLN module 5 (i.e. the genetic and phenotypic similarity of SSC patients, and specifically those with distinct module memberships, is positively correlated across SNPs near genes from PLN module 5) (**Fig 5.6c**), which overlaps Brainspan module 2 ($p = 0.019$, $OR = 2.26$; right-tailed Fisher test) (**Fig S5.3**).

Lastly, the genetic similarity of all SSC patients (with shared or distinct module memberships) across SNPs within 10 kb of histone modifying genes is strongly correlated with their phenotypic similarity (**Fig 5.5**). This suggests CVs affecting histone modifying genes can modulate phenotypic disease variability and the associated effect is strong (i.e. more comparable to that of rare penetrant variants), possibly owing to their high likelihoods of essentiality.

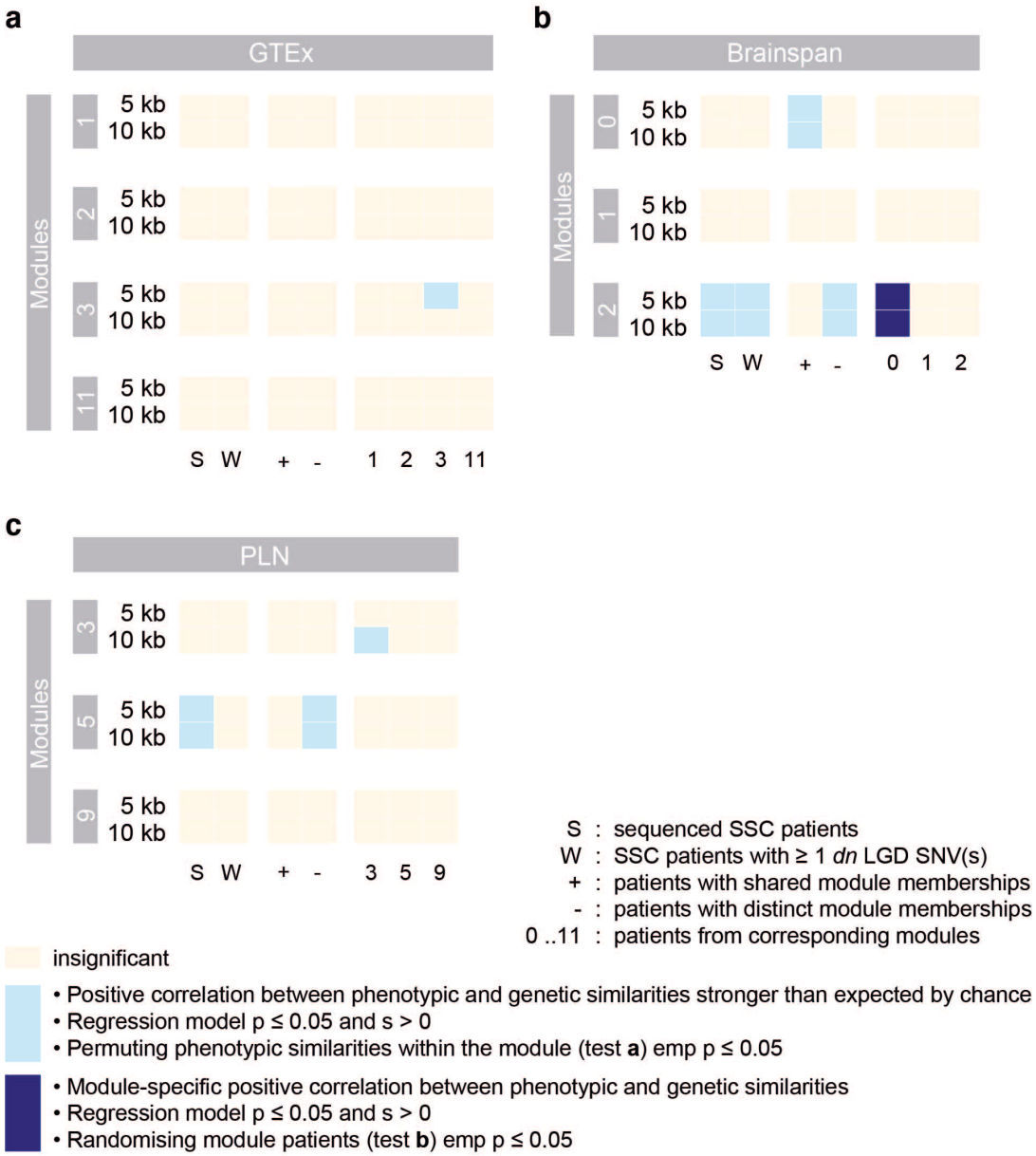


Fig 5.6 | Correlation between phenotypic and genetic similarities across SNPs near module genes. The relationship between phenotypic similarity and genetic similarity across SNPs within 5 or 10 kb of module genes was examined using linear regression. This relationship was tested across sequenced patients (S), patients with *de novo* (*dn*) likely gene-disrupting (LGD) SNVs (W), patients with shared (+) or distinct (-) module memberships and module patients when considering (**a**) GTEx, (**b**) Brainspan and (**c**) PLN modules (depicted by numbers). Beige depicts a lack of correlation between genetic and phenotypic similarities. Light blue depicts a positive correlation between phenotypic and genetic similarities, which is stronger than expected by chance (i.e. CVs act as genetic modifiers of rare variants). Dark blue depicts a significant positive relationship between phenotypic and genetic similarities, which is specific to the module examined (i.e. common variants (CVs) specifically modulate the phenotypes associated with the disruption of module genes). Significance levels were derived as described in **section 5.3.8** (light blue is associated with test **a** and dark blue with test **b**).

CVs specifically acting as modifiers of genes from particular gene modules

We, next, investigated whether CVs could specifically act as genetic modifiers of particular gene modules. For this, we examined the relationship between the phenotypic similarity of module patients and their genetic similarity across SNPs affecting particular gene sets (**Table 5.3**). Moreover, we tested the specificity of the observed relationships, by re-performing the analysis on permuted patient modules (**section 5.3.8**). We observed several instances where CVs influence the positioning of patients along the phenotypic spectra, associated with the corresponding phenotypically coherent modules.

Firstly, the number of common alleles, shared by patients from GTEx module 3 across high confidence ASD genes and brain-specific genes, determines their phenotypic similarity (**Fig 5.5**). In fact, randomising the module (i.e. randomly selecting the same number of SSC patients) suppresses this signal (**Fig 5.5**), suggesting the associated CVs exert a module-specific modulatory effect over phenotypic outcomes. Considering an expanded list of ASD genes (i.e. SFARI genes with scores 2, 3 or syndromic) (**Table 5.3**), we failed to confirm the positive correlation between genetic and phenotypic similarities of module patients (not shown). Nevertheless, given the excess of brain-specific genes among high confidence ASD genes with scores of 1 in the SFARI Gene database (Abrahams *et al.*, 2013), compared to genes with scores ≤ 2 (OR = 1.21) or ≤ 3 (OR = 2.23), brain specificity may be central to the observed relationship. Genes from GTEx module 3 are (1) highly expressed in the pre- and post-natal brain (**Fig S5.4**), (2) are enriched for neuronal GO terms

including *cell projection morphogenesis* (GO:0048858; BH p = 0.0040) (**Table S5.1**), and (3) the targeted disruption of their 1:1 mouse orthologues causes craniofacial mouse phenotypes (**Table S5.4**). Accordingly, CVs in brain-specific genes, may act as genetic modifiers of rare penetrant mutations affecting genes with brain-related functions (i.e. GTEx module 3 genes).

Secondly, we observed a positive relationship between the phenotypic similarity and genetic similarity of patients from GTEx module 1. Specifically, the genetic distance across SNPs within 5 or 10 kb of genes with roles in regulation of gene expression, namely transcription factors and histone modifiers, influences their phenotypic similarity (**Fig 5.5**). This modulatory effect is module-specific (**Fig 5.5**), suggesting the associated loci influence clinical outcomes in the presence of rare penetrant mutations affecting module genes. Genes from GTEx module 1 are likely to have roles during neuronal proliferation and differentiation, given the over-representation of GO terms relating to cell cycle and DNA damage response (**Table S5.1**) and their strong expression in the developing brain (n.b. while not as strongly, they are also expressed in the adult brain) (**Fig S5.3**). These observations suggest that the phenotypic manifestations of rare gene disruptions may be modified by CVs affecting expanded sets of functionally related genes. Correspondingly, the phenotypic similarity of patients from GTEx module 1 is, also, determined by the number shared alleles mapping to RASopathy and chromatin remodelling genes (**Fig 5.5**). Indeed, these sets of genes are directly or indirectly – via indirect regulation of gene expression – involved in processes of cell proliferation and differentiation, respectively. Nevertheless, this

modulatory effect is not module-specific, given the general influence exerted by SNPs near these genes on the clinical variability between SSC patients (**Fig 5.5**). We, then, repeated the analysis using Brainspan module 1, which partially overlaps GTEx module 1 (**Fig S5.3**), and comprises genes involved in gene expression control (**Table S5.2**) in the prenatal and early postnatal brain (**Fig S5.4**). We confirmed the specific and unspecific roles of CVs within histone modifying and RASopathy genes as genetic modifiers of rare penetrant mutations in genes from Brainspan module 1. Nevertheless, no relationship between genetic and phenotypic similarity was observed for transcription factors.

Thirdly, CVs affecting components of the mTOR signalling pathway appear to specifically influence the phenotypic manifestations of patients from GTEx module 11 (**Fig 5.5**). This module comprises genes, enriched for GO terms relating to GTPase activity (**Table S5.1**). Many genes from the mTOR signalling pathway encode GTPases (e.g. *RHEB*, *RAG* genes), GTPase-activating proteins (e.g. *TSC1* and *TSC2*) and GTPase-activated components (e.g. *MTOR*) (Laplante and Sabatini, 2012), supporting the view that the phenotypic effects of rare gene disruptions are modulated by CVs in functionally related genes. Nevertheless, these results are challenged by the lack of signal, when considering SNPs near RASopathy genes (**Fig 5.5**).

Replication using eQTLs

We repeated our analysis using eQTLs acting in *cis* and *trans* to genes from these sets but failed to replicate most signals (**Fig S5.10**). This could be due to differential functional effects between eQTLs, governing gene expression levels, and SNPs, which might be in linkage with variants affecting amino acid sequences, promoter activity, messenger RNA stability or localisation. Inevitably, it could, also, be a consequence of insufficient statistical power, given the relatively small number of patients and large numbers of eQTLs evaluated, or our inability to confidently identify eQTLs. Regardless, the interpretation of these results is difficult. Nonetheless, this work should be regarded as the first study providing supporting evidence for the role of CVs as genetic modifiers of rare penetrant mutations. In particular, we highlight the need to validate the influence of SNPs, affecting particular functional gene classes, over the positioning of globally coherent patients along the associated phenotypic spectra, beyond this study cohort.

5.5 Discussion

This work represents the first systematic study of the influence exerted by the interplay between rare penetrant *dn* gene mutations and common variants over the phenotypic manifestations of ASD. Specifically, we showed that while functionally converging rare *dn* gene disruptions underlie the onset of globally more similar clinical outcomes, CVs could act as modulators of the observed abnormalities in some pathways. Taken together, our work provides supporting evidence for a rare allele genetic model of ASD, incorporating a modifying role for CVs.

Taking advantage of the rigorously phenotyped and sequenced SSC cohort, we examined the phenotypic manifestations of functionally related and co-expressed rare *dn* gene disruptions. Albeit exhibiting coherent phenotypes globally (**Fig 5.3c**), no clinical archetypes could be linked to patients with similar molecular disruptions (**Fig S5.7**). Instead, in line with our findings from **chapter 4**, phenotypic similarity involves the shared presence and absence of many comorbidities and assessment of a large number of individuals is necessary to comprehensively shape the phenotypic spectra associated with the disruption of a specific biological process (**Fig 5.4**).

Our study pipeline uses pathway analysis approaches to detect molecular perturbations, whose phenotypic effects are then evaluated. While our findings indicate different genomic resources are complementary, rather than redundant, in identifying biological pathways, analogous functional modules

were consistently uncovered across NDDs, forms of genetic and inheritance patterns using the same resources (**Figs S4.9 and S5.3**). Specifically, genes involved in early neurodevelopmental and mature synaptic processes are affected by *dn* and inherited CNVs and SNVs in ASD, ID and DD, hence supporting their aetiopathological role across NDDs. Genes involved in early neurodevelopmental process encompass a wide range of functions. For example, disruption of transcription factors, chromatin remodellers and histone modifiers can induce aberrant transcriptional programs, possibly impairing neuroproliferation, migration and differentiation. Additionally, dysfunction of cell cycle-associated genes is likely to cause similar defects. Furthermore, synaptogenesis, occurring throughout neurodevelopment but also in adult life, involves the construction of specialised neuronal structures by a variety of adhesion and scaffolding proteins, as well as receptors and ion channels. Accordingly, disruption of genes encoding such components causes dysfunctional synapses, synaptic firing, transmission and plasticity, ultimately leading to abnormalities of higher mental function. Finally, such phenotypes are also associated with improper synaptic plasticity, involving mutations in genes encoding receptors, ion channel, signal transduction components, as well as transcriptional and translational regulators, modulating the response and adaptation to the external environment. Examples of such genes, many of which are encountered in SSC patients, are displayed in **Table 1.3**.

Our study supports the role of aberrant immunity as an environmental cause of ASD, based on the observation that autoimmune disorders in the patients

and their mothers are less common among patients with *dn* LGD gene disruptions. Accordingly, immunity represents an aetiological confounder in our genotype and phenotype analyses. For example, GTEx module 4 and PLN module 2 comprise phenotypically similar patients, which are depleted of autoimmune and psychiatric disorders. Accordingly, the phenotypic coherence signal derives from the distinction between individuals (and their mothers) with and without these phenotypes (**Fig 5.3c, right panel**). Several epidemiological studies support autoimmunity as an environmental cause of ASD. Firstly, (1) antibodies from autoimmune disorders, such as systemic lupus erythematosus, can be reactive against brain antigens (Trottier *et al*., 1999; Goines and Van der Water, 2010). Secondly, (2) increased levels of such antibodies occur in ASD patients (Ashwood and Van de Water, 2004; Goines and Van der Water, 2010) and SSC mothers (Brimberg *et al*., 2013; Keil *et al.*, 2010). Thirdly, (3) maternal auto-antibodies are able to cross into the foetus and interact with foetal brain antigens (Dalton *et al*., 2003; Singer *et al*., 2009; Braunschweig *et al*., 2012). Lastly, (4) individuals with a family history of autoimmune disorders have an increased risk of ASD (Keil *et al.*, 2010; Wu *et al.*, 2015). Moreover, an association between psychiatric and autoimmune disorders has been suggested (Davidson, 2012; Pego-Reigosa and Isenberg, 2008; Chen *et al.*, 2012; Nicholson *et al.*, 2012), possibly clarifying why autoimmune and psychiatric conditions segregate together in SSC patients (OR > 3.1).

To date, no common ASD risk loci have been successfully replicated through GWAS (Liu *et al.*, 2008; Ma *et al.*, 2009; Anney *et al.*, 2010; Anney *et al.*,

2012; Connolly *et al.*, 2013; Chaste *et al.*, 2015; Kuo *et al.*, 2015; Liu *et al.*, 2016a; Mitra *et al.*, 2016) and the evidence for an elevated polygenic ASD risk has been conflicting (Jansen *et al.*, 2017; Weiner *et al.*, 2017). Moreover, no difference in disease liability has been observed between patients with and without neurodevelopmental comorbidities (Weiner *et al.*, 2017). Consequentially, we hypothesised CVs may have a role in modulating the phenotypic manifestation of ASD, rather than induce disease. Accordingly, we undertook the first systematic investigation of the role of CVs in modifying ASD *dn* LGD pathway disruptions. Specifically, we showed that the fraction of common alleles patients share, across genes involved in control of gene expression (i.e. chromatin remodelling and histone modifying genes), influences their phenotypic resemblance across the evaluated clinical abnormalities (**Fig 5.5**). Studies of rare genetic variation have recurrently implicated regulation of chromatin structure in the aetiology of ASD, such that penetrant mutations in genes displaying such functions are solely sufficient to cause disease (Vorstman *et al.*, 2017). Complementing this view, this work advises common variants in these genes, while of very small individual risk effect sizes, could combinatorially influence phenotypic outcomes. Nevertheless, as expected, the impact of CVs appears to be less that of rare penetrant mutations, which represent the main determinants of the phenotypic manifestations of disease (**Fig 5.3**). Interestingly, CVs in brain and synaptic genes do not appear to modulate disease presentations, despite constituting one of the most frequently disrupted gene categories in ASD (Vorstman *et al.*, 2017). Nevertheless, we observed a strong positive relationship between phenotypic and genetic similarities across brain-specific ASD genes (i.e.

Brainspan module 2) (**Fig 5.6; Table S5.3**). From this we could speculate, that modulatory effect of CVs is confined to genes with higher probability of essentiality, which are likely to cause aberrant phenotypes upon rare mutational events. This is somewhat surprising, as one would expect purifying selection to remove such variants from the population, or maintain them at low frequencies. An alternate hypothesis is that rare and common variants within genes with lower likelihoods of essentiality contribute to healthy phenotypic variability. This view is supported by the observation that brain genes are frequently disrupted by both rare and common CNVs in healthy individuals (de Smith *et al.*, 2007; Raychaudhuri *et al.*, 2010), such that only a subset may be implicated in disease (**chapter 3**).

CVs have been shown to act as genetic modifiers of rare causal genetic lesions in some rare neurological syndromes (Badano *et al.*, 2006; Mlynarski *et al.*, 2015; Moreno-de-Luca *et al.*, 2015). Accordingly, we examined the relevance of this model in elucidating the phenotypic variability observed across phenotypically coherent patients (i.e. from phenotypically similar modules). While the phenotypic distance between individuals from phenotypically concordant modules is not governed by CVs affecting module genes (**Fig 5.6**), we observed genetic similarities across NDD-associated gene classes (**Fig 5.5**). Specifically, CVs affecting regulators of gene expression, brain-specific genes and mTOR signalling components, appear to act as genetic modifiers of rare penetrant ASD mutations affecting genes with analogous functions (i.e. from the corresponding functionally related gene modules) (**Fig 5.5**). Taken together, these observations suggest that while

functionally related rare gene disruptions underlie the onset of globally more similar clinical abnormalities, the expression of particular phenotypic spectra may be determined by CVs.

Our analysis of the role of CVs in the aetiology of ASD presents several limitations worth mentioning. Firstly, (1) our analysis supports a role for CVs as modifiers of phenotypic manifestations of disease for SNPs within 5 or 10 kb of genes but not for local or distal eQTLs (**Fig S5.10**). While this is not necessarily unexpected, given the unclear link between SNPs and eQTLs, we highlight the need to confirm the observed associations using independent datasets. Secondly, (2) our results are inherently subject to the gene sets evaluated, such that the analysis of CVs affecting distinct gene categories may provide additional insights into their modulatory roles. Thirdly, (3) the centre of assessment could inevitably confound the observed correlation between genetic and phenotypic similarities. This is because there are persistent biases in phenotypic impressions between clinics, patients are likely to have been examined at the nearest clinic and geographically closer individuals are presumably more genetically similar. While the assessment centre should not confound the evaluated modules (as they were identified using rare and penetrant *dn* LGD variants), modules could fortuitously comprise an excess of individuals examined at the same clinics. Based on the assumption that patients were examined at their closest centre, we ensured patients from the same modules are not more genetically similar than expected by chance. Furthermore, we showed phenotypic and genome-wide genetic distances are not strongly correlated. Nevertheless, the assessment

centres would ideally be included as covariates in the constructed regression models.

Our study relies on the partitioning of patients into groups of individuals displaying functionally related rare *dn* penetrant gene mutations. We, however, neglect the contribution of variants less likely to be pathogenic, as these are not readily interpretable. Such variants include missense SNVs with CADD C scores < 15, synonymous and inherited rare mutations. The simplex nature of the SSC cohort suggests *dn* variants underlie a substantial proportion of the risk, justifying the exclusion of rare inherited variation when forming gene modules. This is supported by the strong genotype to phenotype convergence observed for *dn* SNVs and the lack of signal for inherited CNVs (**chapter 4**). Thus, our ability to uncover genotype and phenotype relationships depends on the relevance of the examined genetic variation, inherently dependent on the nature of the study cohort (i.e. simplex or multiplex). A comprehensive understanding of disease and genotype and phenotype relationships will undoubtedly require the evaluation of all types of genetic variants. Lastly, as discussed in **chapter 4**, this study presents generic shortcomings relating to the clinical abnormalities evaluated and the phenotypic resolution of the ASD cohort.

Despite the discussed limitations, this work strengthens the paradigm that functionally related molecular perturbations cause similar phenotypic presentations of disease. Moreover, while we observed a role for CVs in modulating the observed clinical heterogeneity of ASD, individual SNPs are

likely to have negligible effects, based on the lack of significant GWAS findings. As a result, the observed patients abnormalities could globally be treated using therapies targeting modules disrupted by the rare genetic variants. In turn, this would involve interventions aimed at a predefined set of biological pathways, predisposing towards relatively well-characterised phenotypic spectra.

5.6 Conclusions

This study represents the first systematic investigation of genotype and phenotype relationships in ASD and advises the disorder involves both environmental (i.e. immunity-related) and genetic causes (i.e. *dn* LGD disruptions). In individuals with LGD disruptions, common molecular perturbations underlie the onset of globally more similar phenotypic outcomes. Furthermore, this work supports a role for CVs affecting genes involved in various NDD-associated processes in modulating the clinical variability between patients.

CHAPTER 6

Mouse neuroanatomical models:

a brain functional genomics study

6.1 Introduction

Beyond the understanding of the fundamental processes governing *Life*, the elucidation of the function of the 20,000 protein-coding genes, or so, annotated in the human genome has strong implications in disease and the development of treatment strategies. Over the past decade, the advent of high throughput technologies and their decreasing cost has enhanced our knowledge of human genetic variation and the mutational spectra in disease. Despite these advancements, the function of many genes remains obscure, inevitably contributing to our inability to distinguish between and interpret neutral and pathogenic variants. Such realisations highlight the value in characterising the ensemble of protein-coding genes using methodologies such as computational predictions, animal models and *in vitro* systems.

Capitalising on the evolutionary conservation of fundamental biological processes across species, such as development, metabolism and physiology (Kirschner and Gerhart, 2006), animal and in particular mouse models provide key organisms to investigate gene function. The utility of mouse models is made possible by construct validity, which relies on the existence of decidedly orthologous genes, present at 1:1 ratios between human and mouse (i.e. 80% of human genes have a 1:1 mouse orthologue; Breschi *et al.*, 2017). Furthermore, the translation of phenotypes and their associated causes between the two species (Buxbaum *et al.*, 2012; Gai *et al.*, 2012; White *et al.*, 2013; Honti *et al.*, 2014; Robinson and Webber, 2014) can support their face validity. Genetic studies of mouse models, enabled by the construction of

gene KnockOuts (KOs) via targeted mutagenesis of embryonic stem cells, led to the functional characterisation of large number of genes. Indeed, as of 2017, phenotypic and molecular data is available for over 9,500 human orthologues (Blake *et al.*, 2017).

Traditionally, several shortcomings have been associated with the use of mouse models in the study of gene function and human disease. Firstly, there are strong ascertainment biases towards well-studied genes and those of interest to investigators (Steinberg *et al.*, 2015). As a result, unknown genes have been mostly neglected, irrespective of their similar likelihoods to contribute to disease (White *et al.*, 2013; Cassa *et al.*, 2017). Secondly, model examination is inherently biased towards phenotypes and molecular assessments of interest. Consequently, the full functional and disease spectrum of genes is often underreported, as a result of their pleiotropic effects. Lastly, the diversity of mouse strains, genetic backgrounds, allelic designs and phenotypic assessments contributes to the overall heterogeneity of results and lack of replicability of KO models. These limitations underpin the necessity for large-scale and standardised mutagenesis screens accompanied by extensive systematic phenotypic assessments to impartially investigate the function of virtually every protein-coding gene from the mouse genome. To this end, many consortia have been established (**Table 1.4**) and their collaborative efforts are facilitating the delivery of vigorous biological data and resources. Much of the current phenotypic and genetic knowledge of mouse KO models, encompassing over 9,500 human orthologues, has been integrated in a community resource, which can be accessed from the Mouse

Genome Informatics database (MGI: http://www.informatics.jax.org) (Blake *et al.*, 2017).

The availability of robust and comprehensive phenotypic data is central to the relevance of mouse models in the study of human disease. In this respect, large-scale projects, such as the International Mouse Phenotyping Consortium (IMPC) and the Wellcome Trust Sanger Institute Mouse Genomics Project (Sanger MGP), have established phenotyping pipelines examining variables relating to 28 physiological systems, ranging from embryogenesis and reproduction to nervous system and behaviour. In turn, such efforts have been facilitated by the development of the Mammalian Phenotype Ontology (MPO) (Smith and Eppig, 2009), which provides a structured set of over 7,000 Mouse Phenotypes (MPs). MPs are organised in a hierarchy, such that more general phenotypic terms are parents to more specific terms, hence enabling systematic and reliable comparisons between models. Using such ontologies, the IMPC has set out to characterise 5,000 mouse KO lines by 2016 and 15,000 by 2021 (Brown and Moor, 2012; Dickinson *et al.*, 2016).

Mouse phenotyping approaches evaluating organ- or stage-specific abnormalities associated with a disease of interest have been instrumental in characterising their underlying pathologies. For example, aiming to explore the role of essential genes, commonly implicated in human disease, Dickinson *et al.* undertook a rigorous and systematic embryonic screen of genes inducing mouse lethality or sub-viability, as part of the IMPC initiative

(Dickinson *et al.*, 2016). Specifically, in addition to gross phenotyping implemented using the MPO, detailed embryonic morphological and quantitative abnormalities (including time of death) were recorded using 3-dimensional brain imaging methods (Dickinson *et al.*, 2016). These efforts revealed additional phenotypes for KO lines, missed by the gross morphological inspection, and shed light onto the developmental trajectories and function of many genes. Moreover, large-scale unbiased mutagenesis screens investigating disease-specific phenotypes can be used to prioritise novel candidate genes. For example, Bassett *et al.*, characterised the skeleton of 100 unselected mouse KO lines, using specific structural and biochemical variables relating to bone mass and strength (Bassett *et al.*, 2012). As a result, they identified several genes leading to various bone abnormalities, further proposing novel candidates for skeletal diseases, such as osteoporosis, osteogenesis imperfecta and sclerosteosis (Bassett *et al.*, 2012).

The use of mouse models in disease-oriented screens is of particular relevance to NeuroDevelopmental Disorders (NDDs), such as Intellectual Disability (ID), and brain malformations. Indeed, mouse models (1) enable the use of invasive approaches and the extraction of brain developmental tissue. Furthermore, (2) only a small portion of the thousands of genes predicted to underlie NDD risk have been validated so far, suggesting such screens are instrumental in identifying novel candidates. Finally, (3) mouse models provide valuable systems to interrogate the pathogenicity of the large number

of variants uncovered through genetic screening of NDD patients and evaluate the underlying disease mechanisms.

ID affects 1% of the population and is characterised by significant limitations in intellectual functioning (evidenced by Intellectual Quotients (IQ) < 70), and adaptive behaviour, arising during early childhood. The disorder varies highly in its severity and rarely develops in isolation, such that additional comorbid disorders, neurologic problems and congenital malformations arise in 30% to 70% of patients (Cooper *et al.*, 2007; Einfeld *et al.*, 2011; Cooper *et al.*, 2015). In particular, Autism Spectrum Disorder (ASD), associated with deficits in social interaction and communication and restricted and repetitive behaviours, is present in 4% to 30% of ID patients (Cooper *et al.*, 2007; Van Naarden Braun *et al.*, 2015). Moreover, although individually rare (i.e. less than 1/10,000 births), brain malformations collectively affect an appreciable number of NDD patients. Accordingly, they occur in up to 15% of epilepsy patients and are commonly associated with rare ID syndromes. Of note, brain malformations are likely to be significantly more common in NDD patients and in apparently healthy individuals than reported, as only a small fraction of individuals are examined for neuroanatomical defects.

The development of the cerebral cortex involves the synchronization of highly complex events governing the differentiation and formation of the tight network of approximately 100 billion neurons constituting the brain. Unsurprisingly, disruptions during various stages of neurodevelopment result in brain malformations and disorders. Specifically, abnormalities affecting the

proliferation of neuroprogenitors has been linked to microcephaly, characterised by a small brain with a circumference of two or more standard deviations below the mean of subjects of the same age, sex and ethnicity, and macrocephaly (i.e. disproportionately large head). Conversely, disorders of migration implicate the inability of neurons to reach their final destinations and encompass cortical lissencephalies, characterised by the lack of brain folds and groove development. Lastly, post-migratory defects of neuronal differentiation have been linked to polymicrogyrias, wherein the brain presents an excess of unusually small folds.

The pathogenesis of NDDs and brain malformations is complex and multifactorial. Nevertheless, epidemiological and twin studies have implicated an underlying genetic aetiology. Many types of genetic variants have been identified and include chromosomal re-arrangements, Copy Number Variants (CNVs; deletions and duplications of DNA segments of > 1 kb) and rare *de novo* single gene mutations. In fact, the advent of large-scale genotyping and sequencing studies are uncovering increasing numbers of deleterious variants. For example, over 750 genes are currently associated with ID (Casanova *et al.*, 2016; Kochinke *et al.*, 2016) and many more are expected to underlie disease risk. In this respect, mouse models represent powerful systems to conduct mutagenesis to validate uncertain and identify novel candidates. In particular, one can investigate the underlying pathogenic mechanisms of disease by examining the shared functional disruptions of animals presenting with the abnormalities of interest. Such analyses rely on the assumption that genes, which participate in common molecular

processes, contribute to the development of analogous phenotypes. Furthermore, comparisons of disease-associated mouse KO lines are highly valuable in disentangling the genetic and phenotypic heterogeneity of NDDs and brain malformations.

6.2 Chapter aims

We aimed to identify genes and biological mechanisms involved NDDs, such as ID and ASD, and brain malformations. In collaboration with Binnaz Yalcin, Institute of Genetics and Molecular and Cellular Biology (IGBMC), University of Strasbourg, France, and Anna Mikhaleva, Centre for Integrative Genomics, University of Lausanne, Switzerland, we analysed high throughput mouse KO data, generated as part of the Sanger MGP. Specifically, using neuroanatomical data regarding 78 brain parameters from 1,057 mouse gene KOs, we (1) identified genes associated with strong mouse neuroanatomical defects, (2) evaluated their functional convergence, described their biological roles. Furthermore, (3) we searched for genotype to phenotype convergence and asked whether animals with disruptions in common molecular processes display similar abnormalities. Finally, (4) we examined the relevance of mouse brain phenotypes in the study of human disease and interrogated their involvement in ID and brain malformations.

6.3 Specific materials and methods

The work described in **sections 6.3.1 to 6.3.4** was performed by Binnaz Yalcin's group, Institute of Genetics and Molecular and Cellular Biology (IGBMC), University of Strasbourg, France and their collaborators. This work includes (1) the generation and neuroanatomical characterisation of mouse mutant lines, (2) the development of the integrative statistical pipeline used to identify genes associated with mouse neuroanatomical defects and (3) the imputation of the mouse phenotypic dataset.

6.3.1 Study samples and datasets

Mouse mutant lines were generated as part of the Sanger MGP (White *et al.*, 2013), partner of the IMPC, using the *knockout-first allele* method (Skarnes *et al.*, 2011) (**Fig S6.1**). For each KO line, at least three WildType (WT) and three HOMozygous (HOM) null or HETerozygous (HET) (when homozygosity was lethal) mutant brain tissues were provided. Only male samples were used and no ascertainment or personal biases influenced the choice of genes to evaluate. Furthermore, there was no bias in their genomic distribution, such that all mouse chromosomes, except Y, were covered. Ultimately, a dataset, comprising 3,464 KO and 1,024 WT samples of mice of fourteen to sixteen weeks of age, was analysed.

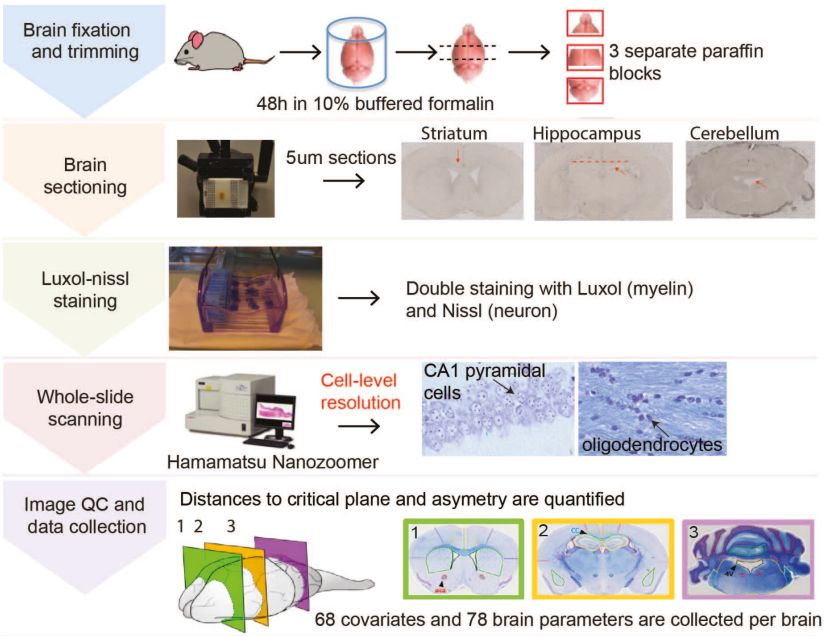


Fig 6.1 | Standardized workflow for morphological phenotyping of the adult mouse brain. The experimental procedure consists of five main steps. Each brain is fixed in 10% buffered formalin for 48 hours and trimmed across two planes for paraffin embedding, thus producing three distinct blocks (step 1). Careful and precise sectioning is prerequisite in this protocol, sections must be well symmetrical and match well defined anatomical features (step 2). Then, each section is double-stained using nissl (violet) for neurons and luxol (blue) for myelin, generating a color contrast between white and grey matter (step 3). Whole slides are scanned using a digital slide Hamamatsu Nanozoomer scanner, producing whole brain images at cell-level resolution (step 4). Finally, the data is collected using a set of 68 experimental co-variates and 78 brain parameters, mapping to Bregma positions +0.98mm, -1.34mm and +5.80mm, and extensive quality control checks and critical evaluation of the entire dataset were carried out (step 5).

In total, 68 covariates and 78 brain parameters of area and length measurements were assessed, using a recently developed standardised and robust histomorphological approach (Mikhaleva *et al.*, 2016). Neuroanatomical data originates from three precise and symmetrical planes of the adult mouse brain, corresponding to stereotactic Bregma positions +0.98 mm, -1.34mm and +5.80mm (herein termed *Bregma sections 1, 2* and *3*, respectively) (**Fig 6.1**), such that the 78 parameters encompass five biological categories, namely brain size, commissures (callosal, anterior and hippocampal), ventricles (lateral and dorsal), cortex (motor, somatosensory,

cingulate, piriform and retrosplenial), and subcortex (hippocampus, amygdala, tracts, caudate putamen, internal capsule, habenula and fimbria). Measurements across left and right hemispheres were, further, combined to increase precision, reducing the number of parameters to 48. Nine parameters were excluded due to high rates of missing data, leaving 39 parameters for downstream analyses.

6.3.2 Experimental flow and statistical analysis

Fig 6.1 shows the overall experimental workflow. Detailed experimental procedures will be available in the published manuscript and can be obtained through personal correspondence.

Statistical analyses were performed using an in-house integrative R pipeline (**Fig 6.2**), carried out separately for each assessment centre (termed *project*), as this covariate explained the highest phenotypic variance (R^2 = 12.6%). Of note, the stereotactic position of the analysed image (R^2 = 0.002%) and its symmetry score (R^2 = 0.006%) explained the least variation, demonstrating the accuracy of the study. Results were, subsequently, merged. A linear mixed model was fitted in R using the PhenStat package (Kurbatova *et al.*, 2015), such that fixed, dependent and random variables consisted of genes, brain parameters and animal necropsy dates, respectively. Of note, the PhenStat package (Kurbatova *et al.*, 2015) was not suitable from a small subset of data, for which a Student t-test was used instead.

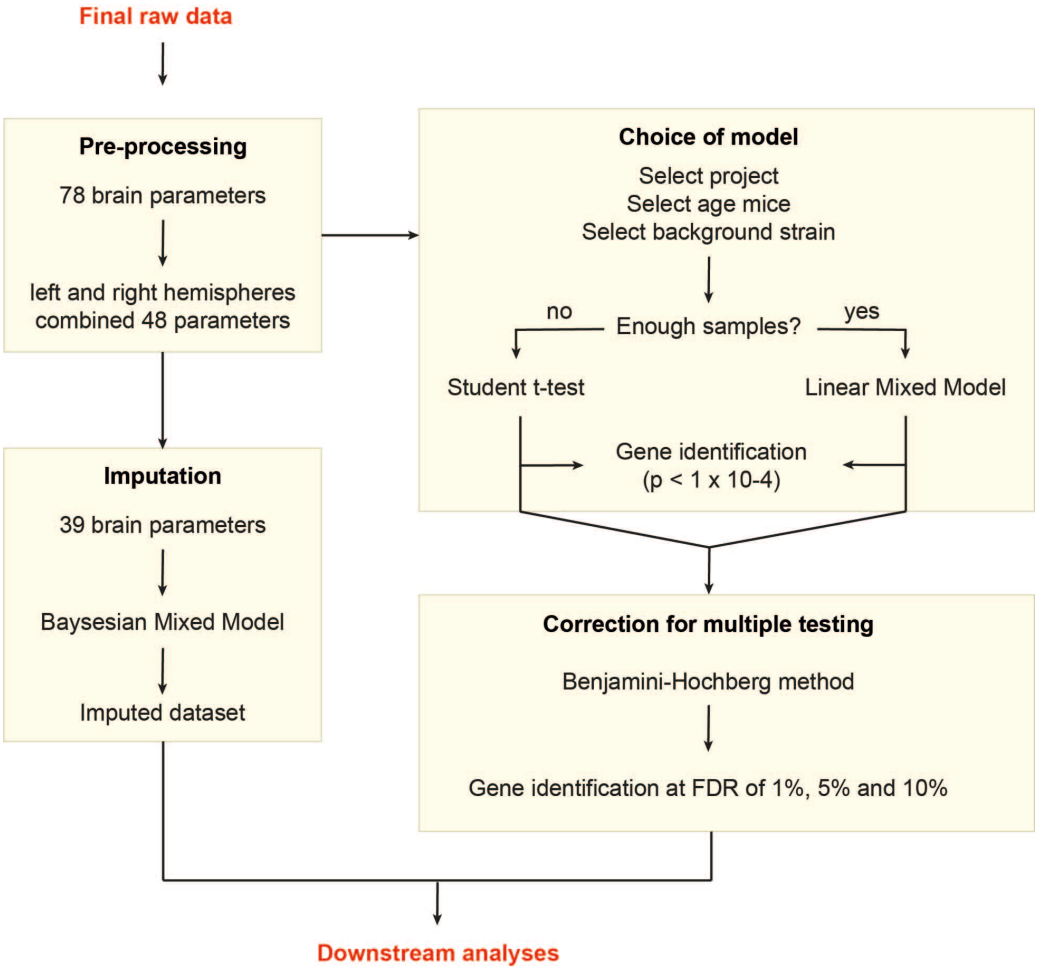


Fig 6.2 | Integrative statistical pipeline. The statistical framework consists of four steps. (1) Pre-processing: To analyze all genes with respect of their genotype and construction, a variable collating gene, mutant type and targeting strategy information was created from the raw data and parameters such as age, background strain, sex and sub project, were used to create batches of analysis. Measurements of left and right hemispheres were combined, reducing the number of parameters from 78 to 48. (2) Choice of model: A linear mixed model was fitted in R (PhenStat package; Kurbatova *et al.*, 2015) using the necropsy date as a random temporal variable. When the model failed due to a lack of samples (about 3% of the data), a two-tailed equal variance Student t-test was used. (3) Multiple corrections: Multiple correction testing was applied using the BH method to control the false positive rate. We produced three gene lists at thresholds of 1%, 5% and 10% FDR. (4) Imputation: a Bayesian multiple-phenotype mixed model method was used to overcome the problem of missing data in the downstream analyses of the study.

Gene association was considered as highly confident when the unadjusted p-value was below the IMPC significance threshold of p = 1 x 10-4 (Kurbatova *et al.*, 2015). Multiple testing corrections were performed using the Benjamini-Hochberg (BH) method, producing three gene lists at the adjusted p-value

thresholds (False Discovery Rates; FDR) of less than 0.01 (1% FDR), 0.05 (5% FDR) and 0.1 (10% FDR). Relevant mouse phenotypes (MP) were used to describe neuroanatomical defects and when needed, new MP terms were created specifically for this study through collaboration with the Jackson Laboratory (http://www.informatics.jax.org/).

6.3.3 Data imputation and imputed datasets

Imputation of missing data was performed using a Bayesian multiple-phenotype mixed model method, termed PHENIX (Dahl *et al.*, 2016). Samples with high data missingness were excluded from the study, leaving 4,356 mouse samples (corresponding to 1,130 alleles (also termed constructs; **section 6.3.4**) from 1,057 genes as well two chromosomal deletions and one CpG island, which were excluded from the study) for imputation. Applying the same statistical pipeline to the imputed and non-imputed datasets produced similar gene lists. Nevertheless, manual and automatic assessment of imputation performance, at the gene and parameter levels, showed an increase in detection power, justifying the use of the imputed dataset for downstream analyses.

6.3.4 Data availability

Raw and processed data (non-imputed and imputed) used throughout this study will be available in the published manuscript (at this stage, data can be shared through agreement with authors). Over 10,000 manually annotated

brain images, handled in this manuscript, are freely available on the IMPC project portal.

6.3.5 Mouse gene KO dataset

This study investigates the neuroanatomical consequences arising from the targeted disruption of 1,057 distinct genes in mouse models.

We note that (1) histopathological examinations were performed in several centres using different protocols (termed *project*). Furthermore, (2) seven mouse strains were used (i.e *B6N*, *B6JTyr;B6N*, *B6Brd;B6Dnk;B6N-Tyr<c-Brd>*, *B6*, *B6JIco*, *C57BL/6N;C57BL/6NTac* and *C3Fe*) and (3) three mouse mutant types were created (i.e. 745 homozygotes, 338 heterozygotes and 47 hemyzygotes, denoted *HOM*, *HET* and *HEM*, respectively) using (4) 26 gene targeting and trapping strategies. To facilitate equitable comparisons, we considered the 1,130 *gene-project-strain-mutant-strategy* constructs, separately. For the 63 genes present in duplicate, triplicate and quadruplicate, one construct was selected per gene. Specifically, we excluded projects lacking neuroanatomical data for features from Bregma section +0.98mm (section 1), leaving 47 genes present as multiple constructs. These were further filtered based on BH-adjusted p-values, such that we maintained those associated with strongest neuroanatomical abnormalities compared to genotype-matched WT mice. The chosen constructs are, hereafter, simply referred to as *genes*. The final dataset consists of 1,057 mouse gene KOs, of

which 702, 316 and 39 are present as HOM, HET and hemizygous mutations, respectively.

Of the 1,057 mouse KO genes, 1,001 have at least one human orthologue, based on homology data downloaded from Ensembl Biomart (Cunningham *et al*., 2015). To optimise translation of gene function between the two species, the human gene analysis is restricted to 960 genes, for which we established 1:1 orthology.

6.3.6 Genes causing strongest and weakest neuroanatomical defects

We measured the average and standard deviation of each neuroanatomical measurement across genotype-matched WT mice mapping to each mouse gene KOs. For each brain feature, we estimated the z-score deviation of gene KO replicates from the average of their matched controls. Ultimately, we considered the average z-score per gene and neuroanatomical feature.

We sought to evaluate whether this large-scale mouse mutagenesis screen results support the view that the disruption of genes with higher likelihoods of deleteriousness have stronger consequences on survival and development. For this, we assigned the maximum absolute z-score value across brain features to each gene and ranked them in decreasing order. Considering only the 960 mutant genes with 1:1 human orthologues, we, then, selected the top 10% of genes associated with the highest and lowest scores. These genes are herein termed *top* and *bottom* 10% genes. The 10% threshold was

chosen as it produced the strongest signal. Nevertheless, analogous results were found when using 1%, 5% and 15% top and bottom genes (not shown). Similar results were found when top and bottom genes were defined as those with highest and lowest cumulative absolute z-scores across brain features, although the signal was weaker (not shown).

As a negative control, we created 1,000 randomised sets of top 10% genes (termed *permuted sets of top 10% genes*) by randomly drawing the same number of genes from all 1,057 mouse KO genes or from the 960 human orthologues.

6.3.7 Phenotypic similarity

Phenotypic similarity between two genes was assessed as the similarity of brain defects resulting from their targeted disruption in mouse models. For each neuroanatomical feature, significant abnormalities (BH $p \leq 0.05$) were re-scored as 1 (presence) or otherwise 0 (BH $p > 0.05$; absence). For each pair of mouse KO genes, phenotypic similarity for a given brain feature was calculated using a simplified version of the *Goodall3* index (**section 2.6**) (Boriah *et al*., 2008). Given the low incidence of significant abnormalities post BH correction, this simplification generates a more conservative index, whereby genes lacking particular sets of neuroanatomical defects do not drive phenotypic similarity. We, thus, simplified the Goodall3 measure to only assess the shared presence of abnormalities between gene pairs, weighted

by their overall frequency. The new simplified measure was computationally implemented as:

$$S(i,j)_k = \begin{cases} 1 - f_k^2 \ , & if\ k\ is\ present\ in\ both\ patients \\ 0 \ , & if\ k\ is\ absent\ in\ both\ or\ present\ in\ one \end{cases}$$

where $S(i,j)_k$ is the Goodall3 measure between genes *i* and *j* for phenotypic variable *k* and f_k is the proportion of all mouse KO genes inducing significant abnormalities of the brain feature represented by *k*.

Phenotypic similarity was measured for all pairwise gene combinations and brain regions. The overall phenotypic similarity of two genes was estimated as the average similarity across Bregma +0.98mm and -1.34 mm, termed sections 1 and 2, respectively.

6.3.8 Human gene intolerance scores and mouse lethal genes

Mutational constraint scores were downloaded for 17,047 and 17,816 human protein-coding genes from Petrovski *et al.*, 2013 and Lek *et al.*, 2016, respectively (**Table 2.1**). These scores are entitled Residual Variation Intolerance Score (RVIS) and probability of being Loss of function Intolerant (pLI), respectively. RVIS evaluates the rate of common functional gene variation using data from the National Heart, Lung and Blood Institute (NHLBI) GO Exome Sequencing Project, such that genes with lower RVIS are more likely to be under purifying selection and thus intolerant to mutations (Petrovski *et al.*, 2013). Conversely, the pLI score assesses the depletion of

penetrant Loss of Function (LoF) rare gene variants, defined as nonsense, splice-site and frameshift mutations, based on data from the Exome Aggregation Consortium (ExAC; Lek *et al.*, 2016).

Selection coefficients (s_{het}) were downloaded for 15,999 human genes (Cassa *et al.*, 2017). S_{het} scores have been estimated based on the expected gene mutation rates, number of observed heterozygous LoF variants in individuals from ExAC and the number of chromosomes sampled. An s_{het} score provides a measure of the reduction in fitness associated with the loss of heterozygosity across gene alleles.

We downloaded the rate of synonymous (dS) and non-synonymous (dN) gene substitutions between mouse and human for 23,147 mouse genes from Ensembl Biomart (Cunningham et al., 2015) and quantified selection pressures as the dN/dS ratio.

Finally, a set of 410 mouse genes, inducing embryonic lethality in mouse KO lines, was downloaded from Dickinson *et al.*, 2016.

6.3.9 Mouse gene expression datasets

We downloaded RNA sequencing expression data for 69,690 mouse genes, generated by the Encode project (ENCODE Project Consortium, 2012), as described in Pervouchine *et al.*, 2015. Gene expression levels were measured

as Fragments Per Kilobase of transcript per Million mapped reads (FPKM) across 25 bodily tissues, spanning five developmental points (E11.5, E14, E14.5, E18, eight weeks). Focusing on the 21,680 protein-coding genes as defined by Ensembl Biomart (Cunningham *et al.*, 2015), we calculated tissue expression of each gene as the mean FPKM value across tissue replicates, collected from age-matched mice. We formed a brain developmental dataset by considering gene expression levels in the CNS at embryonic stages E11.5, E14 and E18, as well as across the adult brain, as given by the average expression level across the cerebellum, cortical plate and frontal cortex in eight-week-old mice. Of note, during gene expression analyses, this dataset was normalised as described in **section 2.7.2**.

6.3.10 Gene expression levels for Bregma +0.98mm and -1.34 mm

Considering only the 960 mutant genes with 1:1 human orthologues, we derived each gene's expression level in critical section 1 (Bregma +0.98mm) and section 2 (Bregma -1.34mm) using the Brainspan expression dataset. Specifically, gene expression levels in Bregma sections 1 and 2 were defined as the maximum expression value across the striatum and neocortex and across the amygdala, hippocampus and neocortex, respectively. For each gene module, we calculated the mean of the expression ratio between Bregma sections 1 and 2 across the corresponding genes.

6.3.11 Mouse gene networks

Using the developing CNS and adult brain gene expression dataset, described in **section 6.3.9**, we constructed a gene co-expression network, termed *mouse developing CNS and adult brain co-expression network* (**Table 2.3**). We considered CNS tissues taken from E11.5, E14 and E18 littermates, as well as cerebellum, cortical plate and frontal cortex tissues collected from eight-week-old mice. Of note, disparate regions constitute embryonic CNS and adult brain tissues and no measurements were available for young mice, causing a large gap between the developmental and postnatal stages examined. Co-expression coefficients were derived by estimating the correlation of expression patterns of pairs of protein-coding genes across all relevant tissues and developmental samples using Pearson's coefficients.

We downloaded a mouse integrative functional gene network, denoted MouseNet v2 (**Table 2.3**) (Kim *et al.*, 2016). MouseNet incorporates various -omics resources, namely (1) Protein-Protein Interactions (PPIs) (i.e. literature-curated interactions or uncovered via high throughput screens), (2) expression data (i.e. 183 GSE maps) and (3) functional annotations (i.e. co-citations and phylogenetic similarities across species). In addition to mouse datasets, this network also includes human, yeast and fly orthologous data to increase coverage. Bayesian statistics were used to integrate these data-sources using mouse Gene Ontology – Biological Process annotations as a benchmark (Kim *et al.*, 2016).

6.4 Results

The work described in **section 6.4.1** was performed and provided by Binnaz Yalcin's group, Institute of Genetics and Molecular and Cellular Biology (IGBMC), University of Strasbourg, France and their collaborators.

6.4.1 Summary of data

Mouse brain samples were obtained through collaboration with the Sanger MGP (White *et al.*, 2013). In total, 1,024 WT and 3,464 KO samples, corresponding to 1,060 distinct genes, chromosomal deletions or CpG islands, were assessed for 39 brain parameters using a standardised and robust histomorphological approach at stereotactic Bregma positions +0.98mm, -1.34mm and -5.80mm (Bregma sections 1, 2 and 3, respectively) (**Fig 6.3a**) (Mikhaleva *et al.*, 2016). We identified 75, 122 and 158 genes (of the 1,057 examined), whose disruption in the mouse yields at least one strong neuroanatomical defect at 1%, 5% and 10% FDR thresholds, respectively (herein termed *NeuroAnatomical Phenotypes genes* or *NAP genes*). Of these, 97 show strong neuroanatomical deficits, relative to WT animals, at the IMPC threshold of $p < 1 \times 10\text{-}4$.

Approximately one third of mouse NAP genes, significant at the IMPC threshold of $p < 1 \times 10\text{-}4$, have an impact on two or more brain regions (**Fig 6.3b**). Examples include *Rnf10*$^{-/-}$ (Ring finger protein 10), *Pfn1*$^{+/-}$ (Profilin 1) and *Camsap3*$^{-/-}$ (Calmodulin regulated spectrin associated protein family

member 3). Conversely, *Sik3*$^{-/-}$ (SIK family kinase 3) affects specifically the anterior commissure (-39.5%, BH p = 1.98 x 10-8), *Anks1b*$^{-/-}$ (Ankyrin repeat and sterile alpha motif domain containing 1B) the dorsal hippocampal commissure (+91.4%, BH p = 1.24 x 10-6), *Sgms1*$^{-/-}$ (Sphingomyelin Synthase 1) and *Hsp90aa1*$^{-/-}$ (Heat Shock Protein 90 Alpha Family Class A Member 1) the ventricles (-166.9%, BH p = 0.01 and +212.5%, BH p = 3.07 x 10-9, respectively), and *Slitrk4*$^{-/Y}$ (SLIT And NTRK Like Family Member 4) the somatosensory cortex (-15.2%, BH p = 0.009). The most common brain malformations involve the commissures (40.8%), followed by the ventricles (27.7%), which are mainly enlarged (ventricular atrophy is rare), cortical malformations (12.6%), brain size defects (6.8%) and hippocampal anomalies (3.1%).

The majority of high confidence mouse NAP genes significant at the IMPC threshold of p < 1 x 10-4 (68.2%) reduce the size of the affected structures, while 20% cause an enlargement of the ventricles, mainly, and 11.8% show dual effects (**Fig 6.3a**). This suggests positive selection towards smaller brain structures. Grouping neuroanatomical parameters, based on the percentage size change relative to WT mice, reveals most structures are very severely affected, with a 727% change in size of lateral ventricles observed in the *Tsga*$^{-/-}$ mutant (**Fig 6.3c**).

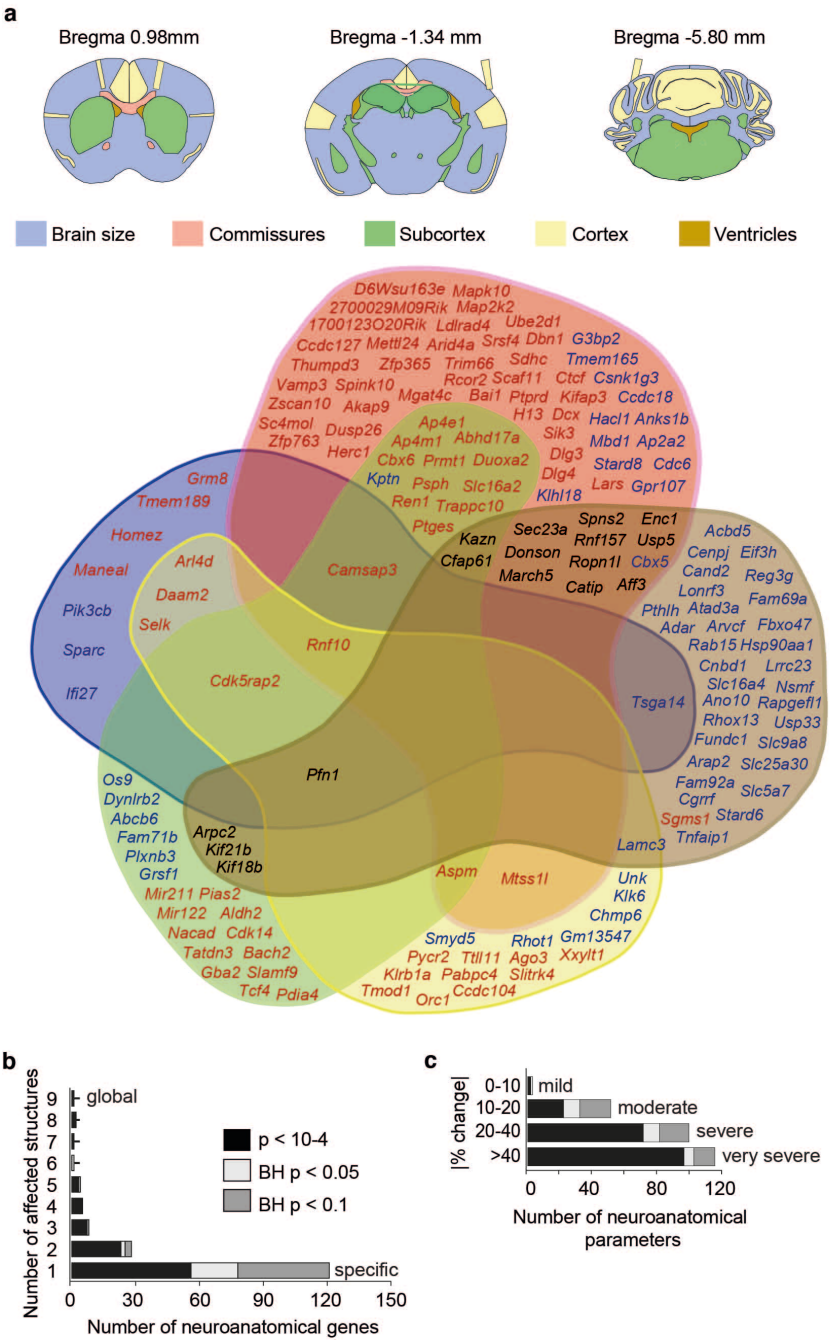


Fig 6.3 | Identification of mouse NAP genes. (**a**) The 78 brain parameters are grouped into five categories and mapped on coronal sections at the indicated Bregma positions. The Venn diagram shows the 158 NAP genes. Genes are mapped on each category and color-coded. Red, blue and black font correspond to decrease or increase in structure size or both, respectively. (**b**) Number of genes affecting, one or more structures. (**c**) Classification of percentage change (absolute value) of neuroanatomical measurements, relative to WT. NAP genes, neuroanatomical phenotype genes, defined as those associated with neuroanatomical deviations in knockout compared to wildtype mice (significant at 10% FDR).

6.4.2 Functional characterisation of mouse NAP genes

To gain a better understanding of the molecular processes essential for healthy brain development and function, we first characterised those genes whose disruption in the mouse yields neuroanatomical defects. We focused on 158 / 1,057 gene disruptions causing significant brain abnormalities compared to WT animals at the FDR threshold of 10% (NAP genes). To ensure a consistent biological signal, we confirmed our findings using NAP genes determined at 5% and 1% FDR thresholds.

Considering all mouse genes with Gene Ontology (GO) annotations, we found that these 158 NAP genes are frequently localised to the *cytoskeleton* (GO:0005856) and *microtubules* (GO:0005874) and involve *cytoskeletal protein binding* (GO:0003779) (BH p = 0.015) (**Fig 6.4a**). Along with a nominal association with *synaptic membrane* (GO:0097060), their functions point to roles in synaptic transmission, as the latter relies on neuronal polarisation, in turn, controlled by the cytoskeleton via regulation of microtubule and actin assembly and organisation (Hoogenraad and Akhmanova, 2010). Conservatively, these enrichments are preserved when using all 1,057 KO genes as a background (**Fig 6.4b**). Concordantly, we also detected an excess of PostSynaptic Density (PSD) genes compared to non-NAP genes (genes with FDR > 10%) (p = 0.017), indicating these roles are specific (**Fig 6.5**). Moreover, we also observed an excess of FMRP target genes (p = 0.0035) (**Fig 6.5**), which have various roles in neurogenesis and synaptic functions (Callan and Zarnescu, 2011; Wang *et al.*, 2012). Indeed, FMRP target genes

encode synaptic membrane components and downstream signalling molecules, whose translation modulates activity-dependent synaptic plasticity. Conservatively, these observations were recapitulated when considering the 75 and 122 NAP genes determined 1% and 5% FDR, respectively (**Figs 6.5, S6.2 and S6.3**).

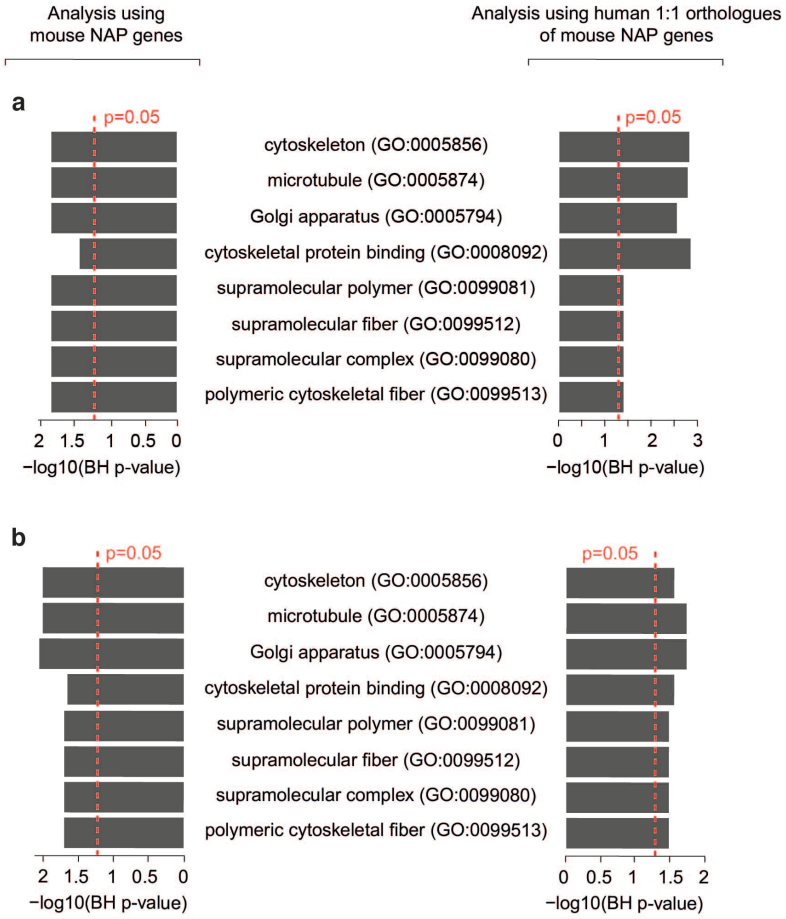


Fig 6.4 | GO enrichment analysis. (**Left side**) NAP genes are enriched for GO terms relating to synaptic and neuronal functions against (**a**) the whole genome and (**b**) KO genes. (**Right side**) Over-represented terms persist when performing the analysis using human 1:1 orthologues. Significance was derived using right-tailed hypergeomtric tests. The x-axes depict the p-values adjusted for the number of GO categories tested using the Benjamini-Hochberg (BH) method. NAP genes, neuroanatomical phenotype genes, defined as those associated with neuroanatomical deviations in knockout compared to wildtype mice (significant at 10% FDR).

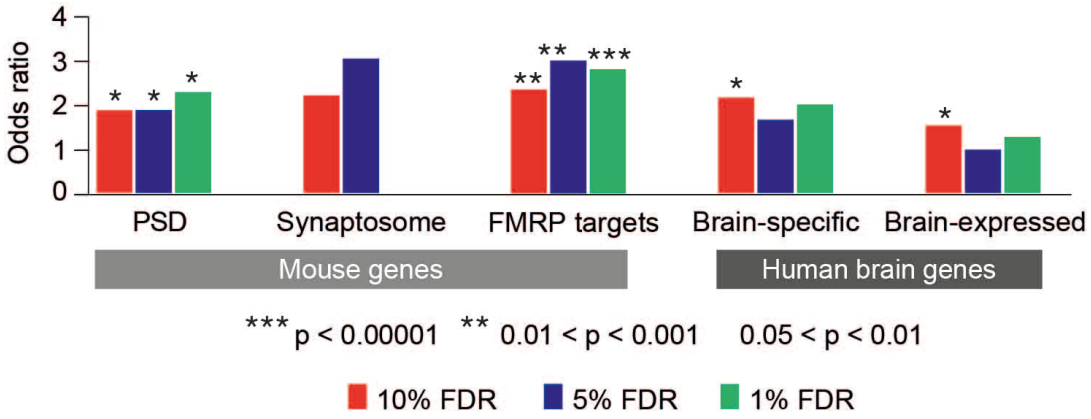


Fig 6.5 | NAP genes are brain genes displaying synaptic functions. Brain, synaptic and FMRP target gene enrichments observed among mouse NAP genes associated with mouse neuroanatomical defects significant at 10%, 5% and 1% FDR, and their human orthologues compared to non-NAP genes (right-tailed Fisher tests). Gene sets are described in **section 2.1** and summarised in **Table 2.1**. NAP genes, neuroanatomical phenotype genes, defined as those associated with neuroanatomical deviations in knockout compared to wildtype mice (significant at 10%, 5% or 1% FDR); PSD, post-synaptic density.

Under the assumption that analogous phenotypes result from the disruption of common biological pathways (Oti and Brunner, 2007), we assessed the likelihood that these NAP genes participate in shared molecular processes. For this, we used a mouse functional gene network, MouseNet v2 (Kim *et al.*, 2016). MouseNet integrates a variety of genomic datasets and resources, such as physical interaction (PPI), expression and functional annotations, in order to infer functional couplings between pairs of mouse genes. Based on 1,000 permutations, the 158 NAP genes exhibit an excess of connections within this network compared to the same number of random genome genes, matched for Coding DNA Sequence lengths (CDS) and network connectivities (empirical p = 0.021) (**Fig 6.6**). While, NAP genes determined at 1% FDR also cluster in MouseNet v2 (p = 0.01), those determined at 5% FDR do not (p = 0.107) (**Fig 6.6; Table S6.1**).

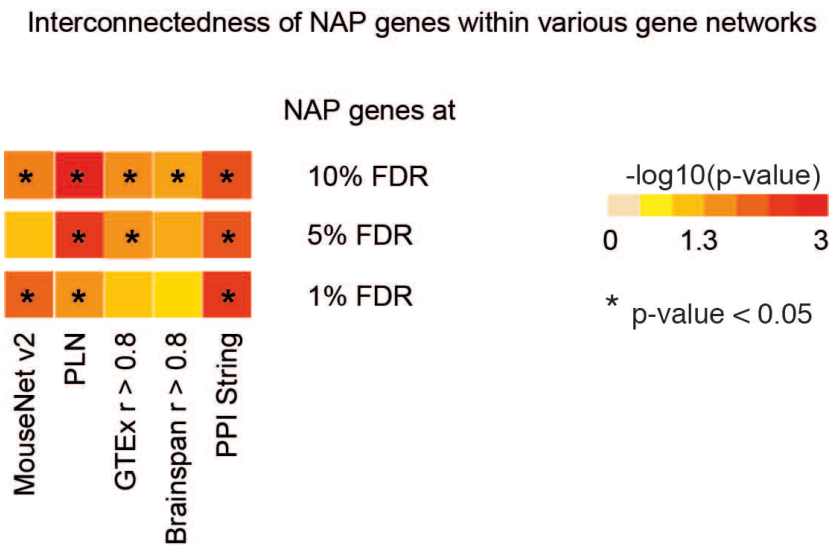


Fig 6.6 | NAP genes are functionally similar and co-expressed. Interconnectedness of mouse NAP genes, determined at 10%, 5% and 1% FDR, and their human orthologues within various gene networks as compared to randomly sampled genome genes, matched for coding DNA sequence length and network connectivity. MouseNet v2 and the PLN (Phenotypic Linkage network) are gene functional networks; GTEx and Brainspan represent gene co-expression network and PPI refers to a protein-protein interaction (PPI) network, retrieved from String. MouseNet v2 uses mouse genes, while the remaining networks use human unique 1:1 orthologues (**Table 2.3**). NAP genes, neuroanatomical phenotype genes, defined as those associated with neuroanatomical deviations in knockout compared to wildtype mice (significant at 10%, 5% or 1% FDR).

NAP genes are not preferentially expressed in the adult mouse brain compared to genome genes of similar lengths. Nevertheless, they are strongly up-regulated in developing CNS, and in particular, during early and mid prenatal development (E11.5-E14) (**Fig 6.7a**).

Using the mouse developing CNS and adult brain gene expression datasets (**section 6.3.9**), we created a co-expression network linking genes whose expression patterns are correlated with r > 0.8 (**section 6.3.11**). Within this co-expression network (i.e. mouse developing CNS and adult brain co-expression network), we observed 2,115 links between 155 NAP genes (**Fig S6.4**). The r > 0.8 threshold was chosen to maximise the number of genes

present in the network, and its modularity; nevertheless, similar results were found when varying the minimum correlation coefficient (not shown). This co-expression sub-network is modular (modularity of network Q = 0.49), and could, thus separated into smaller sub-networks (i.e. modules), such that genes from the same modules display an excess of intra- compared to inter-module network connections. Clustering with the Louvain algorithm revealed that NAP genes associate into 5 separate clusters of genes (named *mouse modules*) comprising between 4 and 51 genes (**Fig S6.4**). Gene expression analysis revealed modules exhibit distinct temporal brain expression patterns (**Fig 6.7b**). Genes from module 0 are broadly up-regulated through mid/late development as well as in the adult brain. Accordingly, while they detain some developmental roles, as evidenced by a nominal GO association with *positive regulation of nervous system development* (GO:0051962; p = 9.36 x 10-5), they generally display mature synaptic functions, with enriched GO terms including *postsynaptic density* (GO:0014096; BH p = 0.018) (**Table S6.2**). Conversely, although their exact roles are unclear (**Table S6.2**), genes from modules 2 and 3 are embryonically expressed and specifically up-regulated in the early and mid/late embryonic mouse CNS, respectively (**Fig 6.7b**). Conservatively, these results are preserved when comparing expression patterns between module genes and remaining NAP genes (**Fig 6.7b**). Taken together, these findings suggest genes, whose disruption causes neuroanatomical abnormalities in mouse models, have essential neurodevelopmental roles. Moreover, they are also likely to also participate in mature brain processes, such as synaptic plasticity.

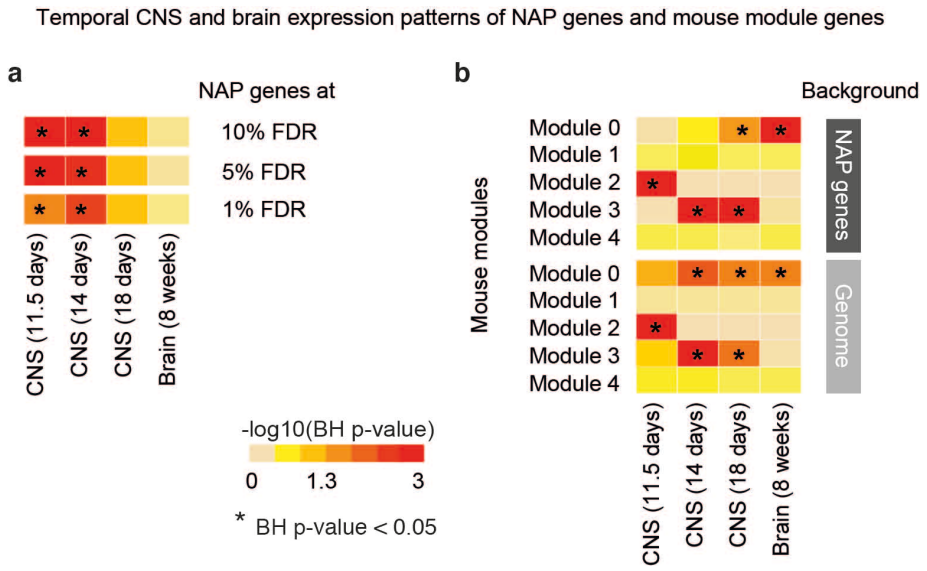


Fig 6.7 | Expression patterns of mouse NAP genes and module genes. (**a**) Expression patterns of NAP genes, determined at 10%, 5% and 1% FDR, in the mouse developing central nervous system (CNS) and postnatal brain compared to randomly sampled genome genes matched for coding DNA sequence length (CDS) to NAP genes. (**b**) NAP gene sub-clustering into modules with distinct spatiotemporal expression dynamics in the mouse embryonic CNS or adult brain co-expression network compared random genome genes of similar CDS or remaining NAP genes. P-values were derived using a permutation-based approach and right-tailed Mann-Whitney U test (**section 2.7.2**), when comparisons were performed using genome genes and remaining NAP genes, respectively. Benjamini-Hochberg (BH) corrections were applied to control for the number of stages tested. NAP genes, neuroanatomical phenotype genes, defined as those associated with neuroanatomical deviations in knockout compared to wildtype mice (significant at 10%, 5% or 1% FDR).

6.4.3 Functional characterisation of human NAP genes

Aiming to employ mouse neuroanatomical data to study ID and brain malformations in humans, we considered the 960 human 1:1 orthologues of the 1,057 mouse KO genes. Of these, 147 are orthologues of the 158 NAP genes and are enriched for the same GO terms, hence supporting a strong functional translation between mouse and human (**Figs 6.4, S6.2 and S6.3**).

The human orthologues of NAP genes are strongly expressed in the human brain, along with a small number of other tissues, including female reproductive organs (cervix, vagina, uterus and fallopian tubes) (**Fig 6.8a**). Aiming to refine the associated brain expression profiles, we evaluated transcriptomic data across six developmental stages and in six brain regions, using the Brainspan dataset. Consistent with observations made using mouse datasets, the human NAP orthologues are strongly up-regulated during early/mid prenatal development compared to genome genes, hence supporting neurodevelopmental roles (**Fig 6.8b**). Additionally, they are strongly expressed throughout infancy and childhood (**Fig 6.8b**) but do not show region-specific expression (**Fig 6.8c**).

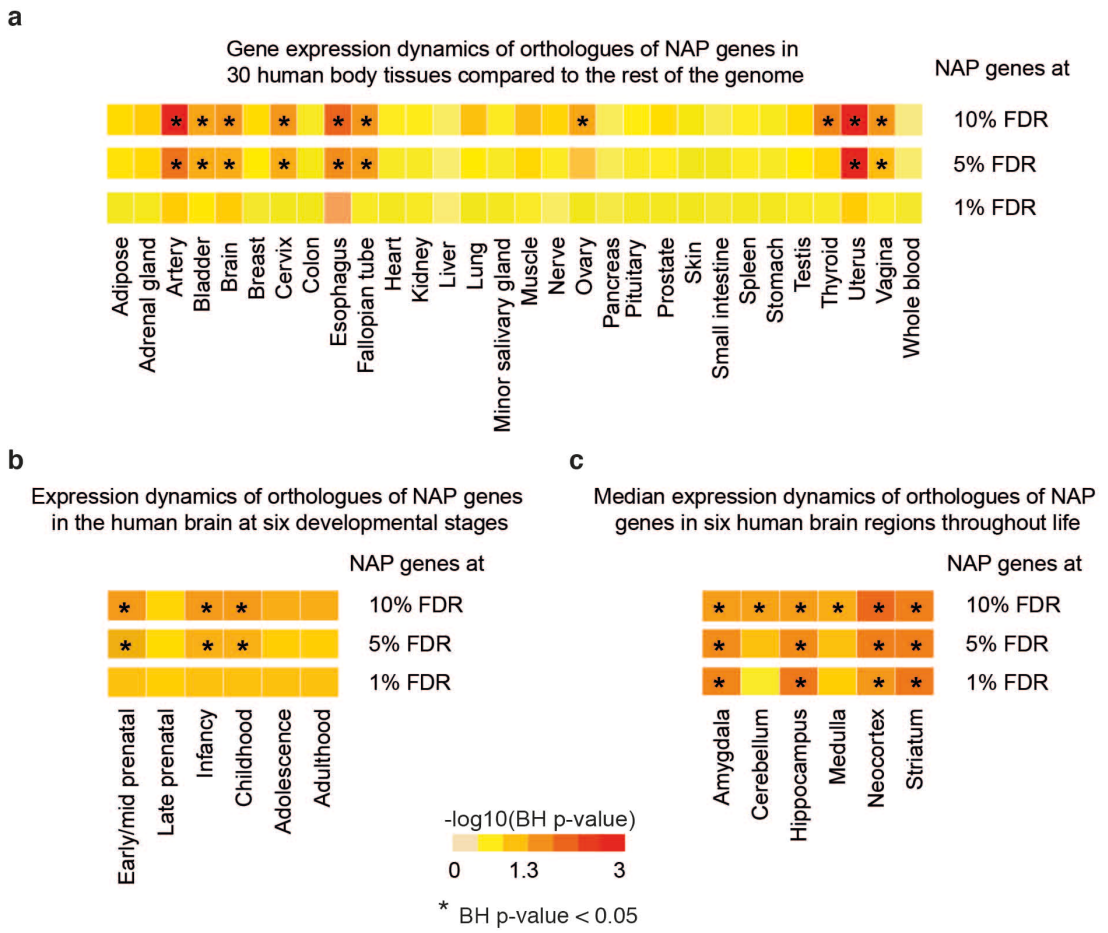


Fig 6.8 | Body-wide and brain expression patterns of human orthologues of mouse NAP genes. (**a**) Expression patterns of human orthologues of NAP genes, determined at 10%, 5% and 1% FDR in (**a**) 30 human body tissues (GTEx dataset) and (**b, c**) in the brain at (**b**) six human developmental stages, and (**c**) in six human brain regions (Brainspan datasets). To derive p-values, expression patterns were compared between orthologues of NAP genes and random genome genes, matched for coding DNA sequence length to NAP orthologues using a permutation-based approach (**section 2.7.2**). Benjamini-Hochberg (BH) corrections were applied to control for the number of tissues, stages or regions tested. NAP genes, neuroanatomical phenotype genes, defined as those associated with neuroanatomical deviations in knockout compared to wildtype mice (significant at 10%, 5% or 1% FDR).

By combining human genomic data-sources, including expression, PPI and functional (i.e. GO, KEGG and Reactome) datasets, we constructed a Phenotypic Linkage Network (PLN), as described in Honti *et al.*, 2014 (**sections 2.4.3 and S2**). The human orthologues of NAP genes show strong clustering within the PLN, suggesting they are functionally more similar to one

another than randomly sampled genome genes (empirical $p < 0.001$) (**Fig 6.6**). This result was reproduced using NAP genes determined at 1% and 5% FDR and persisted when restricting the PLN to only the strongest functional links or considering a PPI network alone (**Fig 6.6; Table S6.1**). As PPI datasets and functional annotations (i.e. GO, MGI, KEGG, Reactome) incorporated in the PLN are prone to ascertainment biases, we repeated the analysis using less biased high-coverage co-expression networks, constructed using data from the GTEx (Lonsdale *et al.*, 2013) and Brainspan (Sunkin *et al.*, 2013) projects (**section 2.4.1**). Considering only gene pairs with co-expression coefficients $r > 0.8$, NAP genes determined at 10% FDR display more similar body-wide and brain spatiotemporal expression patterns than similar genome genes (empirical $p < 0.05$) (**Fig 6.6**). These observations are robust to varying the baseline correlation coefficient in Brainspan but not GTEx (**Table S6.1**), possibly owing to the brain phenotypes associated with these orthologues.

6.4.4 Neuroanatomical functional gene modules

The functional similarity network formed by human orthologues of mouse NAP genes consists of 381 links within the PLN between 121 genes. Given its modularity ($Q = 0.45$), we partitioned this network using the Louvain algorithm. We identified nine gene modules comprising between four and 23 genes, with a median of 12 genes per module (**Figs 6.9 and S6.5**).

While our PLN network incorporates many and diverse functional genomics resources to identify functional convergence, we can examine individual resource contributions for interpretation. Examining GO annotations (**Table S6.3**), we found that genes from module 2 are enriched for cell cycle-related GO annotations, such that they are specifically associated to the G2/M checkpoint (**Table S6.3**), suggesting they participate in early neurodevelopmental processes. Correspondingly, they are nominally up-regulated during early/mid prenatal development ($p = 0.01$), and in the striatum compared to genome genes (**Fig 6.10a**), implicating aberrant neuroproliferation and differentiation of striatum cells. Similarly, genes from module 3 are up-regulated during early/mid prenatal development, nominally in the striatum (**Fig 6.10a**) and detain cell cycle-related functions, as indicated by nominal GO enrichments of *mitotic cell cycle process* ($p = 1.15 \times 10\text{-}4$) and *DNA replication* ($p = 1.66 \times 10\text{-}4$) (**Table S6.3**).

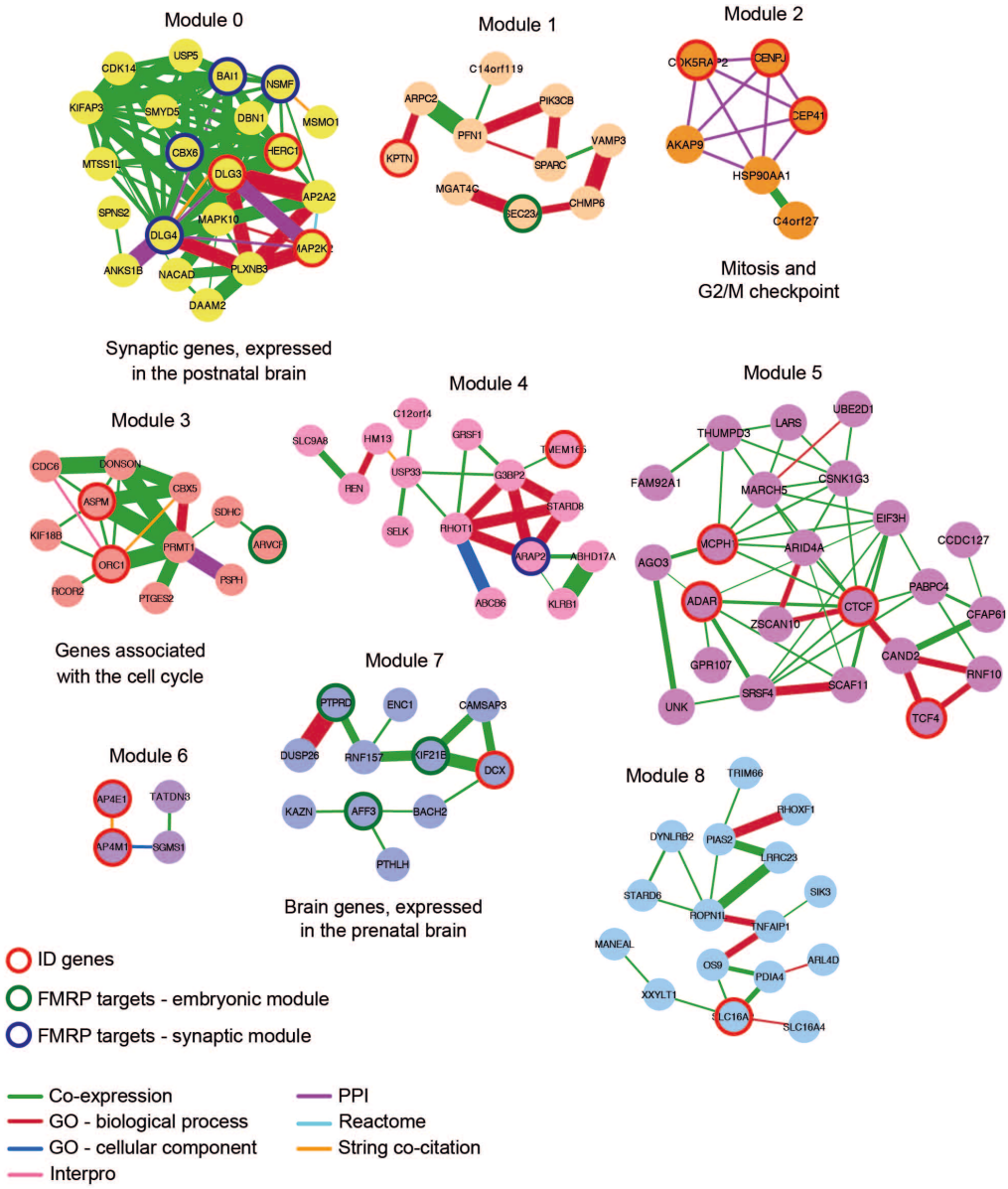


Fig 6.9 | Functional similarity network of human orthologues of mouse NAP genes. Using the PLN (Phenotypic Linkage Network), we identified 381 functional links between 121 human orthologues of NAP genes, which we partitioned into the nine modules of closely-related genes shown (illustrated by the colour of the nodes; see **Fig S6.5** for the unpartitioned network). The thickness of each edge is proportional to the functional similarity score, as given by the PLN, and the colour of each edge indicates the largest contributing information source. Red node borders depict known ID genes, whereas blue and green borders refer to the set of embryonically-expressed and mature FMRP target genes, respectively. Module descriptions were added, where clearly discernable. NAP genes, neuroanatomical phenotype genes, defined as those associated with neuroanatomical deviations in knockout compared to wildtype mice (significant at 10% FDR); GO, Gene Ontology.

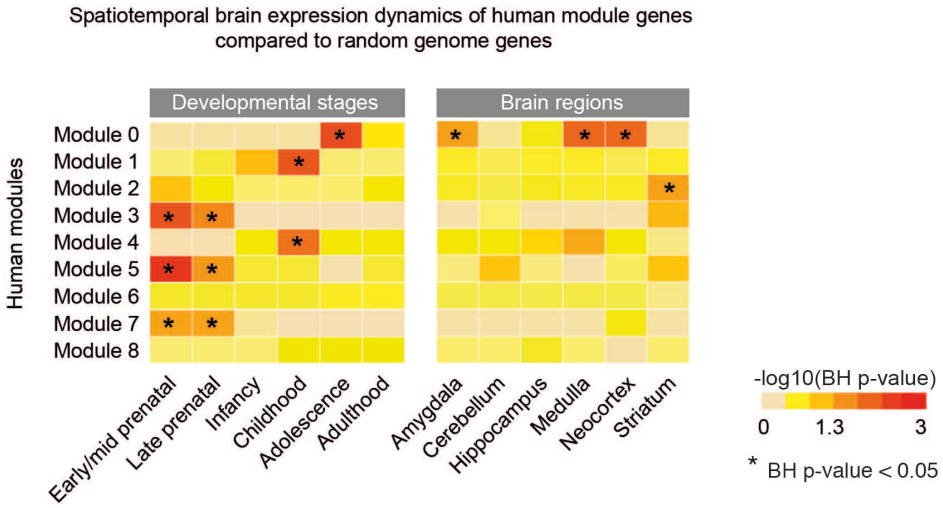


Fig 6.10 | Spatiotemporal brain expression patterns of human module genes. Spatiotemporal gene expression dynamics in the human brain at six developmental stages, and in six regions (Brainspan datasets). Expression dynamics were compared between human module genes and random genome genes, matched for coding DNA sequence length to module genes. P-values were derived using a permutation-based approach (**section 2.7.2**). Benjamini-Hochberg (BH) corrections were applied to control for the number of stages tested.

By contrast, human modules 0 and 7 exhibit an excess of brain-specific ($p = 0.002$ and $p = 0.045$, respectively) and human orthologues of FMRP target genes ($p = 7.81 \times 10\text{-}4$ and $p = 0.030$, respectively) (**Fig 6.11**). Moreover, module 0 is also enriched for human orthologues of mouse PSD and synaptosome genes ($p = 7.81 \times 10\text{-}4$ and $p = 0.019$, respectively) (**Fig 6.11**) and the associated genes are functionally annotated with the GO terms *postsynaptic density* (BH $p = 0.0015$), *cell projection* (BH $p = 0.040$) and *cytoskeleton* (BH $p = 0.026$) (**Table S6.3**). While both modules possess neuronal functions, genes from module 0 are up-regulated postnatally, while those in module 7 are up-regulated in the foetal brain (**Fig 6.10a**), suggesting each module contributes to distinct neurodevelopmental and mature synaptic functions. This is further supported by the observation that genes from modules 7 and 0 are enriched in different subsets of foetally- and postnatally-

expressed FMRP target genes (p = 0.0038 and p = 0.0055, respectively) (**Fig 6.11**). We note that while genes from module 5 are also up-regulated during prenatal development (**Fig 6.10a**), we are unable to assign them with specific functionalities, given the lack of interpretable GO enrichments (**Table S6.3**) and particular gene sets (**Fig 6.11**).

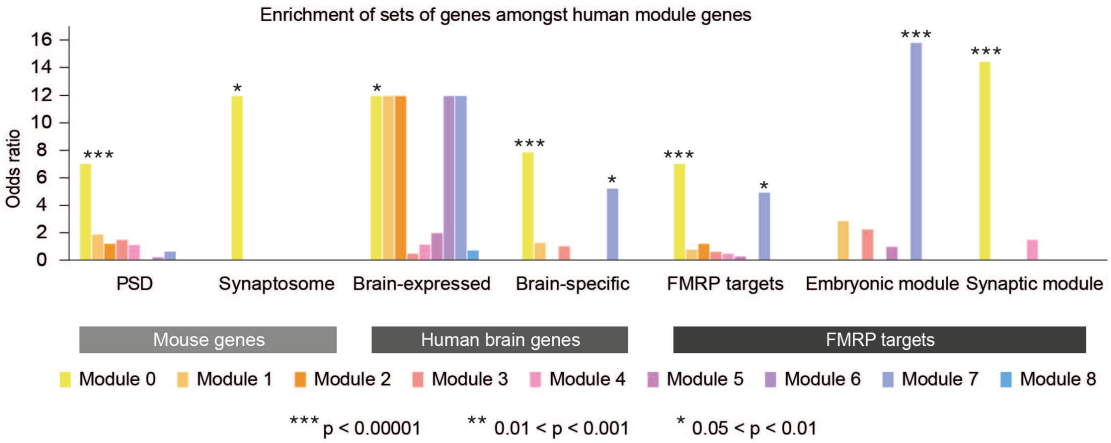


Fig 6.11 | Gene set enrichment across human modules. Brain, synaptic and FMRP target gene enrichments observed in human module genes compared to remaining NAP genes (right-tailed Fisher tests). Gene sets are described in **section 2.1** and summarised in **Table 2.1**. NAP genes, neuroanatomical phenotype genes, defined as those associated with neuroanatomical deviations in knockout compared to wildtype mice (significant at 10% FDR); PSD, post-synaptic density; Embryonic module, subset of embryonically-expressed FMRP target genes; Synaptic module, subset of FMRP target genes expressed during postnatal life.

Despite the functional convergence of genes within modules, no single neuroanatomical defect is enriched amongst mouse KOs of orthologous module genes (not shown). However, when scoring the overall similarity of neuroanatomical abnormalities across Bregma sections 1 and 2 using a simplified version of the *Goodall3* index (**section 6.3.1**) (Boriah *et al.*, 2008), phenotypic convergence was observed for modules 2, 7 and 8 (**section 6.3.7**). Specifically, mouse orthologues of modules 2 and 7 yielded more similar brain abnormalities than other NAP genes across Bregma +0.98mm

but not Bregma -1.34mm, while the opposite pattern was found for module 8 orthologues (**Fig 6.12a**). Correspondingly, while genes from human modules 2 and 7 are more strongly expressed across tissues mapping to Bregma +0.98mm than Bregma -1.34mm, those from module 8 are preferentially expressed across Bregma -1.34mm tissues (**Fig 6.12b**). Taken together, our results are consistent with the observation that disruption of individual neurodevelopmental pathways in human yield pleiotropic but distinct patterns of phenotypic consequences (Andrews *et al.*, 2015b).

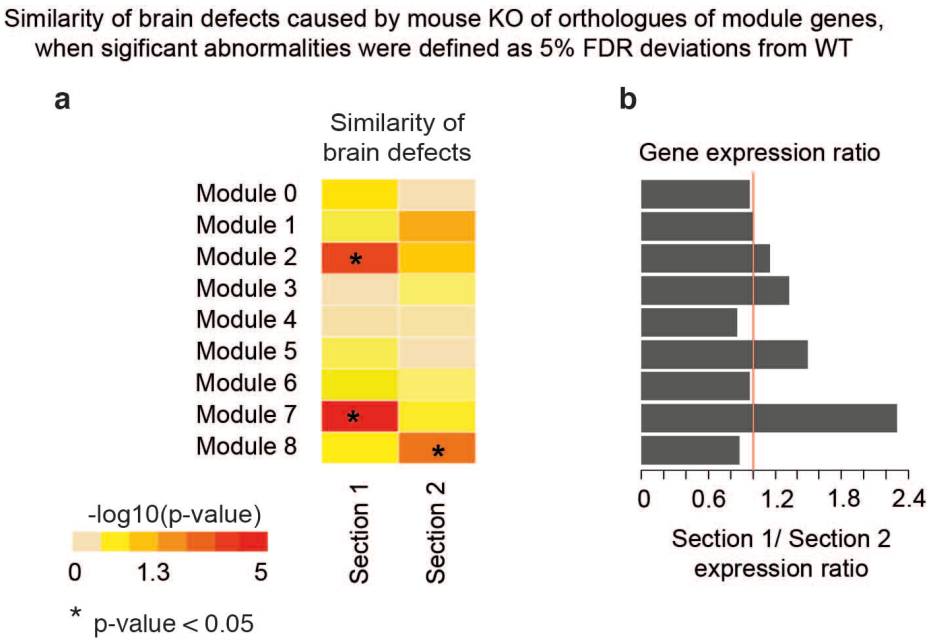


Fig 6.12 | Similarity of neuroanatomical defects caused by disruption of human module genes. (**a**) The heatmap depicts the similarity of brain abnormalities caused by module genes in the critical sections 1 and 2 (Bregma +0.98mm and -1.34mm, respectively). (**b**) The barplot shows the normalised section 1 and section 2 expression ratio, based on Brainspan data (**section 6.3.10**).

6.4.5 Functional translation between mouse and human

Based on the consistency of the GO and transcriptomic findings using mouse and human genes, this study supports the notion of functional translation between the two species. To further validate our findings, we examined the coherence of the uncovered human PLN module and mouse modules identified using the developing CNS and adult brain gene co-expression network. Considering the 960 mouse genes with 1:1 human orthologues, we confirmed similar modules were identified, such that genes from human modules are not over-represented in more than one mouse module (**Fig 6.13**). For example, human module 0 and mouse module 0 overlap at a 1:1 ratio and comprise postnatal synaptic genes (**Figs 6.13, 6.7b, 6.10 and S6.6; Tables S6.2 and S6.3**). Moreover, human modules 2, 3 and 5 involved in neurodevelopmental processes overlap mouse module 2, which correspondingly includes genes expressed during early CNS development (**Figs 6.13, 6.7b, 6.10 and S6.6**). Lastly, genes from mouse module 3, up-regulated close to birth, are over-represented in human modules 4 and 7, which comprise genes expressed in the pre- and post-natal brain, respectively (**Figs 6.13, 6.7b, 6.10 and S6.6**). Of note, comparisons between brain expression dynamics between human and mouse modules are only qualitative, given the inevitable discrepancies between embryonic and postnatal mouse and human stages.

Visual inspection of neuroanatomical gene networks suggests modules are more explicitly uncovered using the PLN than using the developing CNS and

adult brain mouse co-expression network (**Figs 6.9, S6.4 and S6.5**). Accordingly, the inference of gene couplings based on multiple lines of genomic evidence enables a clearer functional delineation between modules. This view is supported by the observation that mouse and human modules overlap at many-to-one ratios (**Fig 6.13**) and that PLN modules display more subtle functions and expression patterns (**Figs 6.7b, 6.10 and S6.6; Tables S6.2 and S6.3**). Inevitably, this could also be a result to a suboptimal choice of brain developmental stages to interrogate and the limited number of CNS tissues collected. Indeed, given the relative scarcity of mouse compared to human genomics datasets, human data can be used to provide more comprehensive biological views.

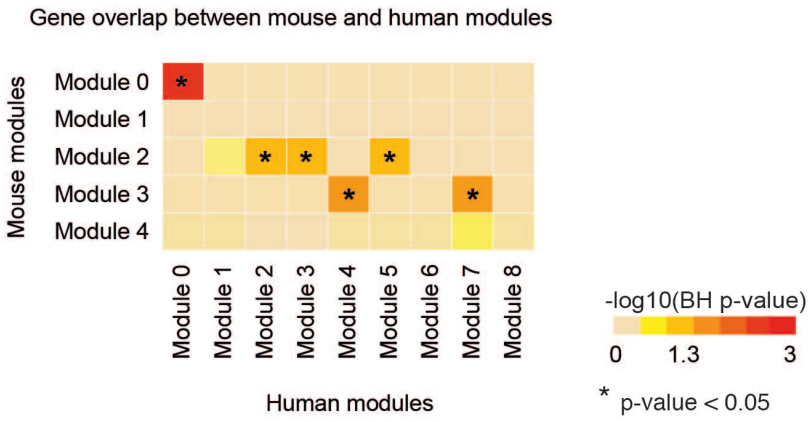


Fig 6.13 | Overlap between mouse and human gene modules. Overlap between mouse gene modules (identified using a gene co-expression network, derived from mouse developing CNS and adult brain expression data) and human PLN (Phenotypic Linkage Network) modules. Overlaps were tested using right-tailed Fisher tests. Corrections for multiple testing were performed using the Benjamini-Hochberg (BH) method to control for the numbers of modules tested.

6.4.6 Neuroanatomical deficits and gene deleteriousness

To examine the relevance of genes associated with strong mouse neuroanatomical phenotypes in human disease, we compared the features of human unique (1:1) orthologues of genes whose disruption in mouse models yielded the largest neuroanatomical variation to those showing the least. Specifically, we compared the top 10% to the bottom 10% of a total of 1,057 mouse KO genes (or the corresponding 960 1:1 human orthologues) and evaluated differences to permutated gene sets. Genes associated with the strongest neuroanatomical variation (1) undergo strong negative selection (dN/dS) (**Figs 6.14a**). Moreover, their 1:1 human orthologues (2) have relatively high selection coefficients (s_{het}; Cassa *et al.*, 2017), (3) are depleted of common genetic variation (RVIS; Petrovski *et al.*, 2013), (4) are predicted to be intolerant to protein-truncating mutations (pLI; Lek *et al.*, 2016) and (5) are more likely to be HaploInSufficient (HIS) (Huang *et al.*, 2010) (**Fig 6.14b**). More generally, we found a significant correlation between these four measures and the strength of brain abnormalities induced by gene disruptions ($p = 0.0394$, $p = 0.049$, $p = 6.7 \times 10\text{-}4$, $p = 0.028$ and $p = 6 \times 10\text{-}4$ for dN/dS, s_{het}, HIS, RVIS and pLI, respectively; linear regression). While not enriched in known human disease genes (**Fig 6.14c**), the properties of the human orthologues of these mouse genes suggest it would not be unsurprising to discover a role for many in human disease.

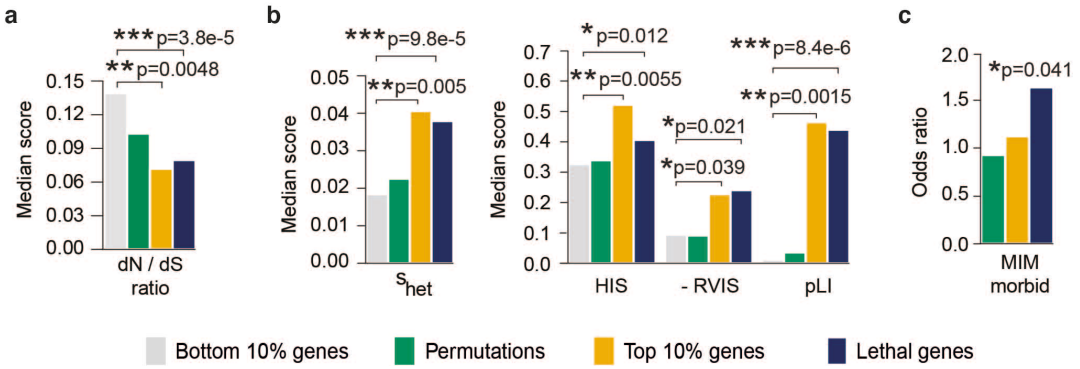


Fig 6.14 | Predicted gene deleteriousness in humans is associated with stronger mouse neuroanatomical abnormalities. (**a**) Mouse lethal genes (Dickinson *et al.*, 2016) and the top 10% of mutants with the largest neuroanatomical abnormalities are compared to the bottom 10% genes and to randomly permuted sets of these genes (**section 6.3.6**). The dN/dS ratio examines selection pressures by comparing the rate of synonymous and non-synonymous substitutions. (**b**) Comparison of the gene properties of the human orthologues of lethal genes (Dickinson *et al.*, 2016) and human orthologues of the top 10% of mutants with the largest neuroanatomical abnormalities as compared to genes in the bottom 10% and to randomly permuted sets of these genes. s_{het} (selection coefficient associated with the loss of heterozygosity), RVIS (residual variation intolerance ccore) and pLI (probability of being loss of dunction Intolerant) are all indicators of purifying selection pressure, while HIS are predicted haploinsufficiency scores (**section 6.3.8**). (**c**) The overlap between gene sets from (**a**) and (**b**) with MIM Morbid Map genes.

6.4.7 NAP genes and human disease

A substantial number of NAP genes have been implicated in neurodevelopmental disorders, consistent with their predicted developmental and synaptic roles. In particular, we found an excess of ID genes from three independently curated gene lists (Casanova *et al.*, 2016; Kochinke *et al.*, 2016; Pinto *et al.*, 2014) among human orthologues of NAP genes, determined at 10% FDR ($p < 0.001$) (**Fig 6.15a**). Specifically, module 2 comprises a substantial proportion of these genes ($0.009 < p < 0.046$) (**Fig 6.15b**), consistent with the implication of early developmental and gene expression regulators in the pathogenesis of ID. Conversely, genes implicated in other brain disorders, such as epilepsy, hyperactivity or schizophrenia,

were not enriched among human NAP orthologues (not shown). Nevertheless, while no excess of autism genes was found, we observed an over-representation of a subset of ID genes comorbid with ASD ($p = 0.0015$ and $OR = 8.50$; right-tailed Fisher test) (Casanova *et al.*, 2016). The excess of functional PLN (or co-expression; not shown) links between genes from these sets and NAP genes further corroborates the link between neuroanatomical orthologues and ID (**Figs S6.6**).

Amongst the 147 human orthologues of NAP genes determined at 10% FDR, 28 (19%) overlap with ID associated genes (Kochinke *et al.*, 2016), including 17 (*AP4E1*, *AP4M1*, *ASPM*, *CDK5RAP2*, *CENPJ*, *CEP41*, *CTCF*, *DCX*, *DONSON*, *HERC1*, *KPTN*, *LAMC3*, *MCPH1*, *ORC1*, *SLC16A2*, *TCF4* and *TMEM165*) linked to highly penetrant structural abnormalities in a large number of patients (**Fig 6.15c**). Amongst these 17, only 5 mouse models have been previously reported (*ASPM*, *CENPJ*, *DCX*, *HERC1* and *LAMC3*), such that the remaining 12 represent new *in vivo* mouse models.

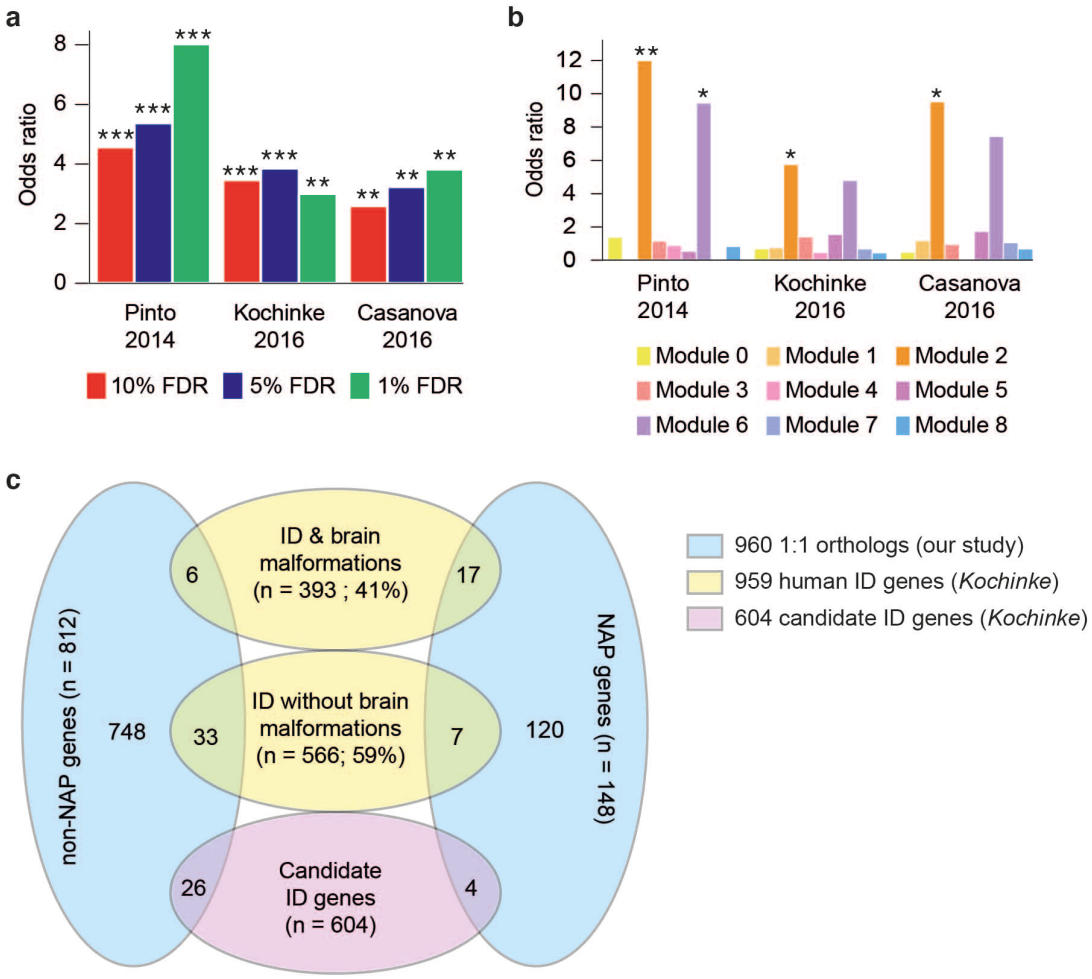


Fig 6.15 | NAP genes overlap ID genes. (**a, b**) Enrichments of ID genes from three independent sets of ID genes (Pinto *et al.*, 2014; Kochinke *et al.*, 2016; Casanova *et al.*, 2016; **section 2.1**; **Table 2.1**) (**a**) among human orthologues of NAP genes compared to human orthologues of non-NAP genes, and (**b**) among human module genes (**Fig 6.11**) compared to remaining NAP orthologues (right-tailed Fisher tests). (**c**) Comparative study between our study and Kochinke *et al.*, 2016. NAP genes, neuroanatomical phenotype genes, defined as those associated with neuroanatomical deviations in knockout compared to wildtype mice (significant at 10% FDR).

Of the 28 ID-associated genes present among human orthologues of NAP genes, 19 have been annotated in OMIM with accompanying cranial malformation (**Table 6.1**). Mutations in the *ADAR* and *ANO10* genes cause Aicardi-Goutières syndrome (MIM:615010) and cerebellar ataxias (MIM:613728), which are marked by cerebral and cerebellar (Vermeer *et al.*, 2010) atrophies, respectively. *CEP41* is associated with Joubert syndrome

(MIM:614464), which is characterised by brainstem malformations and underdevelopment of the cerebellar vermis (Brancati *et al.*, 2010). Mutations in *LAMC3*, encoding a protein critical for the formation of cerebral cortex convolutions (Barak *et al.*, 2011), cause cortical malformations (MIM:614115). Similarly, mutations *DCX*, encoding a doublecortin critical for neuronal migration, induce lissencephaly (MIM:300067). Many of these ID genes are also linked to brain size abnormalities. For example, the *AP4E1*, *AP4M1* genes encoding adaptor complex subunits (spastic paraplegia, MIM:213744 and MIM:612936, respectively) as well as *ORC1* (Meier-Gorlin syndrome, MIM:224690) and *TMEM165* (congenital glycosylation disorder, MIM:614727) are associated with conditions which present with microcephaly amongst other features (Moreno-De-Luca *et al.*, 2011; Jamra *et al.*, 2011; Shalev and Hall, 2003; Bicknell *et al.*, 2011; Foulquier *et al.*, 2012). Additionally, *ASPM*, *CDK5RAP2* and *CENPJ*, encoding proteins essential for mitotic spindle formation during neurogenesis as well as *MCPH1*, encoding a DNA damage protein, are known causes of microcephaly (MIM:608716, MIM:604804, MIM:613676 and MIM:251200, respectively). Lastly, *KPTN*, enconding a protein involved in neuromorphogenesis, has been linked to macrocephaly (Baple *et al.*, 2014; Pajusalu *et al.*, 2015).

Table 6.1 | ID associated genes among human NAP orthologues.

Gene	OMIM human disease MIM	Description	Inheritance	Gene description
ACBD5	NA	NA	NA	Acyl-CoA-binding protein with roles in in cell metabolism and gene regulation
ADAR	615010	Aicardi-Goutieres syndrome 6	AR	Regulation of gene expression vis A-to-I RNA editing.
	127400	Dyschromatosis symmetrica hereditaria	AD	
ANO10	613728	Spinocerebellar ataxia, autosomal recessive 10	AR	Calcium-activated chloride channel.
AP4E1	613744	Spastic paraplegia 51, autosomal recessive	AR	Subunits of adaptor protein complex-4, involved in targeting proteins from the trans-Golgi network to the endosomal-lysosomal system.
	184450	Stuttering, familial persistent, 1	AD	
AP4M1	612936	Spastic paraplegia 50, autosomal recessive	AR	
ASPM	608716	Microcephaly 5, primary, autosomal recessive	AR	Involved in neurogenesis by regulation mitotic spindle functions in embryonic neuroblasts.
C12orf4	NA	NA	NA	cytoplasmic protein that is involved in mast cell degranulation
CDK5RAP2	604804	Microcephaly 3, primary, autosomal recessive	AR	Plays a role in cell division as a regulator of cyclin-dependent kinase 5 activity.
CENPJ	613676	Seckel syndrome 4	AR	Plays a role in cell division and centrosome function via maintenance of centrosome integrity and normal spindle morphology.
	608393	Microcephaly 6, primary, autosomal recessive	AR	
CEP41	614464	Joubert syndrome 15	AR	Centrosomal and microtubule-binding protein, involved in cell division.
CTCF	615502	Mental retardation, autosomal dominant 21	AD	Transcriptional regulator. It binds chromatin insulators which prevent interaction between promoters and enahnacers/silencers, thus acting as both a transcriptional activator and repression.
DCX	300067	Lissencephaly, X-linked	XL	Doublecortin protein, which binds microtubules. By regulating their organisation and stibility, it directs neuronal migration in the developing cortex.
	300067	Subcortical laminal heteropia, X-linked	XL	
DLG3	300850	Mental retardation, X-linked 90	XLR	Membrane-associated guanylate kinase protein from the MAGUK family, involved in clustering of NMDA receptors at excitatory synapse.
DLG4	NA	NA	NA	
DONSON	NA	NA	NA	Unclear function but appear to have roles in cell proliferation during development
GBA2	614409	Spastic paraplegia 46, autosomal recessive	AR	Putative transmembrane protein with roles in carbohydrate transport and metabolism
HACL1	NA	NA	NA	Involved in the catalysis of carbon-carbon cleavage reactions
HERC1	617011	Macrocephaly, dysmorphic facies, and osychomotor retardation	AR	HERC protein involved in membrane transport processes
KPTN	615637	Mental retardation, autosomal recessive 41	AR	Filamentous-actin-associated protein involved in neuromorphogenesis.
LAMC3	614115	Cortical malformations, occipital	AR	Glycoprotein localising to basement membranes involved in regulation of cell migration, attachment and organisation. It plays a role in forming the convolution of the cerebral cortex, as suggested by high expression in the human fetal brain.
MAP2K2	615280	Cardiofaciocutaneous syndrome 4	NA	ERK/MAP kinase kinase gene.
MAPK10	NA	NA	NA	MAP kinase with roles in a wide variety of cellular functions
MCPH1	251200	Microcephaly 1, primary, autosomal recessive	AR	DNA damage response protein acting at the G2/M cell cycle checkpoint, during neurogenesis.
MTSS1L	NA	NA	NA	Protein involved in plasma membrane dynamics
ORC1	224690	Meier-Gorlin syndrome 1	AR, AD	Component of the origin recognition complex, essential for initiation of DNA replication. Additionally it is involved in the control of cellular centriole and centrosome copy number.
SLC16A2	300523	Allan-Herndon-Dudley syndrome	XLD	Thyroid hormone transporter, with an important role in nervous system development.
TCF4	613267	Corneal dystrophy, Fuchs endothelial, 3	AD	Transcription factor with an important function in initiation of neuronal differentiation.
	610954	Pitt-Hopkins syndrome	AD	
TMEM165	614727	Congenital disorder of glycosylation, type IIk	AR	Transmembrane protein, which may function as a calcium/proton trnasporter. It is involved in terminal Golgi glycosylation.

Human orthologues of our NAP are not preferentially affected by *de novo* copy-number (*dn* CNVs) and/or single nucleotide (*dn* SNVs) variants in patients with developmental disorders (ASD, ID and/or Developmental Delay – DD) compared to non-NAP genes (we examined the ASD Iossifov, ID DeLigt, ID Hamdam, Nijmegen and DECIPHER datasets; **section 2.1.4; Table 2.1**) (not shown). Nevertheless, the 119 (81%) human NAP orthologues that did not overlap ID associated genes, provide a rich new resource to dissect CNV regions and SNVs associated to brain defects down to single gene resolution (**Table 6.2**). For example, *ARVCF,* localized to the 22q11.21 region deleted in the velocardiofacial syndrome (MIM:192430), has been linked to schizophrenia (Sanders *et al.*, 2005). Our study associates disruption of *Arvcf* with structural abnormalities in the striatum (smaller), while in the Simons Simplex Collection (SSC) (Fischbach and Lord, 2010), we identified a *de novo* nonsense mutation affecting *ARVCF* in an ASD patient presenting with microcephaly. Moreover, seven patients with DD or ID, recruited by the Department of Human Genetics at the Radboud University, Nijmegen, Netherlands (Nijmegen dataset; **Table 2.1**; Vulto-van Silfhout *et al*. 2013a; Andrews *et al.*, 2015b) display *dn* CNVs affecting *ARVCF*, of which two have microcephaly and two facial abnormalities (**Table 6.2**). Additionally, 26 patients from the Database of GenomiC varIation and Phenotype in Humans using Ensembl Resources (DECIPHER) (Bragin *et al.*, 2014; Firth *et al.*, 2009) with *dn* CNVs affecting *ARVCF* present with craniofacial abnormalities. Specifically macrocephaly, pachygyria (involving the under-development of brain gyri; also called incomplete lissencephaly), basal ganglia calcification and general craniosynostosis occur in one individual each, agenesis of the

corpus callosum and cortical dysplasias are observed in two individuals each and microcephaly and plagiocephaly (i.e. asymmetrically flattened skull) are each present in three patients. More generally, 24 patients have facial abnormalities, including cleft palate, microtia and other morphological ear defects, facial shape abnormalities, nasal ridge defects and others. Beyond this example, four other genes (*AKAP9*, *BAI1*, *CTCF* and *SIK3*) are affected by *de novo* loss of function variants in SSC patients, and an additional 14 by multigenic CNVs in a neurodevelopmental cohort described in Vulto-van Silfhout *et al.*, 2013a (**Table 6.2**). As many of these patients are recorded as having craniofacial abnormalities, these mouse models might direct future studies towards a causal understanding.

Table 6.2 | List of genes mapping to ID, ASD and/or DD patients with cranial abnormalities.

Gene	Patients with ID, ASD and/or DD			Gene description
	Dataset	#	Description	
ABHD17A	Nijmegen	1	1 patient with severe ID, microcephaly and seizures	• Hydrolase
AGO3	Nijmegen	1	1 patient with impaired language development and several morphological facial abnormalities	• Member of the Argonaute family of proteins with a role in RNA-mediated gene silencing.
AKAP9	ASD Iossifov	1	1 patient with ASD and anxiety disorder	Scaffolding gene, important for synaptic transmission. Mutations in this gene cause Long syndrom (MIM:611820)
ARVCF	Nijmegen	7	6 patients with severe or moderate ID 4 patients with impaired language development 5 patients with facial abnormalities 2 patients with microcephaly 1 patient with lissencephaly, ventriculomegaly, seizures and autism 1 patient with brachycephaly	• Catenin involved in the formation of adherens junction complexes. • Gene deleted in the Velo-Cardio-Facial syndrome (22q11 microdeletion), characterised by a variable phenotypic spectrum with abnormalities including cleft palate, heart defects and facial dysmorphology.
	ASD Iossifov	1	*De novo* nonsense mutation in ASD patient with microcephaly and language delay	
BAI1	ASD Iossifov	1	-	• Receptor enabling the engulfment of apoptotic cells. Functions as an angiogenesis inhibitor in the brain.
CBX5	Nijmegen	1	1 patient with ID and facial abnormalities	• Heterochromatin protein involved in gene silencing
CTCF	ASD Iossifov	1	1 patient with language delay	• Chromatin regulator, that acts as both a transcriptional activator and repressor.
DBN1	Nijmegen	2	2 patients with severe ID, scaphocephaly, macrocephaly and dolichocephaly 1 patient with impaired language development	• Member of the drebrin family involved in cell migration, extension of neuronal processes and plasticity of dendrites. • DBN1 is a cytoplasmic actin-binding protein with a role in neuronal growth in the developing brain.
FAM71B	Nijmegen	1	1 patient with ID and ASD	• Possible role in RNA biogenesis.
HM13	Nijmegen	1	1 patient with ID, impaired language development and macrocephaly	• Integral membrain protein involved in processing the hepatitis C virus core protein.
KAZN	Nijmegen	1	1 patient with brachycephaly, dolichocephaly and additional morphological facial abnormalities	• Involved in desmosome assembly, cell adhesion, cytoskeletal organization, and epidermal differentiation
MANEAL	Nijmegen	1	1 patient with impaired language development and several morphological facial abnormalities	• Hydrolase
MAP2K2	Nijmegen	1	1 patient with severe ID, microcephaly and seizures	• MAPK kinase kinase protein cause cardiofaciocutaneous syndrome 4 (MIM:615280)
PTHLH	Nijmegen	1	1 patients with morphological abnormality of the cerebrum	• Neuroendocrine peptide involved in bone development, causing brachydactyly type E2 (MIM:613382)
RHOT1	Nijmegen	2	2 patients with ID and impaired language development	• Mitochondrial GTPase with a role mitpchondrial homeostasis and apoptosis
SGMS1	Nijmegen	1	1 patient with ID and macrocephaly	• Transmembrane protein, predominantly expressed in the brain, involved in the synthesis of the sphingolipid, sphingomyelin.
SIK3	ASD Iossifov	1	-	• AMPK protein kinase
TCF4	Nijmegen	2	2 patients with severe ID and impaired language development 1 patient with skull asymetry 1 patient with ventriculomegaly, ASD and seizures	• Transcription factor • Causes corneal dystrophy, Fuchs endothelial 3 (MIM:613267) • Causes Pitt-Hopkins syndrome (MIM:610954)
TNFAIP1	Nijmegen	1	1 patient with ID	• Part of the E3 ubiquitin-protein ligase complex involved in regulation of cytoskeleton structure

Nijmegen refers to the set of ID and/or DD patients with *de novo* CNVs, described in Vulto-van Silfhout *et al.*, 2013a. ASD Iossifov refers to SSC patients with ASD that harbour *de novo* loss of function (nonsense, frameshift, splice-site) variants (**section 2.1.4; Table 2.1**).

6.5 Discussion

In an effort to generate a functional map of all genes from the mouse genome, high throughput mutagenesis screens accompanied by extensive and systematic phenotyping have been undertaken under the Sanger MGP initiative. This study demonstrates the utility of such large-scale data extends beyond the construction of a reference phenotypic resource of mutant alleles, and into the study of human disease. Specifically, we highlight the relevance of mouse neuroanatomical phenotypes in identifying genes whose disruption in humans can induce brain disorders, such as ID, and malformations, such as microcephaly.

Resources such as the mouse GO database enabled us to implicate neuronal and synaptic dysfunction in the pathogenesis of neuroanatomical mouse defects. Moreover, using developmental brain transcriptional gene profiles, generated by the Encode project (ENCODE Project Consortium, 2012; Pervouchine *et al.*, 2015), we showed such genes generally participate in neurodevelopmental processes. The incompleteness, biases and relative scarcity of auxiliary comprehensive mouse genomic datasets, however, hindered further characterisation of the associated genes. For example, the network analysis using MouseNet v2 shows limitations relating to low reliability. The reasons for this are three-fold: (1) a substantial fraction of links consist of homology and phylogenetic inferences, (2) most GO annotations represent computational predictions and (3) close to 5,000 mouse genes lack any type of data (Kim *et al.*, 2016). Additionally, a limited number of brain

tissues have been examined at few developmental points to generate the developing CNS and adult brain transcriptomic mouse gene profiles generated (Pervouchine *et al.*, 2015). Distinctly, the Brainspan project provides a more comprehensive and systematic examination of human gene expression across 27 brain tissues spanning twelve developmental stages (Sunkin *et al.*, 2013). As a result, our analysis principally exploited large-scale transcriptomic, proteomic and functional human resources, under the expectation of biological translation between species. This assumption aligns with the view that genes with conserved unique orthologues are inclined to share function (the *orthologue conjecture*; Dolinski and Botstein, 2007). Furthermore, it is supported, in our context, by the observation that mouse neuroanatomical deficits are informative of human gene deleteriousness and hence disease (**Fig 6.14**). Additionally, mouse NAP genes and their human orthologues converge at the functional (**Figs 6.4 to 6.6**) and transcriptomic levels (**Figs 6.6 to 6.8**). Taken together, these considerations demonstrate the translational potential of research between human and mouse, suggesting the integration of multi-species genomics data enables a more comprehensive understanding of human disease.

This study makes use of systematically phenotyped mouse KO lines to identify 158 genes inducing morphological brain abnormalities. Many of these NAP genes have been implicated in ID and human brain malformations (**Figs 6.14 and 6.15; Table 6.1; section S1**). For example, microcephaly, a condition linked to aberrant neuroproliferation, has been associated with NAP genes encoding DNA damage response (i.e. *MCPH1*), centrosomal (i.e.

CENPJ and *CDK5RAP2*) and mitotic spindle (i.e. *ASPM*) proteins. Many of the uncovered genes are also known causes of lissencephaly, a cortical malformation resulting from improper neuroprogenitor and neuronal migration. Multiple pathogenic mechanisms have been uncovered. These include (1) dysfunctional neuronal pulling forces mediated by processes of cytoskeleton organisation and remodelling, which have been associated with mutations in *DCX*. Additionally, the migratory defects observed in lissencephaly can be caused by (2) the neuronal failure to adhere to destined trajectories, which is facilitated by physical scaffolding structures, hence implicating extracellular matrix components, such as laminins encoded by the *LAMC3* gene. In line with the strong co-occurrence of brain malformations with ID, we additionally observed high confidence and known ID genes. These include the *CTCF* gene, encoding a transcriptional regulator, and *DLG3*, encoding a membrane receptor subunit mediating synaptic plasticity at excitatory synapses. These findings validate the use of mouse neuroanatomical models as an effective approach to uncover novel causes for these conditions and study the underlying pathogenic mechanisms. In this respect, many NAP genes are observed among ID and ASD patients with cortical malformations, supporting their implication in disease (**Table 6.2**). In particular, the *ARVCF* gene, encoding a catenin at the 22q11 locus, deleted in the Velo-Cardio-Facial syndrome, maps to *dn* CNVs from many NDD patients presenting with brain morphological anomalies. Additionally, a penetrant nonsense mutation in *ARVCF* has been detected in an ASD patient with microcephaly. The protein encoded by this gene is likely to play a role in synaptogenesis and synaptic plasticity by mediating the adhesion of pre- and post-synaptic structures.

Hence, *ARVCF* represents a good candidate, warranting its study in the aetiology of NDDs and accompanying malformations. More generally, our *in vivo* mouse models show a great capacity for novel disease gene discovery with up to 81% (119) of our NAP genes not yet linked to any cognitive disorders, constituting a new pool of candidate genes testable against human diseases. Several of the NAP mouse models (for example *Kptn*$^{-/-}$, *Donson*$^{+/-}$ or *Aff3*$^{-/-}$) already turned out to be associated with causative mutations in recent human studies (Baple *et al.*, 2014; Moore *et al.*, 2014; Evrony *et al.*, 2017), suggesting that other genes identified by us could be validated in future studies. Ultimately, these models offer a tangible chance of improving disease gene discovery rates, addressing the problem of missing pathogenicity that most large-scale sequencing studies encounter and helping with disease stratification and treatment.

Our study indicates that genetic mutations causing neuroanatomical defects might distort known synaptic connections. Indeed, functional examination of human orthologues of genes inducing mouse neuroanatomical defects point to brain genes with synaptic functions. Specifically, our findings implicate genes with roles in cytoskeleton remodelling, crucial for processes of axonal growth and morphogenesis, synaptogenesis and synaptic plasticity (**Figs 6.4 and 6.5**). In fact, these genes reveal two primary modules, consisting of (1) neurodevelopmental genes acting in the prenatal brain, which regulate cell proliferation, via mitotic control. Conversely, (2) the second group comprises mature brain genes, expressed during post-natal life and involved in synaptic plasticity via regulation of cytoskeleton organisation at postsynaptic densities

(**Figs 6.10 and 6.11; Table S6.3**). These modules reiterate previous findings from ASD (Vorstman *et al.*, 2017; **chapter 5**), Gilles de la Tourette syndrome (**chapter 3**) and developmental delay (**chapter 4**), supporting the notion of aetiological overlap between NDDs. Furthermore, they demonstrate the efficacy of our approach to recapitulate major themes of NDDs, further recommending the use of mouse models to study the underlying pathogenic mechanisms of human disease.

Central to the field of disease genomics is the systematic examination of genotype and phenotype relationships. While the same biological pathways have consistently been globally linked to NDDs, the aetiopathology behind the onset of variably expressive forms of disease remains elusive. Here, we addressed this question by evaluating the similarity of neuroanatomical defects caused by functionally related genes in a controlled genetic background. We found an overall agreement between genotype and phenotype, such that regional brain expression patterns were indicative of the abnormalities observed. Indeed, genes strongly expressed across tissues mapping to Bregma +0.98mm and Bregma -1.34mm cause similar morphological abnormalities in the respective regions (**Fig 6.12**).

This study presents several limitations including (1) the combined analysis of homozygous and heterozygous animals, as the loss of one or two copies of dosage-sensitive genes can lead to different phenotypic outcomes (de Angelis *et al.*, 2015). We, however, reasoned zygosity should not impact our findings as this study was not intended for an in-depth phenotypic inspection of mouse

models and their systematic comparison, but, rather, to uncover genes whose disruption is more likely to yield morphological brain anomalies and NDDs in humans. Secondly, (2) the phenotyping pipeline focuses on a subset of morphological features, regarding only two brain regions, and neglects behavioural and qualitative phenotypes, suggesting more comprehensive assessments could reveal additional disease-relevant abnormalities. Thirdly, (3) given our interest in brain morphogenesis and NDDs, the use of adult, eight-week-old, mice, equivalent to 20-year-old humans, presents limitations and calls for validation of our findings in embryonic or infant animals. More generally, (4) the study of KO models is uniquely informative for loss of function-associated phenotypes, but may not inform on gain of function mutations, such as those observed in rare microduplication syndromes. Lastly, (5) while the translation of biological findings between human and mouse is strongly supported, comparative genomic and genetic research has identified both similarities and divergences at the sequence (Chinwalla *et al.*, 2002), transcriptomic (Miller *et al.*, 2010b; Zheng-Bradley *et al.*, 2010; Brawand *et al.*, 2011; Prasad *et al.*, 2013; Yue *et al.*, 2014; Lin *et al.*, 2014; Monaco *et al.*, 2015; Pervouchine *et al.*, 2015) and functional (Nehrt *et al.*, 2011; Chen and Zhang, 2012) levels, highlighting the need for further validation using human *in vivo* and *in vitro* models.

Despite these limitations, the data presented here represents a rich resource to investigate gene function and demonstrates the relevance of mouse models in the study of human disease, leading to the discovery of potential candidates for ID and microcephaly. Indeed, our resource opens up the

possibility of exploiting neuroanatomy as a common denominator of cognition, offering endophenotypes for neurological or psychiatric disorders. With the advent of high throughput technologies, such as CRISPR/Cas-9, and systematic phenotyping, the Sanger MGP will inevitably enhance the functional characterisation of the mouse genome and the investigation of pathological mechanisms of disease.

6.6 Conclusion

In this study, we analysed 1,057 mouse KOs, assessed for a large number of neuroanatomical features, such that we identified 158 NAP genes, whose disruption yields strong brain defects. The 1:1 human orthologues of these NAP genes encompass the two major NDD themes, namely (1) prenatal brain genes with roles in neurodevelopment and regulation of cell proliferation and (2) mature brain genes specifically associated with synaptic transmission and plasticity. Phenotypic examination of mouse models reveals genotype to phenotype convergence, such that spatial gene expression specificity is indicative of the strength of brain regional anomalies. In the context of human disease, NAP genes gather an appreciable number of microcephaly and ID genes. Furthermore, by taking advantage of systematically phenotyped human disease cohorts, we propose several novel potential ID candidates. In particular, the *ARVCF* gene, encoding a protein potentially mediating adhesion of pre- and postsynaptic cells, is recurrently observed in patients with ID, ASD and microcephaly.

CHAPTER 7

From genomics to the clinic:

conclusions and perspectives

Each one of the estimated one hundred trillion cells forming the human body contains three billion maternally and three billion paternally inherited DNA bases, spanning a total of 24 billion possible nucleotides. Considering (1) the immense number of possible inherited base combinations, (2) the large number of DNA alterations caused by germline, somatic and epigenetic changes (e.g. 44 to 82 *de novo* point mutations arise in the genome of an average individual; Acuna-Hidalgo *et al.*, 2016), as well as (3) environmental influences, two individuals are unlikely to ever be identical. Inevitably, such variation can manifest itself with strong deviations from *normality*, resulting in anomalies of human functions or structures and producing abnormal phenotypes that define a disorder. As a result, the understanding of human disease and the development of effective medical treatments rely on the rigorous genetic and non-genetic characterisation of individuals presenting with the disorders of interest.

In the context of NeuroDevelopmental Disorders (NDDs), the genetic basis of several conditions is relatively well understood. Examples include Williams syndrome, caused by the hemizygous deletion of 27 genes from the q11.23 region of chromosome 7. Conversely, the underlying causes of more common yet broader disorders, including Developmental Delay (DD), Intellectual Disability (ID), Autism Spectrum Disorder (ASD) and Gilles de la Tourette Syndrome (GTS), remain incomplete. As a result, genetic screening of patients with these conditions is necessary to shape and advance our understanding of their underlying architecture. In fact, various forms of genetic variation should be examined, including coding and non-coding point

mutations, small insertions and deletions, as well as structural variants, under the form of Copy Number Variants (CNVs; deletions and duplication of DNA segments > 1 kb). These types of variants can occur at both high and low population frequencies (i.e. common and rare variants present in more or less than 1% of the population, respectively) and can be inherited or arise *de novo* (*dn*).

GTS is an NDD affecting 1% of the population and is characterised by the presence of motor and vocal tics. Beyond the discovery of neurodevelopmental CNVs (i.e. conferring risk to several NDDs) in GTS patients, CNV studies have mostly generated limited insights into the aetiological mechanisms of disease (Sundaram *et al.*, 2010; Fernandez *et al.*, 2012; Nag *et al.*, 2013; McGrath *et al.*, 2014; Bertelsen *et al.*, 2016; Huang *et al.*, 2016). In this respect, we have undertaken a genome-wide CNV scan of Danish GTS cases and healthy controls (**chapter 3**). Under the assumption that causal genetic lesions target a common set of disease-associated functions, we identified a subset of CNV-affected genes more likely to contribute to GTS pathogenesis. Specifically, we implicated genes encoding transcriptional regulators of early neurodevelopmental processes and post-natal brain genes associated with synaptic functions. As these themes have consistently been associated with many conditions, including ID and ASD (Vortsman *et al.*, 2017), we provide corroborative evidence for the notion of aetiological overlap across disorders. Accordingly, our findings support the view that NDDs manifest as a phenotypic spectrum rather than a collection of phenotypically and genetically distinct conditions. Consequently, the

examination of individuals presenting with a wider range of neuropsychiatric abnormalities should gain considerable power for the detection of NDD-associated mechanisms, which can further be used to infer causal aetiopathologies in patients.

The discovery of disease-causing variants relies on the identification of those recurrently arising in patients compared to healthy individuals. While our GTS study did not seemingly uncover disease-associated genetic lesions, CNVs affecting the *DLG2* and *MYT1L* genes were more frequently observed in disease subjects (**chapter 3**). While these signals remain to be replicated in independent cohorts, the discovery of *DLG2* variants in ASD and schizophrenia and the implication of *MYT1L* in the aetiology of ID (MIM:616621) support roles for these genes in GTS and across NDDs. Furthermore, based on the overlap with variants from ASD and ID patients, several additional genes, from the subset we identified as more likely to relate to disease, warrant attention. Indeed, the theory of aetiological overlap across NDDs coupled to the *guilt by association* principle suggest that genes previously observed in individuals presenting with NDDs and/or linked to NDD themes are more likely to be causal, hence facilitating the prioritisation of novel candidates.

Stepping away from the evaluation of patient cohorts, mouse models can also be used to predict novel disease susceptibility genes, based on their face validity and construct validity with human disease characteristics and causes (Buxbaum *et al.*, 2012; Gai *et al.*, 2012; White *et al.*, 2013; Honti *et al.*, 2014).

Accordingly, genes whose disruption induces NDD-associated physical, cognitive or behavioural abnormalities are prioritised for further assessments. Aiming to expand our understanding of gene function, large-scale mouse mutagenesis projects have been undertaken to systematically phenotype knockout models of virtually all genes from the mouse genome. Here, in collaboration with the Wellcome Trust Sanger Institute Mouse Genomics Project (Sanger MGP) initiative, we analysed over 1,000 mouse neuroanatomical knockouts, demonstrating the relevance of these models in the study of human disease (**chapter 6**). Specifically, as human orthologues of genes causing stronger mouse neuroanatomical defects reveal higher likelihoods of essentiality, we examined their contribution to disease. Orthologues of genes inducing significant morphological mouse brain abnormalities involve both early neurodevelopmental processes (i.e. cell proliferation and synaptogenesis) and mature neuronal functions (i.e. synaptic transmission and plasticity). Given the substantial overlap between these orthologues and ID susceptibility genes, our results (1) support the implication of these two major NDD themes in the pathogenesis of ID and (2) inversely establish the efficacy of mouse neuroanatomical models to study ID. As a result, orthologues of genes associated with mouse neuroanatomical defects represent a rich resource to uncover novel potential ID genes. In particular, by taking advantage of patient NDD cohorts, we support a role for the *ARVCF* gene in the aetiology of NDDs with accompanying cranial defects.

Beyond the identification of novel candidate disease genes, comprehensive mouse phenotyping initiatives, such as the Sanger MGP, and the integration

of the associated findings in the Mouse Genome Informatics data resource (http://www.informatics.jax.org; Blake *et al.*, 2017) facilitate the systematic interrogation of gene function. Moreover, the relevance of mouse models extends into the assessment of causality of genetic variants from patients, through the evaluation of mouse phenotypic outcomes. Finally, neuro-biological and -pathological studies of developmental mouse brains enable the study and elucidation of pathogenic disease mechanisms. These avenues have, however, not been exploited in this thesis.

While the genetic architecture of NDDs, such as GTS or language impairments, remains relatively obscure, our knowledge of other disorders, including ASD and ID, has been greatly enriched over the past decade. Indeed, the development of data-sharing genomics communities facilitated the recruitment of large numbers of patients under initiatives such as the DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources (DECIPHER) (Firth *et al.*, 2009), the European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations (ECARUCA) (Vulto-van Silfhout *et al.*, 2013b), the Simons Simplex Collection (SSC) (Fischbach and Lord, 2010), the Autism Genetic Resource Exchange (AGRE) (Lajonchere and the AGRE Consortium, 2010), the Autism Genome Project (AGP) (Hu-Lince *et al.*, 2005) and the Autism Simplex Collection (ASC) (Buxbaum *et al.*, 2014). Enabled by the development of high throughput molecular technologies and analytical tools, genotyping and sequencing of these cohorts have (1) shaped the mutational landscape of NDDs, (2) led to the discovery of susceptibility genes and (3) shed light onto the various

dysregulated biological pathways. Nevertheless, our mechanistic understanding of disease remains elusive, hindering the development or effective and safe therapeutic strategies.

Disentangling the genetic heterogeneity of NDDs is complicated by the equally striking phenotypic heterogeneity. The qualitative impairments of the diagnostic features of ID (characterised by limitations in cognitive functioning and adaptive function) and ASD (characterised by deficits in social interaction and communication and restricted and repetitive behaviours) are highly variable. Moreover, these conditions are often accompanied by additional comorbid disorders, malformations and medical problems. Indeed, while classified under the same categories from the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) (American Psychiatric Association, 2013), ID and ASD patients, respectively, uncover a tremendous clinical variability. Under the hypothesis that the disruption of common functional processes leads to the development of analogous phenotypes, this heterogeneity advises the examination of tens of thousands of individuals is necessary to elucidate the onset of variably expressive forms of disease.

Based on the view that DSM-5 categories provide artificial classifications of diseases (Van Heugten-van der Kloet and van Heugten, 2015), which can manifest themselves with similar cognitive, behavioural, physical and medical abnormalities, we considered a cohort of individuals with DD, ID, ASD and congenital abnormalities (**chapter 4**). Taking advantage of the systematic phenotyping of these patients using standardised clinical vocabulary, provided

by the Human Phenotype Ontology (HPO) (Robinson *et al.*, 2008; Köhler *et al.*, 2017), we explored the paradigm that analogous abnormalities arise from the disruption of common biological pathways. For this, we formed sub-groups of patients with similar CNV disturbances. While we demonstrated global genotype and phenotype convergence, no clinical archetype could be associated with the gene clusters identified. Instead, common functional disruptions appeared to predispose towards a phenotypic continuum, characterised by the presence and absence of more homogeneous sets of abnormalities. These results highlight the need to interrogate large numbers of patients to reconcile the plethora of variants expected to underlie the onset of variably expressive forms of disease with their associated interactions, incomplete penetrance and pleiotropy.

By examining the clinical manifestations of functionally converging rare *de novo* gene variants in ASD patients from the SSC, we extended our findings regarding the coherence between genotype and phenotype across different types of genetic variation (**chapter 5**). Taken together, our findings indicate structural and point variants affecting genes, which participate in shared molecular processes, underlie the development of globally more similar clinical abnormalities, whose combinations and associated expressivity define a phenotypic spectrum. As a result, this work advises the integrated study of patients diagnosed with DSM-5-derived NDD categories across the ensemble of rare genetic variants will enhance our understanding of the genetic and phenotypic landscapes of disease.

The phenotypic variability of individuals displaying analogous genetic lesions suggests other factors, such as genetic background, modulate the clinical manifestations of disease. Taking advantage of the phenotypically coherent groups of SSC patients we identified, we investigated the influence of common genetic variation over the associated phenotypic spectra (**chapter 5**). By evaluating the relationship between phenotypic resemblance and genetic distance, we showed common variants in proximity of genes displaying NDD-associated functions (e.g. regulation of chromatin state and synaptic genes) are likely to influence disease outcomes. In particular, they appear to specifically modulate the effects of rare penetrant mutations affecting genes from the respective classes. As our study represents the first systematic investigation of the ability of common variants to determine the expression of various ASD traits, the significance of our findings remains to be validated in independent cohorts. Notwithstanding, this work advises that in line with polygenic models of disease, the additive, combinatorial and modifying effects of large numbers of genetic variants, encompassing various types, frequencies and modes of inheritance are likely to underlie the development of NDDs.

As discussed here, the translation of genetic data into the mechanistic understanding of NDD aetiopathologies involves (1) the description of mutational disease landscapes, (2) the discovery and functional validation of novel candidate genes as well as (3) the integrated evaluation of genotypes and phenotypes to elucidate how they influence each other (**Fig 7.1**). In turn, such research relies on the continuous efforts of genomics communities to

recruit patient cohorts, generate high-quality sequencing data, provide systematic and rigorous patient phenotypic measures and develop safe data-sharing platforms for data delivery.

In the context of disease prevention, NDDs call for early therapeutic interventions, given the associated early neurodevelopmental abnormalities. In fact, it is unclear whether behavioural and pharmacological therapies can restore function following the onset of symptoms, or whether the system has already degenerated to a point of no or only partial return. Over the past decades, however, treatments have mostly been administered at the end of infancy and throughout childhood, following firm NDD diagnoses. Nevertheless, the observation that foetal DNA can be isolated from maternal blood makes non-invasive sequencing of foetal genomes a technical reality (Fan *et al.*, 2012; Kitzman *et al.*, 2012). With respect to the use foetal screening to predict illnesses, the field of disease genomics promises the discovery of risk variants and the elucidation of aetiological pathways underlying NDDs. In turn, such advancements pave the road towards the development of personalised medicine, wherein pharmacologic therapies are specifically administered to address particular lesions. For example, individuals carrying high-risk *SCN2A* mutations could potentially be treated using allosteric modulators of GABA receptors (Yuen *et al.*, 2017). Nevertheless, individualised treatments targeted to each genetic cause of disease are not an economically and medically viable option. An alternate, hopeful, therapeutic strategy involves the development of a finite number of interventions to address a predefined set of biological pathways, predisposing

towards relatively well-characterised phenotypic spectra. The prospect of this scenario is supported by the work performed in this thesis.

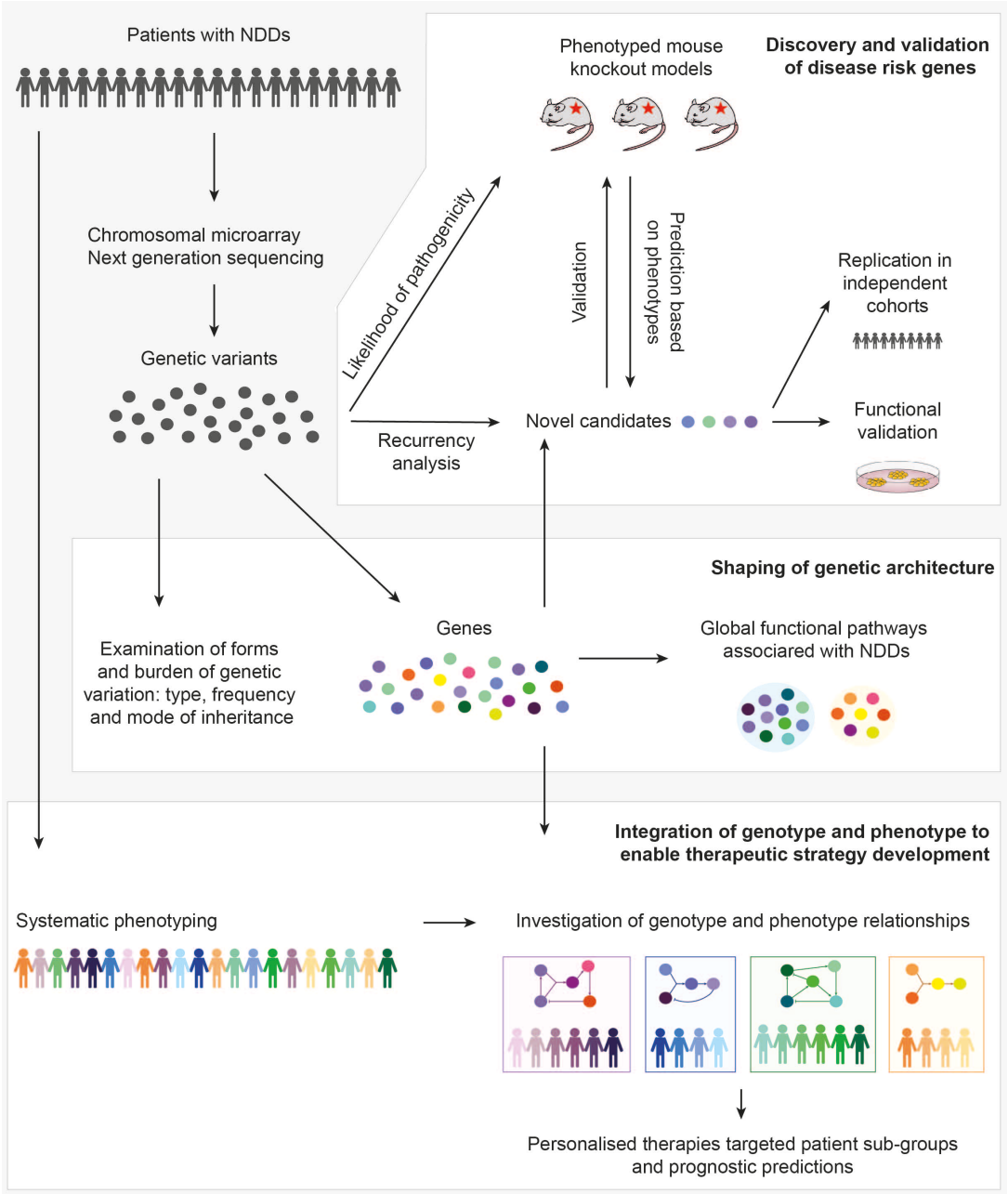


Fig 7.1 | From genomics to the clinic. In this thesis, we tackled three major aspects of translational genomic research in NDDs, namely the genetic architecture of GTS and ASD (**chapters 3 and 5**), the discovery of novel GTS and ID candidate genes (**chapters 3 and 6**) and the investigation of genotype and phenotype relationships in DD, ID and ASD (**chapters 4 and 5**).

BIBLIOGRAPHY

Abelson, J.F., Kwan, K.Y., O'Roak, B.J., Baek, D.Y., Stillman, A.A., Morgan, T.M., Mathews, C.A., Pauls, D.L., Rasin, M.R., Gunel, M. *et al.* (2005). Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science*, 310(5746): 317-320. doi:10.1126/science.1116502.

Abraham, W.C. (2008). Metaplasticity: tuning synapses and networks for plasticity. *Nature Reviews Neuroscience*, 9(5): 387. doi:10.1038/nrn2356.

Acuna-Hidalgo, R., Veltman, J.A. and Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. *Genome Biology*, 17(1): 241. doi:10.1186/s13059-016-1110-1.

Adzhubei, I., Jordan, D.M. and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, Chapter 7: Unit7.20. doi:10.1002/0471142905.hg0720s76.

Alarcon, M., Abrahams, B.S., Stone, J.L., Duvall, J.A., Perederiy, J.V., Bomar, J.M., Sebat, J., Wigler, M., Martin, C.L., Ledbetter, D.H. *et al.* (2008). Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. *American Journal of Human Genetics*, 82(1): 160-164. doi:10.1016/j.ajhg.2007.09.005.

Alazami, A.M., Patel, N., Shamseldin, H.E., Anazi, S., Al-Dosari, M.S., Alzahrani, F., Hijazi, H., Alshammari, M., Aldahmesh, M.A., Salih, M.A. *et al*. (2015). Accelerating novel candidate gene discovery in neurogenetic disorders via whole-exome sequencing of prescreened multiplex consanguineous families. *Cell Reports*, 10(2): 148-161. doi:10.1016/j.celrep.2014.12.015.

Albin, R.L., and Mink, J.W. (2006). Recent advances in Tourette syndrome research. *Trends in Neuroscience*, 29(3): 175-182. doi:10.1016/j.tins.2006.01.001.

Al-Shahrour, F., Minguez, P., Marqués-Bonet, T., Gazave, E., Navarro, A. and Dopazo, J. (2010). Selection upon genome architecture: conservation of functional neighborhoods with changing genes. *PLoS Computational Biology*, 6(10): e1000953. doi:10.1371/journal.pcbi.1000953.

American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders (4th ed. Text Revision). Washington, DC: American Psychiatric Publishing.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, VA: American Psychiatric Publishing.

Amir, R.E., Van den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U. and Zoghbi, H.Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature Genetics*, 23(2): 185-188.

Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P. and Zondervan, K.T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols*, 5(9): 1564-1573. doi:10.1038/nprot.2010.116.

Anderson, C.A., Soranzo, N., Zeggini, E. and Barrett, J.C. (2011). Synthetic Associations Are Unlikely to Account for Many Common Disease Genome-Wide Association Signals. *PLoS Biology*, 9(1): e1000580. doi:10.1371/journal.pbio.1000580.

Andrews, T., Honti, F., Pfundt, R., de Leeuw, N., Hehir-Kwa, J., Vulto-van Silfhout, A., de Vries, B. and Webber, C. (2015a). The clustering of functionally related genes contributes to CNV-mediated disease. *Genome Research*, 25(6): 802-813. doi:10.1101/gr.184325.114.

Andrews, T., Meader, S., Vulto-van Silfhout, A., Taylor, A., Steinberg, J., Hehir-Kwa, J., Pfundt, R., de Leeuw, N., de Vries, B.B. and Webber, C. (2015b). Gene networks underlying convergent and pleiotropic phenotypes in a large and systematically-phenotyped cohort with heterogeneous developmental disorders. *PLoS Genetics*, 11(3): e1005012. doi: 10.1371/journal.pgen.1005012.

Anney, R., Klei, L., Pinto, D., Almeida, J., Bacchelli, E., Baird, G., Bolshakova, N., Bölte, S., Bolton, P.F., Bourgeron, T. *et al.* (2012). Individual common variants exert weak effects on the risk for autism spectrum disorders. *Human Molecular Genetics*, 21(21): 4781-4792. doi:10.1093/hmg/dds301.

Anney, R., Klei, L., Pinto, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., Sykes, N., Pagnamenta, A.T. *et al.* (2010). A genome-wide scan for common alleles affecting risk for autism. *Human Molecular Genetics*, 19(20): 4072-4082. doi:10.1093/hmg/ddq307.

Anttila, V., Winsvold, B.S., Gormley, P., Kurth, T., Bettella, F., McMahon, G., Kallela, M., Malik, R., de Vries, B., Terwindt, G. *et al.* (2013). Genome-wide meta-analysis identifies new susceptibility loci for migraine. *Nature Genetics*, 45(8): 912-917. doi:10.1038/ng.2676.

Ardhanareeswaran, K., Mariani, J., Coppola, G., Abyzov, A. and Vaccarino, F.M. (2017). Human induced pluripotent stem cells for modelling neurodevelopmental disorders. *Nature Reviews Neurology*, 13(5): 265-278. doi:10.1038/nrneurol.2017.45.

Arking, D.E., Cutler, D.J., Brune, C.W., Teslovich, T.M., West, K., Ikeda, M., Rea, A., Guy, M., Lin, S., Cook, E.H. and Chakravarti, A. (2008). A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. *American Journal of Human Genetics*, 82(1): 160-164. doi:10.1016/j.ajhg.2007.09.015.

Ashwood, P. and Van de Water, J. (2004). Is autism an autoimmune disease? *Autoimmunity Reviews*, 3(7-8): 557-562.

Autism Genome Project Consortium, Szatmari, P., Paterson, A.D., Zwaigenbaum, L., Roberts, W., Brian, J., Liu, X.Q., Vincent, J.B., Skaug, J.L., Thompson, A.P. *et al.* (2007). Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nature Genetics*, 39(3): 319-28. doi:10.1038/ng1985.

Ayadi, A., Birling, M.C., Bottomley, J., Bussell, J., Fuchs, H., Fray, M., Gailus-Durner, V., Greenaway, S., Houghton, R., Karp, N. *et al.* (2012). Mouse large-scale phenotyping initiatives: overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. *Mammalian Genome*, 23(9-10): 600-610. doi:10.1007/s00335-012-9418-y.

Badano, J.L., Leitch, C.C., Ansley, S.J., May-Simera, H., Lawson, S., Lewis, R.A., Beales, P.L., Dietz, H.C., Fisher, S. and Katsanis, N. (2006). Dissection of epistasis in oligogenic Bardet-Biedl syndrome. *Nature,* 439(7074): 326-30. doi:10.1038/nature04370.

Bailey, A., Le Couteur, A., Gottesman, I., Bolton, P., Simonoff, E., Yuzda, E. and Rutter, M. (1995). Autism as a strongly genetic disorder: evidence from a British twin study. *Psychological Medicine*, 25(1):63-77.

Bakken, T.E., Bloss, C.S., Roddey, J.C., Joyner, A.H., Rimol, L.M., Djurovic, S., Melle, I., Sundet, K., Agartz, I., Andreassen, O.A. *et al.* (2011). Association of genetic variants on 15q12 with cortical thickness and cognition in schizophrenia. *JAMA Psychiatry*, 68(8): 781-790. doi:10.1001/archgenpsychiatry.2011.81.

Baple, E.L., Maroofian, R., Chioza, B.A., Izadi, M., Cross, H.E., Al-Turki, S., Barwick, K., Skrzypiec, A., Pawlak, R., Wagner, K. *et al.* (2014). Mutations in KPTN Cause Macrocephaly, Neurodevelopmental Delay, and Seizures. *American Journal of Human Genetics*, 94(1): 87-94. doi:10.1016/j.ajhg.2013.10.001.

Barak, T., Kwan, K.Y., Louvi, A., DemirBilek, V., Saygi, S., Tusysuz, B., Choi, M., Boyaci, H., Doerschner, K., Zhu, Y. *et al.* (2011). Recessive LAMC3 mutations cause malformations of occipital cortical development. *Nature Genetics*, 43(6): 590-594. doi:10.1030/ng.836.

Barbaresi, W.J., Colligan, R.C., Weaver, A.L., Voigt, R.G., Killian, J.M. and Katusic, S.K. (2013). Mortality, ADHD, and psychosocial adversity in adults with childhood ADHD: a prospective study. *Pediatrics*, 131(4): 637-644. doi:10.1542/peds.2012-2354.

Barkovich, A.J., Guerrini, R., Kuzniecky, R.I., Jackson, G.D. and Dobyns, W.B. (2012). A developmental and genetic classification for malformations of cortical development: update 2012. *Brain,* 135(Pt 5): 1348-1369. doi:10.1093/brain/aws019.

Barnard, R.A., Pomaville, M.B. and O'Roak, B.J. (2015). Mutations and Modeling of the Chromatin Remodeler CHD8 Define an Emerging Autism Etiology. *Frontiers in Neuroscience*, 9: 477. doi:10.3389/fnins.2015.00477.

Barros, C.S., Franco, S.J. and Müller, U. (2011). Extracellular matrix: functions in the nervous system. *Cold Spring Harbour Perspectives in Biology*, 3(1): a005108. doi:10.1101/cshperspect.a005108.

Bartlett, C.W., Flax, J.F., Logue, M.W., Smith, B.J., Vieland, V.J., Tallal, P. and Brzustowicz, L.M. (2004). Examination of potential overlap in autism and language loci on chromosomes 2, 7, and 13 in two independent samples ascertained for specific language impairment. *Human Heredity*, 57(1):10-20. doi:10.1159/000077385

Baspinar, A., Cukuroglu, E., Nussinov, R., Keskin, O. and Gursoy, A. (2014). PRISM: a web server and repository for prediction of protein-protein interactions and modeling their 3D complexes. *Nucleic Acids Research*, 42(Web Server issue): W285-W289. doi:10.1093/nar/gku397.

Basu, S.N., Kollu, R. and Banerjee-Basu, S. (2009). AutDB: a gene reference resource for autism research. *Nucleic Acids Research*, 37(Database issue): D832-D836. doi:10.1093/nar/gkn835.

Baum, A.E., Akula, N., Cabanero, M., Cardona, I., Corona, W., Klemens, B., Schulze, T.G., Cichon, S., Rietschel, M., Nöthen, M.M. *et al.* (2008). A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Molecular Psychiatry*, 13(2): 197-207.

Baym, C.L., Corbett, B.A.m Wright S.B, and Bunge, S.A. (2008). Neural correlates of tic severity and cognitive control in children with Tourette syndrome. *Brain,* 131(Pt 1):165-79. doi:http://dx.doi.org/10.1093/brain/awm278

Beaudoin, M., Goyette, P., Boucher, G., Lo, K.S., Rivas, M.A., Stevens, C., Alikashani, A., Ladouceur, M., Ellinghaus, D., Törkvist, L. *et al.* (2013). Deep resequencing of GWAS loci identifies rare variants in CARD9, IL23R and RNF186 that are associated with ulcerative colitis. *PLoS Genetics*, 9(9): e1003723. doi:10.1371/journal.pgen.1003723.

Ben-David, E. and Shifman, S. (2013). Combined analysis of exome sequencing points toward a major role for transcription regulation during brain development in autism. *Molecular Psychiatry*, 18(10): 1054-1056. doi:10.1038/mp.2012.148.

Berg, J.M. and Geschwind, D.H. (2012). Autism genetics: searching for specificity and convergence. *Genome Biology*, 13(7): 247. doi:10.1186/gb4034.

Bernier, R., Golzio, C., Xiong, B., Stessman, H.A., Coe, B.P., Penn, O., Witherspoon, K., Gerdts, J., Baker, C., Vulto-van Silfhout, A.T. *et al.* (2014). Disruptive CHD8 mutations define a subtype of autism early in development. *Cell,* 158(2): 263-276. doi:10.1016/j.cell.2014.06.017.

Bertelsen, B., Stefánsson, H., Riff Jensen, L., Melchior, L., Mol Debes, N., Groth, C., Skov, L., Werge, T., Karagiannidis, I., Tarnok, Z. *et al.* (2016). Association of AADAC Deletion and Gilles de la Tourette Syndrome in a Large European Cohort. *Biological Psychiatry*, 79(5): 383-391. doi:10.1016/j.biopsych.2015.08.027.

Betancur, C. and Mitchell, K.J. (2015). Synaptic disorders. In: The Genetics of neurodevelopmental disorders (1st ed). Hoboken, NJ: John Wiley & Sons, Inc, pp. 195-238.

Betancur, C., Sakurai, T. and Buxbaum, J.D. (2009). The emerging role of synaptic cell-adhesion pathways in the pathogenesis of autism spectrum disorders. *Trends in Neuroscience*, 32(7): 402-412. doi:10.1016/j.tins.2009.04.003.

Bicknell, L.S., Bongers, E.M., Leitch, A., Brown, S., Schoots, J., Harley, M.E., Aftimos, S., Al-Aama, J.Y., Bober, M., Brown, P.A. *et al.* (2011). Mutations in the pre-replication complex cause Meier-Gorlin syndrome. *Nature Genetics*, 43(4): 356-359. doi:10.1038/ng.775.

Biederman, J., Faraone, S.V., Keenan, K., Knee, D. and Tsuang, M.T. (1990). Family-genetic and psychosocial risk factors in DSM-III attention deficit disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, 29(4): 526-533. doi:10.1097/00004583-199007000-00004.

Bilder, D., Botts, E.L., Smith, K.R., Pimentel, R., Farley, M., Viskochil, J., McMahon, W.M., Block, H., Ritvo, E., Ritvo, R.A. and Coon, H. (2013). Excess mortality and causes of death in autism spectrum disorders: a follow up of the 1980s Utah/UCLA autism epidemiologic study. *Journal of Autism and Developmental Disorders*, 43(5): 1196-1204. doi:10.1007/s10803-012-1664-z.

Bilgüvar, K., Oztürk, A.K., Louvi, A., Kwan, K.Y., Choi, M., Tatli, B., Yalnizoğlu, D., Tüysüz, B., Cağlayan, A.O., Gökben, S. *et al.* (2010). Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature,* 467(7312): 207-210. doi:10.1038/nature09327.

Bishop, D.V., North, T. and Donlan, C. (1995). Genetic basis of specific language impairment: evidence from a twin study. *Developmental Medicine and Child Neurology,* 37(1): 56-71.

Blake, J.A., Eppig, J.T., Kadin, J.A., Richardson, J.E., Smith, C.L., Bult, C.J. and the Mouse Genome Database Group. (2017). Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Research*, 45(D1): D723-D729.

Blondel, V.D., Guillaune, J.-L., Lambiotte, R. Lefebvre (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, P10008. doi:10.1088/1742-5468/2008/10/P10008.

Boriah, S, Varun, C. and Kuman, V. (2008). Similarity measures for categorical data: A commparative evaluation. *Proceedings of the eighth SIAM International Conference on Data Mining*, 30: 234–254.

Bourgeron, T. (2015). From the genetic architecture to synaptic plasticity in autism spectrum disorder. *Nature Reviews Neuroscience*, 16(9): 551-563. doi:10.1038/nrn3992.

Boyle, C.A., Boulet, S., Schieve, L.A., Cohen, R.A., Blumberg, S.J., Yeargin-Allsopp, M., Visser, S. and Kogan, M.D. (2011). Trends in the prevalence of developmental disabilities in US children, 1997-2008. *Pediatrics*, 127(6): 1034-1042. doi:10.1542/peds.2010-2989.

Bragin, E., Chatzimichali, E.A., Wright, C.F., Hurles, M.E., Firth, H.V., Bevan, A.P. and Swaminathan, G.J. (2014). DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Research*, 42(Database issue): D993-D1000. doi:10.1093/nar/gkt937.

Bramon, E., Pirinen, M., Strange, A., Lin, K., Freeman, C., Bellenguez, C., Su, Z., Band, G., Pearson, R., Vukcevic, D. *et al.* (2014). A Genome-wide Association Analysis of a Broad Psychosis Phenotype Identifies Three Loci for Further Investigation. *Biological Psychiatry*, 75(5): 386–397. doi:10.1016/j.biopsych.2013.03.033.

Brancati, F., Dallapiccola, B. and Valente, E.M. (2010). Joubert Syndrome and related disorders. *Orphanet Journal of Rare Diseases*, 5: 20. doi:10.1186/1750-1172-5-20.

Braunschweig, D., Golub, M.S., Koenig, C.M., Qi, L., Pessah, I.N., Van de Water, J. and Berman, R.F. (2012). Maternal autism-associated IgG antibodies delay development and produce anxiety in a mouse gestational transfer model. *Journal of Neuroimmunology*, 252: 56–65.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M. *et al.* (2011). The evolution of gene expression levels in mammalian organs. *Nature,* 278: 343-348. doi:10.1038/nature10532.

Breschi, A., Gingeras, T.R. and Guigo, R. (2017). Comparative transcriptomics in human and mouse. *Nature Reviews Genetics,* 18(7): 425-440. doi:10.1038/ntg.2017.19.

Brimberg, L., Sadiq, A., Gregersen, P.K. and Diamond, B. (2013) Brain-reactive IgG correlates with autoimmunity in mothers of a child with an autism spectrum disorder. *Molecular Psychiatry*, 18: 1171–1177.

Brown, S.D.M. and Moore, M.W. (2012). The International Mouse Phenotypic Consortium: past and future perspectives on mouse phenotyping. *Mammalian Genome*, 23(0): 632-640. doi:10.1007/s00335-012-9427-x.

Bubb, K.L., Bovee, D., Buckley, D., Haugen, E., Kibukawa, M., Paddock, M., Palmieri, A., Subramanian, S., Zhou, Y., Kaul, R. *et al.* (2006). Scan of human genome reveals no new Loci under ancient balancing selection. *Genetics*, 173(4): 2165-2177. doi:10.1534/genetics.106.055715.

Buescher, A.V., Cidav, Z., Knapp, M. and Mandell, D.S. (2014). Costs of autism spectrum disorders in the United Kingdom and the United States. JAMA *Pediatrics*, 168(8): 721-728. doi:10.1001/jamapediatrics.2014.210.

Burgu, B., Aydogdu, O., Gurkan, K., Uslu, R. and Soygur, T. (2011). Lower urinary tract conditions in children with attention deficit hyperactivity disorder: correlation of symptoms based on validated scoring systems. *Journal of Urology*, 185(2):663-8. doi:10.1016/j.juro.2010.09.116.

Burton, C.L. (2016). 16.1 A Hypothesis-Driven Genome-Wide Association Study of an Obsessive-Compulsive Quantitative Trait in a Community-Based Sample of Children and Adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry*, 55(10): S282. doi:10.1016/j.jaac.2016.07.212.

Buxbaum, J.D., Betancur, C., Bozdagi, O., Dorr, N.P., Elder, G.A. and Hof, P.R. (2012). Optimizing the phenotyping of rodent ASD models: enrichment analysis of mouse and human neurobiological phenotypes associated with high-risk autism genes identifies morphological, electrophysiological, neurological, and behavioral features. *Molecular Autism*, 3: 1. doi:10.1186/2040-2392-3-1.

Buxbaum, J.D., Bolshakova, N., Brownfeld, J.M., Anney, R.J., Bender, P., Bernier, R., Cook, E.H., Coon, H., Cuccaro, M., Freitag, C.M. *et al.* (2014). The Autism Simplex Collection: an international, expertly phenotyped autism sample for genetic and phenotypic analyses. *Molecular Autism*, 5: 34. Doi:10.1186/2040-2392-5-34.

Callan M.A. and Zarnescu, D.C. (2011). Heads-up: new roles for the fragile X mental retardation protein in neural stem and progenitor cells. *Genesis*, 49(6): 424-440. doi:10.1002/dvg.20745.

Calzolar, E., Barisic, I., Loane, M., Morris, J., Wellesley, D., Dolk, H., Addor, M.C., Arriola, L., Bianchi, F., Neville, A.J. *et al.* (2014). Epidemiology of multiple congenital anomalies in Europe: a EUROCAT population-based registry study. *Birth Defects Research Part A: Clinical and Molecular Teratology*, 100(4): 270-6. doi:10.1002/bdra.23240.

Cantor, R.M., Navarro, L., Won, H., Walker, R.L., Lowe, J.K. and Geschwind, D. (2017). ASD restricted and repetitive behaviors associated at 17q21.33: genes prioritized by expression in fetal brains. *Molecular Psychiatry*, Advance Online Publication. doi10.1038/mp.2017.114.

Cappi, C., Brentani, H., Lima, L., Sanders, S.J., Zai, G., Diniz, B.J., Reis, V.N., Hounie, A.G., Conceição do Rosário, M., Mariani, D. *et al.* (2016). Whole-exome sequencing in obsessive-compulsive disorder identifies rare mutations in immunological and neurodevelopmental pathways. *Translational Psychiatry*, 6: e764. doi:10.1038/tp.2016.30.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C.R., Lim, E.P., Kalyanaraman, N. *et al.* (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, 22(3): 231-238. doi:10.1038/10290.

Carrié, A., Jun, L., Bienvenu, T., Vinet, M.C., McDonell, N., Couvert, P., Zemni, R., Cardona, A., Van Buggenhout, G., Frints, S. *et al.* (1999). A new member of the IL-1 receptor family highly expressed in hippocampus and involved in X-linked mental retardation. *Nature Genetics*, 23(1):25-31.

Casanova, E.L., Sharp, J.L., Chakraborty, H., Sumi, N.S. and Casanova, M.F. (2016). Genes with high penetrance for syndromic and non-syndromic autism typically function within the nucleus and regulate gene expression. *Molecular Autism*, 7: 18. doi:10.1186/s13229-016-0082-z.

Cassa, C., Weghorn, D., Balick, D.J., Jorfan, D.M., Nusinow, D., Samocha, K.E., O'Donnell-Luria, A., MacArthur, D.G., Daly, M.J *et al.* (2017). Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nature Genetics*, 49(5): 806-810. doi:10.1038/ng3831.

Ceroni, F., Simpson, N.H., Francks, C., Baird, G., Conti-Ramsden, G., Clark, A., Bolton, P.F., Hennessy, E.R., Donnelly, P., Bentley, D.R. *et al.* (2014). Homozygous

microdeletion of exon 5 in ZNF277 in a girl with specific language impairment. *European Journal of Human Genetics*, 22(10): 1165-1171. doi:10.1038/ejhg.2014.4.

Chahrour, M.H., Yu, T.W., Lim, E.T., Ataman, B., Coulter, M.E., Hill, R.S., Stevens, C.R., Schubert, C.R., ARRA Autism Sequencing Collaboration, Greenberg, M.E. *et al.* (2012). Whole-exome sequencing and homozygosity analysis implicate depolarization-regulated neuronal genes in autism. *PLoS Genetics*, 8(4): e1002635. doi:10.1371/journal.pgen.1002635.

Chakraborti, S., Natarajan, K., Curiel, J., Janke, C. and Liu, J. (2016). The emerging role of the tubulin code: From the tubulin molecule to neuronal function and disease. *Cytoskeleton*, 73(10): 521-550. doi:10.1002/cm.21290.

Chang, D. and Keinan, A. (2012). Predicting signatures of "synthetic associations" and "natural associations" from empirical patterns of human genetic variation. *PLoS Computational Biology*, 8(7): e1002600. doi:10.1371/journal.pcbi.1002600.

Chang, J., Gilman, S.R., Chiang, A.H., Sanders, S.J. and Vitkup, D. (2015). Genotype to phenotype relationships in autism spectrum disorders. *Nature Neuroscience*, 18(2): 191-198. doi:10.1038/nn.3907.

Chaste, P., Klei, L., Sanders, S.J., Hus, V., Murtha, M.T., Lowe, J.K., Willsey, A.J., Moreno-De-Luca, D., Yu, T.W., Fombonne, E. *et al.* (2015). A genome-wide association study of autism using the Simons Simplex Collection: Does reducing phenotypic heterogeneity in autism increase genetic homogeneity? *Biological Psychiatry*, 77(9): 775-784. doi:10.1016/j.biopsych.2014.09.017.

Chatr-Aryamontri, A., Breitkreutz, B.J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O'Donnell, L. *et al.* (2013). The BioGRID interaction database: 2013 update. *Nucleic Acids Research*, 41(Database issue): D816-D823. doi:10.1093/nar/gks1158.

Chavali, P.L., Pütz, M. and Gergely, F. (2014). Small organelle, big responsibility: the role of centrosomes in development and disease. *Philosophical Transactions of the Royal Society of London, Series B: Biological Science*s, 369(1650): 20130468. doi:10.1098/rstb.2013.0468.

Chen, D.T., Jiang, X., Akula, N., Shugart, Y.Y., Wendland, J.R., Steele, C.J., Kassem, L., Park, J.H., Chatterjee, N., Jamain, S. *et al.* (2013). Genome-wide association study meta-analysis of European and Asian-ancestry samples identifies three novel loci associated with bipolar disorder. *Molecular Psychiatry*, 18(2): 195-205. doi:10.1038/mp.2011.157.

Chen, J.F., Zhang, Y., Wilde, J., Hansen, K.C., Lai, F. and Niswander, L. (2014). Microcephaly disease gene Wdr62 regulates mitotic progression of embryonic neural stem cells and brain size. *Nature Communications*, 5: 3885. doi:10.1038/ncomms4885.

Chen, S.-J., Chao, Y.-L., Chen, C.-Y., Chang, C.-M., Wu, E.C.-H, Wu, C.-S., Yeh, H.-H., Chen, C.-H. and Tsai, H.-J. (2012). Prevalence of autoimmune disease in in-patients with schizophrenia: nationwide population-based study. *The British Journal of Psychiartry*, 200(5): 374-380. doi:10.1192/bjp.bp.111.092098.

Chen, X. and Zhang, J. (2012). The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. *PLoS Computational Biology*, 8(11): e1002784. doi:10.1371/journal.pcbi.1002784.

Chinwalla, A.T., Cook, L.L., Delehaunty, K.D., Fewell, G.A., Fulton, L.A., Fulton, R.S., Graves, T.A., Hillier, L.W., Mardis, E.R., McPherson, J.D. *et al.* (2002). Initial

sequencing and comparative analysis of the mouse genome. *Nature,* 420: 520-562. doi:10.1038/nature01262.

Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., Smith, J.D., Harrell, T.M., McMillin, M.J., Wiszniewski, W., Gambin, T. *et al.* (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *American Journal of Human Genetics*, 97(2): 199-215. doi:10.1016/j.ajhg.2015.06.009.

Christensen, D.L., Bilder, D.A., Zahorodny, W., Pettygrove, S., Durkin, M.S., Fitzgerald, R.T., Rice, C., Kurzius-Spencer, M., Baio, J. and Yeargin-Allsopp, M. (2016). Prevalence and Characteristics of Autism Spectrum Disorder Among 4-Year-Old Children in the Autism and Developmental Disabilities Monitoring Network. Journal of Developmental and Behavioural *Pediatrics*, 37(1):1-8. doi:10.1097/DBP.0000000000000235.

Citri, A. and Malenka, R.C. (2008). Synaptic plasticity: multiple forms, functions, and mechanisms. *Neuropsychopharmacology,* 33(1): 18-41. doi:10.1038/sj.npp.1301559.

Clayton, D.G. (2009). Sex chromosomes and genetic association studies. *Genome Medicine*, 1(11): 110. doi:10.1186/gm110.

Coe, B.P., Witherspoon, K., Rosenfeld, J.A., van Bon, B.W., Vulto-van Silfhout, A.T., Bosco, P., Friend, K.L., Baker, C., Buono, S., Vissers, L.E. *et al.* (2014). Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nature Genetics*, 46(10): 1063-1071. doi: 10.1038/ng.3092.

Coghlan, S., Horder, J., Inkster, B., Mendez, M.A., Murphy, D.G. and Nutt, D.J. (2012). GABA system dysfunction in autism and related disorders: from synapse to symptoms. *Neuroscience and Biobehavioural Reviews*, 36(9): 2044-2055. doi:10.1016/j.neubiorev.2012.07.005.

Cohen, S., Conduit, R., Lockley, S.W., Rajaratnam, S.M.W. and Cornish, K.M. (2014). The relationship between sleep and behaviour in autism spectrum disorder (ASD): a review. *Journal of Neurodevelopmental Disorders*, 6(1): 44. doi:10.1186/1866-1955-6-44.

Collins, M.O., Husi, H., Yu, L., Brandon, J.M., Anderson, C.N., Blackstock, W.P., Choudhary, J.S. and Grant, S.G. (2006). Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. *Journal of Neurochemistry*, 97(Suppl 1): 16-23. doi:10.1111/j.1471-4159.2005.03507.x.

Connolly, J.J., Glessner, J.T. and Hakonarson, H. (2013). A genome-wide association study of autism incorporating autism diagnostic interview-revised, autism diagnostic observation schedule, and social responsiveness scale. *Child Development,* 84(1): 17-33. doi:10.1111/j.1467-8624.2012.01838.x.

Constantino, J.N., Zhang, Y., Frazier, T., Abbacchi, A.M. and Law, P. (2010). Sibling recurrence and the genetic epidemiology of autism. *American Journal of Psychiatry*, 167(11): 1349-1356. doi:10.1176/appi.ajp.2010.09101470.

Cooper, D.N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. and Kehrer-Sawatzki, H. (2013). Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Human Genetics*, 132(10): 1077-1130. doi:10.1007/s00439-013-1331-2.

Cooper, G.M., Coe, B.P., Girirajan, S., Rosenfeld, J.A., Vu, T.H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V. *et al.* (2011). A copy number variation

morbidity map of developmental delay. *Nature Genetics*, 43(9): 838-846. doi:10.1038/ng.909.

Cooper, G.M., Stone, E.A., Asimenos, G., NISC Comparative Sequencing Program, Green, E.D., Batzoglou, S. and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, 15(7): 901-913. doi:10.1101/gr.3577405.

Cooper, S.-A., McLean, G., Guthrie, B., McConnachie, A., Mercer, S., Sullivan, F. and Morrison, J. (2015). Multiple physical and mental health comorbidity in adults with intellectual disabilities: population-based cross-sectional analysis. *BMC Family Practice*, 16:110. doi:10.1186/s12875-015-0329-3.

Cooper, S.-A., Smiley, E., Morrison, J., Williamson, A. and Allan, L. (2007). Mental ill-health in adults with intellectual disabilities: prevalence and associated factors. *The British Journal of Psychiatry*, 190: 27-35.

Corominas, R., Yang, X., Lin, G.N., Kang, S., Shen, Y., Ghamsari, L., Broly, M., Rodriguez, M., Tam, S., Trigg, S.A. *et al.* (2014). Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. *Nature Communications*, 5: 3650. doi:10.1038/ncomms4650.

Cotney, J., Muhle, R.A., Sanders, S.J., Liu, L., Willsey, A.J., Niu, W., Liu, W., Klei, L., Lei, J., Yin, J. *et al.* (2015). The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. *Nature Communications*, 6: 6404. doi:10.1038/ncomms7404.

Coutinho, A.M., Sousa, I., Martins, M., Correia, C., Morgadinho, T., Bento, C., Marques, C., Ataíde, A., Miguel, T.S., Moore, J.H. *et al.* (2007). Evidence for epistasis between SLC6A4 and ITGB3 in autism aetiology and in the determination of platelet serotonin levels. *Human Genetics*, 121(2):243-56. doi:10.1007/s00439-006-0301-3.

Cristino, A.S., Williams, S.M., Hawi, Z., An, J.Y., Bellgrove, M.A., Schwartz, C.E., Costa Lda, F. and Claudianos, C. (2014). Neurodevelopmental and neuropsychiatric disorders represent an interconnected molecular system. *Molecular Psychiatry*, 19(3): 294-301. doi:10.1038/mp.2013.16.

Crocker-Buque, A., Currie, S.P., Luz, L.L., Grant, S.G., Duffy, K.R., Kind, P.C. and Daw, M.I. (2016). Altered thelamocortical development in the SAP102 knockout model of intellectual disability. *Human Molecular Genetics*, 25(18): 4052-4061. doi:10.1093/hmg/ddw244.

Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R. *et al.* (2014). The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42(Database issue): D472-477. doi:10.1093/nar/gkt1102.

Croning, M.D., Marshall, M.C., McLaren, P., Armstrong, J.D. and Grant, S.G. (2009). G2Cdb: the Genes to Cognition database. *Nucleic Acids Research*, Database issues. doi:10.1093/nar/gkn700.

Crosbie, J. (2016). 16.2 A Genome-Wide Association Study of a Cognitive Endophenotype of Attention-Deficit/Hyperactivity Disorder in a Community-Based Pediatric Sample. *Journal of the American Academy of Child and Adolescent Psychiatry*, 55(10): S282. doi:10.1016/j.jaac.2016.07.213.

Cross-Disorder Group of the Psychiatric Genomics Consortium, Lee, S.H., Ripke, S., Neale, B.M., Faraone, S.V., Purcell, S.M., Perlis, R.H., Mowry, B.J., Thapar, A.,

Goddard, M.E. *et al.* (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*, 45(9): 984-994. doi:10.1038/ng.2711.

Cross-Disorder Group of the Psychiatric Genomics Consortium (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet*, 381(9875): 1371-1379. doi:10.1016/S0140-6736(12)62129-1.

Cukier, H.N., Dueker, N.D., Slifer, S.H., Lee, J.M., Whitehead, P.L., Lalanne, E., Leyva, N., Konidari, I., Gentry, R.C., Hulme, W.F. *et al.* (2014). Exome sequencing of extended families with autism reveals genes shared across neurodevelopmental and neuropsychiatric disorders. *Molecular Autism*, 5(1): 1. doi:10.1186/2040-2392-5-1.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S. *et al.* (2015). Ensembl 2015. *Nucleic Acids Research*, 43(Database Issue): D662-D669. doi:10.1093/nar/gku1010.

Cuthbert, B.N. and Insel, T.R. (2013). Towards the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Medicine*, 11: 126. doi:10.1186/1741-7015-11-126.

Dabell, M.P., Rosenfeld, J.A., Bader, P., Escobar, L.F., El-Khechen, D., Vallee, S.E., Dinulos, M.B., Curry, C., Fisher, J., Tervo, R. *et al.* (2013). Investigation of NRXN1 deletions: clinical and molecular characterization. American *Journal of Medical Genetics* A, 161A(4): 717-731. doi:10.1002/ajmg.a.35780.

Dahl, A., Iotchkova, V., Baud, A., Johansson, A., Gyllensten, U., Soranzo, N., Mott, R., Kranis, A. and Marchini, J. (2016). A multiple-phenotype imputation method for genetic studies. *Nature Genetics*, 48(4): 466-472. doi:10.1038/ng.3513.

Dalsgaard, S., Ostergaard, S.D., Leckman, J.F., Mortensen, P.B. and Pedersen, M.G. (2015). Mortality in children, adolescents, and adults with attention deficit hyperactivity disorder: a nationwide cohort study. *The Lancet*, 385(9983): 2190-2196. doi:10.1016/S0140-6736(14)61684-6.

Dalton,P., Deacon, R., Blamire, A., Pike, M., McKinlay, I., Stein, J., Styles, P. and Vincent, A. (2003). Maternal neuronal antibodies associated with autism and a language disorder. *Annals of Neurology,* 53: 533–537.

Dang, T., Duan, W.Y., Yu, B., Tong, D.L., Cheng, C., Zhang, Y.F., Wu, W., Ye, K., Zhang, W.X. *et al.* (2017). Autism-associated Dyrk1a truncation mutants impair neuronal dendritic and spine growth and interfere with postnatal cortical development. *Molecular Psychiatry*, doi: 10.1038/mp.2016.253.

Darnell, J.C., Van Driesche, S.J., Zhang, C., Hung, K.Y., Mele, A., Fraser, C.E., Stone, E.F., Chen, C., Fak, J.J., Chi, S.W. *et al.* (2011). FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell,* 146: 246-250.

Davidson, K. (2012). Autoimmunity in Psychiatry. *The British Journal of Psychiatry*, 200(5): 353-355. doi:10.1192/bjp.bp.111.104471.

Davies, G., Marioni, R.E., Liewald, D.C., Hill, W.D., Hagenaars, S.P., Harris, S.E., Ritchie, S.J., Luciano, M., Fawns-Ritchie, C., Lyall, D. *et al.* (2016). Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N=112 151). *Molecular Psychiatry*, 21(6): 758-767. doi:10.1038/mp.2016.45.

Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S.E., Liewald, D., Ke, X., Le Hellard, S., Christoforou, A., Luciano, M. *et al.* (2011). Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Molecular Psychiatry*, 16(10): 996-1005. doi:10.1038/mp.2011.85.

Dawson, S., Glasson, E.J., Dixon, G. and Bower, C. (2009). Birth defects in children with autism spectrum disorders: a population-based, nested case-control study. *American Journal of Epidemiology*, 169(11): 1296-1303. doi:10.1093/aje/kwp059.

de Angelis, M.H., Nicholson, G., Selloum, M., White, J.K., Morgan, H., Ramirez-Solis, R., Sorg, T., Wells, S., Fuchs, H., Fray, M. *et al.* (2015). Analysis of mammalian gene function through broad-based phenotypic screens across a consortium of mouse clinics. *Nature Genetics*, 47(9): 969-978. doi:10.1038/ng.3360.

De la Torre-Ubieta, L., Won, H., Stein, J.L. and Geschwind, D.H. (2016). Advancing the understanfing of autism disease mechanisms through genetics. *Nature Medicine*, 22(4): 345-361. doi:10.1038/nm.4071.

Del Hoyo, L., Xicota, L., Langohr, K., Sánchez-Benavides, G., de Sola, S., Cuenca-Royo, A., Rodriguez, J., Rodríguez-Morató, J., Farré, M., Dierssen, M. *et al.* (2016). VNTR-DAT1 and COMTVal158Met Genotypes Modulate Mental Flexibility and Adaptive Behavior Skills in Down Syndrome. *Frontiers in Behavioural Neuroscience*, 10:193. doi:10.3389/fnbeh.2016.00193.

de Ligt, J., Willemsen, M.H., van Bon, B.W., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C. *et al.* (2012). Diagnostic exome sequencing in persons with severe intellectual disability. *New England Journal of Medicine*, 367(20): 1921-1929. doi:10.1056/NEJMoa1206524.

Deng, H., Gao, H. and Jankovic, J. (2012). The genetics of Tourette syndrome. *Nature Reviews Neurology*, 8: 203-2013. doi:10.1038/nmeurol.2012.26.

De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S. *et al.* (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature,* 515(7526): 209-215. doi:10.1038/nature13772.

de Smith, A.J., Tsalenko, A., Sampas, N., Scheffer, A., Yamada, N.A., Tsang, P., Ben-Dor, A., Yakhini, Z., Ellis, R.J., Bruhn, L. *et al.* (2007). Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Human Molecular Genetics*, 16(23): 2783-2794.

Developmental Disabilities Monitoring Network Surveillance Year 2010 Principal Investigatiors (DMNSY) and Centers for Disease Control and Prevention (CDC). (2014). Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2010*. MMWR Surveillance Summary*, 28(63): 1-21.

Dickinson, M.E., Flenniken, A.N., Xiao, J., Teboul, L., Wong, M.D., White, J.K., Meehan, T.F., Weninger, W.J., Westerberg, H., Adissu, H. *et al.* (2016). High-throughput discovery of novel developmental phenotypes. *Nature,* 537(7621): 508-514. doi:10.1038/nature19356.

Dickson, B.J. (2002). Molecular mechanisms of axon guidance. *Science*, 298(5600): 1959-1964. doi:10.1126/science.1072165.

Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. and Goldstein, D.B. (2010). Rare variants create synthetic genome-wide associations. *PLoS Biology*, 8(1): e1000294. doi:10.1371/journal.pbio.1000294.

Dietmann, S., Georgii, E., Antonoy, A., Tsuda, K. and Mewes, H.W. (2009). The DICS repository: module-assisted analysis of disease-related gene lists. *Bioinformatics*, 25(6): 830-831. doi:10.1093/bioinformatics/btp055.

Doelken, S.C., Köhler, S., Mungall, C.J., Gkoutos, G.V., Ruef, B.J., Smith, C., Smedley, D., Bauer, S., Klopocki, E., Schofield, P.N. *et al.* (2013). Phenotypic overlap in the contribution of individual genes to CNV pathogenicity revealed by cross-species computational analysis of single-gene mutations in humans, mice and zebrafish. *Disease Models and Mechanisms*, 6: 358-372. doi:10.1242/dmm.010322.

Dolcetti, A., Silversides, C.K., Marshall, C.R., Lionel, A.C., Stavropoulos, D.J., Scherer, S.W. and Bassett, A.S. (2013). 1q21.1 Microduplication expression in adults. *Genetics in Medicine*, 15: 282–289. doi:10.1038/gim.2012.129.

Dolinski, K. and Botstein, D. (2007). Orthology and functional conservation in eurkaryotes. *Annual Review of Genetics*, 41: 465-507. doi:10.1146/annurev.genet.40.110405.090439.

Dolk, H., Loane, M. and Garne, E. (2010). The prevalence of congenital anomalies in Europe. *Advances in Experimental Medicine and Biology*, 686: 349-64. doi:10.1007/978-90-481-9485-8_20.

Doyle, A., McGarry, M.P., Lee, N.A. and Lee, J.J. (2013). The Construction of Transgenic and Gene Knockout/Knockin Mouse Models of Human Disease. *Transgenic Research*, 21(2): 327–349. doi:10.1007/s11248-011-9537-3.

Dror, A.A. and Avraham, K.B. (2010). Hearing impairment: a panoply of genes and functions. *Neuron,* 68(2): 293-308. doi:10.1016/j.neuron.2010.10.011.

Edmiston, E., Ashwood, P. and Van der Water, J. (2017). Autoimmunity, Autoantibodies, and Autism Spectrum Disorder. *Biological Psychiatry*, 81(5): 383-390. doi:10.1016/j.biopsych.2016.08.31.

Egger, G., Roetzer, K.M., Noor, A., Lionel, A.C., Mahmood, H., Schwarzbraun, T., Boright, O., Mikhailov, A., Marshall, C.R., Windpassinger *et al.* (2014). Identification of risk genes for autism spectrum disorder through copy number variation analysis in Austrian families. *Neurogenetics*, 15(2): 117-127. doi:10.1007/s10048-014-0394-0.

Eicher, J.D., Powers, N.R., Miller, L.L., Akshoomoff, N., Amaral, D.G., Bloss, C.S., Libiger, O., Schork, N.J., Darst, B.F., Casey, B.J. *et al.* (2013). Genome-wide association study of shared components of reading disability and language impairment. *Genes, Brain and Behaviour*, 12(8): 792-801. doi:10.1111/gbb.12085.

Einfeld, S.L., Ellis, L.A., and Emerson, E. (2011). Comorbidity of intellectual disability and mental disorder in children and adolescents: a systematic review. *Journal of Intellectual and Developmental Disability*, 36(2): 137-143. doi:10.1080/13668250.2011.572548.

Elia, J., Gai, X., Xie, H.M., Perin, J.C., Geiger, E., Glessner, J.T., D'arcy, M., deBerardinis, R., Frackelton, E., Kim, C. *et al.* (2010). Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. *Molecular Psychiatry*, 15(6): 637-646. doi:10.1038/mp.2009.57.

Elia, J., Glessner, J.T., Wang, K., Takahashi, N., Shtir, C.J., Hadley, D., Sleiman, P.M., Zhang, H., Kim, C.E., Robison, R. *et al.* (2011). Genome-wide copy number variation study associates metabotropic glutamate receptor gene networks with attention deficit hyperactivity disorder. *Nature Genetics*, 44(1): 78-84. doi:10.1038/ng.1013.

Ellison, K.A., Fill, C.P., Terwilliger, J., DeGennaro, L.J., Martin-Gallardo, A., Anvret, M., Percy, A.K., Ott, J. and Zoghbi, H. (1992). Examination of X chromosome

markers in Rett syndrome: exclusion mapping with a novel variation on multilocus linkage analysis. *American Journal of Human Genetics*, 50(2): 278-287.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489(7414): 57-74. doi:10.1038/nature11247.

Enriquez-Barreto, L. and Morales, M. (2016). The PI3K signaling pathway as a pharmacological target in Autism related disorders and Schizophrenia. *Molecular and Cellular Therapy*, 4: 2. doi:10.1186/s40591-016-0047-9.

Epi4K Consortium and Epilepsy Phenome/Genome Project (EPGP) (2013). De novo mutations in epileptic encephalopathies. *Nature,* 501(7466): 217-221. doi:10.1038/nature12439.

Ercan-Sencicek, A.G., Stillman, A.A., Ghosh, A.K., Bilguvar, K., O'Roak, B.J., Mason, C.E., Abbott, T., Gupta, A., King, R.A., Pauls, D.L. *et al.* (2010). L-histidine decarboxylase and Tourette's syndrome. *New England Journal of Medicine*, 362(20): 1901-1908. doi:10.1056/NEJMoa0907006.

Evrony, G.D., Cordero, D.R., Shen, J., Partlow, J.N., Yu, T.W., Rodin, R.E., Hill, R.S., Coulter, M.E., Lam, A.N., Jayaraman, D. *et al.* (2017). Integrated genome and transcriptome sequencing identifies a noncoding mutation in the genome replication factor *DONSON* as the cause of microcephaly-micromelia syndrome. *Genome Research*, 7(8): 1323-1335. doi:10.1101/gr.219899.

EuroEPINOMICS-RES Consortium, Epilepsy Phenome/Genome Project and Epi4K Consortium (2014). De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. *American Journal of Human Genetics*, 95(4): 360-370. doi:10.1016/j.ajhg.2014.08.013.

EuroEPINOMICS-RES Consortium, Epilepsy Phenome/Genome Project and Epi4K Consortium (2017). De Novo Mutations in Synaptic Transmission Genes Including DNM1 Cause Epileptic Encephalopathies. *American Journal of Human Genetics*, 100(1): 179. doi:10.1016/j.ajhg.2016.12.012.

Fabrecat, A., Sidiropoulous, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S. *et al.* (2016). The Reactome pathway Knowledgebase. *Nucleic Acids Research*, 44(D1): D481-487. doi:10.1093/nar/gkv1351.

Fairthorne, J., Hammond, G., Bourke, J., Jacoby, P. and Leonard, H. (2014). Early mortality and primary causes of death in mothers of children with intellectual disability or autism spectrum disorder: a retrospective cohort study. *PLoS One*, 9(12): e113430. doi:10.1371/journal.pone.0113430.

Falconer, D.S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*, 29(1): 51-76. doi:10.1111/j.1469-1809.1965.tb00500.x

Fan, H.C., Gu, W., Wang, J., Blummenfeld, Y.J., El-Sayed, Y.Y. and Quake, S.R. (2012). Non-invasive prenatal measurement of the fetal genome. *Nature,* 487(7407): 320-324. doi:10.1038/nature11251.

Fanous, A.H., Zhou, B., Aggen, S.H., Bergen, S.E., Amdur, R.L., Duan, J., Sanders, A.R., Shi, J., Mowry, B.J., Olincy, A. *et al.* (2012). Genome-wide association study of clinical dimensions of schizophrenia: polygenic effect on disorganized symptoms. *American Journal of Psychiatry*, 169(12): 1309-1317. doi:10.1176/appi.ajp.2012.12020218.

Faraone, S.V., Biederman, J., Keenan, K. and Tsuang, M.T. (1991). Separation of DSM-III attention deficit disorder and conduct disorder: evidence from a family-genetic study of American child psychiatric patients. *Psychological Medicine*, 21(1):109-121.

Faraone, S.V., Perlis, R.H., Doyle, A.E., Smoller, J.W., Goralnick, J.J., Holmgren, M.A. and Sklar, P. (2005). Molecular genetics of attention-deficit/hyperactivity disorder. *Biological Psychiatry*, 57(11):1313-1323. doi:10.1016/j.biopsych.2004.11.024.

Fernandez, T.V., Sanders, S.J., Yurkiewicz, I.R., Ercan-Sencicek, A.G., Kim, Y.S., Fishman, D.O., Raubeson, M.J., Song, Y., Yasuno, K., Ho, W.S. *et al.* (2012). Rare copy number variants in tourette syndrome disrupt genes in histaminergic pathways and overlap with autism. *Biological Psychiatry*, 71(5): 392-402. doi:10.1016/j.biopsych.2011.09.034.

Ferreira, M.A., O'Donovan, M.C., Meng, Y.A., Jones, I.R., Ruderfer, D.M., Jones, L., Fan, J., Kirov, G., Perlis, R.H., Green, E.K. *et al.* (2008). Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nature Genetics*, 40(9): 1056-1058. doi:10.1038/ng.209.

Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R.M. and Carter, N.P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources. *American Journal of Human Genetics*, 84(4): 524-533. doi:10.1016/j.ajhg.2009.03.010.

Fischbach, G.D. and Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron,* 68(2): 192-195. doi:10.1016/j.neuron.2010.10.006.

Foong, J., Girdea, M., Stavropoulos, J. and Brudno, M. (2015). Prioritizing Clinically Relevant Copy Number Variation from Genetic Interactions and Gene Function Data. *PLoS One*, 10(10):e0139656. doi:10.1371/journal.pone.0139656.

Foulquier, F. Amyere, M., Jaerken, J., Zeevaert, R., Schollen, E., Race, V., Bammens, R., Morelle, W., Rosnoblet, C., Legrand, D. *et al.* (2012). TMEM165 deficiency causes a congenital disorder of glycosylation. *American Journal of Human Genetics*, 91(1): 15-26. doi:10.1016/j.ajhg.2012.05.002.

Freitag, C.M. (2007). The genetics of autistic disorders and its clinical relevance: a review of the literature. *Molecular Psychiatry*, 12(1): 2-22. doi:10.1038/sj.mp.4001896.

Fromer, M., Pocklington, A.J., Kavanagh, D.H., Williams, H.J., Dwyer, S., Gormley, P., Georgieva, L., Rees, E., Palta, P., Ruderfer, D.M. *et al.* (2014). De novo mutations in schizophrenia implicate synaptic networks. *Nature,* 506(7487): 179-184. doi:10.1038/nature12929.

Fuchsberger, C., Flannick, J., Teslovich, T.M., Mahajan, A., Agarwala, V., Gaulton, K.J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D.J. *et al.* (2016). *Nature,* 536(7614): 41-47. doi:10.1038/nature18642.

Gai, X., Xie, H.M., Perin, J.C., Takahashi, N., Murphy, K., Wenocur, A.S., D'arcy, M., O'Hara, R.J., Goldmuntz, E., Grice, D.E. *et al.* (2012). Rare structural variation of synapse and neurotransmission genes in autism. *Molecular Psychiatry*, 17(4): 402-411. doi:10.1038/mp.2011.10.

Garbett, K., Ebert, P.J., Mitchell, A., Lintas, C., Manzi, B., Mirnics, K. and Persico, A.M. (2008). Immune transcriptome alterations in the temporal cortex of subjects with autism. *Neurobiological Disease*, 30(3): 303-311. doi:10.1016/j.nbd.2008.01.012.

Gaugler, T., Klei, L., Sanders, S.J., Bodea, C.A., Goldberg, A.P., Lee, A.B., Mahajan, M., Manaa, D., Pawitan, Y., Reichert, J. *et al.* (2014). Most genetic risk for autism resides with common variation. *Nature Genetics*, 46(8): 881-885. doi:10.1038/ng.3039.

Gauthier, J. and Rouleau, G.A. (2012). De novo mutations in neurological and psychiatric disorders: effects, diagnosis and prevention. *Genome Medicine*, 4(9): 71. doi:10.1186/gm372.

Gauthier, J., Siddiqui, T.J., Huashan, P., Yokomaku, D., Hamdan, F.F., Champagne, N., Lapointe, M., Spiegelman, D., Noreau, A., Lafrenière, R.G. *et al.* (2011). Truncating mutations in NRXN2 and NRXN1 in autism spectrum disorders and schizophrenia. *Human Genetics*, 130(4): 563-573. doi:10.1007/s00439-011-0975-z.

Gazzellone, M.J., Zarrei, M., Burton, C.L., Ealker, S., Uddin, M., Shaheen, S.M., Coste, J., Rajendram, R., Schachter, R.J., Colasanto, M. *et al.* (2016). Uncovering obsessive-compulsive disorder risk genes in pediatric cohort by high-resolution analysis of copy number variation. *Journal of Neurodevelopmental Disorders*, 8: 36. doi.10.1186/s11689-016-9170-9.

Geller, D.A. (2006). Obsessive-compulsive and spectrum disorders in children and adolescents. *Psychiatric Clinics of North American*, 29(2): 353-370. doi:10.1016/j.psc.2006.02.012.

Gene Ontology Consortium. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1): 25-9.

Germine, L., Robinson, E.B., Smoller, J.W., Calkins, M.E., Moore, T.M., Hakonarson, H., Daly, M.J., Lee, P.H., Holmes, A.J., Buckner, R.L. *et al.* (2016). Association between polygenic risk for schizophrenia, neurocognition and social cognition across development. *Translational Psychiatry*, 6(10): e924. doi:10.1038/tp.2016.147.

Gershon, E.S. and Grennan, K.S. (2015). Genetic and genomic analyses as a basis for new diagnostic nosologies. *Dialogues in Clinical Neuroscience*, 17(1):69-78.

Geschwind, D.H. (2011). Genetics of autism spectrum disorders. *Trends in Cognitive Sciences*, 15(9): 409-416. doi:10.1016/j.tics.2011.07.003.

Geschwind, D.H. and State, M.W. (2015). Gene hunting in autism spectrum disorder: on the path to precision medicine. *Lancet Neurology,* 14(11): 1109-1120. doi:10.1016/S1474-4422(15)00044-7.

Gialluisi, A., Newbury, D.F., Wilcutt, E.G., Olson, R.K., DeFries, J.C., Brandler, W.M., Pennington, B.F., Smith, S.D., Scerri, T.S., Simpson, N.H. *et al.* (2014). Genome-wide screening for DNA variants associated with reading and language traits. *Genes, Brain and Behaviour*, 13(7): 686-701. doi:10.1111/gbb.12158.

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13(2):135-145. doi:10.1038/nrg3118.

Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A. *et al.* (2014). Genome sequencing identifies major causes of severe intellectual disability. *Nature,* 511(7509): 344-347. doi:10.1038/nature13394.

Gilman, S.R., Chang, J., Xu, B., Bawa, T.S., Gogos, J.A., Karayiorgou, M. and Vitkup, D. (2012). Diverse types of genetic variation converge on functional gene networks involved in schizophrenia. *Nature Neuroscience*, 15(12): 1723-1728. doi:10.1038/nn.3261.

Gilman, S.R., Iossifov, I., Levy, D., Ronemus, M., Wigler, M. and Vitkup, D. (2011). Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron,* 70(5): 898-907. doi:10.1016/j.neuron.2011.05.021.

Girirajan, S., Brkanac, Z., Coe, B.P., Baker, C., Vives, L., Vu, T.H., Shafer, N., Bernier, R., Ferrero, G.B., Silengo, M. *et al.* (2011). Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS Genetics*, 7(11): e1002334. doi:10.1371/journal.pgen.1002334.

Girirajan, S., Rosenfeld, J.A., Coe, B.P., Parikh, S., Friedman, N., Goldstein, A., Filipink, R.A., McConnell, J.S., Angle, B., Meschino, W.S. *et al.* (2012). Phenotypic heterogeneity of genomic disorders and rare copy-number variants. *New England Journal of Medicine*, 367(14): 1321-1331. doi:10.1056/NEJMoa1200395.

Girirajan, S., Rosenfeld, J.A., Cooper, G.M., Antonacci, F., Siswara, P., Itsara, A., Vives, L., Walsh, T., McCarthy, S.E., Baker, C. *et al.* (2010). A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nature Genetics*, 42: 203–209. doi:10.1038/ng.534.

Glessner, J.T., Wang, K., Cai, G., Korvatska, O., Kim, C.E., Wood, S., Zhang, H., Estes, A., Brune, C.W., Bradfield, J.P. *et al.* (2009). Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature,* 459(7246): 569-573. doi:10.1038/nature07953.

Goines, P. and Van der Water, J. (2010) The Immune System's Role in the Biology of Autism. Current Opinion in *Neurology,* 23(2): 111–117.

Goodbourn, P.T., Bosten, J.M., Bargary, G., Hogg, R.E., Lawrance-Owen, A.J. and Mollon, J.D. (2014). Variants in the 1q21 risk region are associated with a visual endophenotype of autism and schizophrenia. *Genes, Brain and Behaviour*, 13(2): 144-151. doi:10.1111/gbb.12096.

Gotha, K., Pickles, A. and Lord, C. (2009). Standardising ADOS scores for a measure of severity in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 39(5): 693-705. doi:10.1007/s10803-008-0674-3.

Gottesman, I.I. and Gould, T.D. (2003). The endophenotype concept in psychiatry: etymology and strategic intentions. *American Journal of Psychiatry*, 160(4): 636-645.

Gratten, J., Wray, N.R., Keller, M.C. and Visscher, P.M. (2014). Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nature Neuroscience*, 17(6): 782-790. doi:10.1038/nn.3708.

Grayton, H.M., Fernandes, C., Rujescu, D. and Collier, D.A. (2012). Copy number variations in neurodevelopmental disorders. *Progress in Neurobiology*, 99(1): 81-91. doi:10.1016/j.pneurobio.2012.07.005.

Greer, P.L. and Greenberg, M.E. (2008). From synapse to nucleus: calcium-dependent gene transcription in the control of synapse development and function. *Neuron,* 59(6): 846-860. doi:10.1016/j.neuron.2008.09.002.

Gregory, A., Polster, B.J. and Hayflick, S.J. (2009). Clinical and genetic delineation of neurodegeneration with brain iron accumulation. *Journal of Medical Genetics*, 46:73-80 doi:10.1136/jmg.2008.061929.

Grice, S.J., Liu, J.L. and Webber, C. (2015). Synergistic interactions between Drosophila orthologues of genes spanned by de novo human CNVs support multiple-hit models of autism. *PLoS Genetics*, 11(3):e1004998. doi:10.1371/journal.pgen.1004998.

Grillo, E., Lo Rizzo, C., Bianciardi, L., Bizzarri, V., Baldassarri, M., Spiga, O., Furini, S., De Felice, C., Signorini, C., Leoncini, S. *et al.* (2013). Revealing the complexity of a monogenic disease: rett syndrome exome sequencing. *PLoS One*, 8(2): e56599. doi:10.1371/journal.pone.0056599.

Guilmatre, A., Dubourg, C., Mosca, A.L., Legallic, S., Goldenberg, A., Drouin-Garraud, V., Layet, V., Rosier, A., Briault, S., Bonnet-Brilhault, F. *et al.* (2009). Recurrent rearrangements in synaptic and neurodevelopmental genes and shared biologic pathways in schizophrenia, autism, and mental retardation. *Archives of General Psychiatry*, 66(9): 947-956. doi:10.1001/archgenpsychiatry.2009.80.

Gupta, S., Ellis, S.E., Ashar, F.N., Moes, A., Bader, J.S., Zhan, J., West, A.B. and Arking, D.E. (2014). Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nature Communications*, 5: 5748. doi:10.1038/ncomms6748.

Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., Sakaguchi, A.Y. *et al.* (1983). A polymorphic DNA marker genetically linked to Huntington's disease. *Nature,* 306(5940): 234-238.

Hahn, M.W. and Kern, A.D. (2004). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution*, 22(4): 803-806. doi:10.1093/molbev/msi072.

Hamdan, F.F., Gauthier, J., Araki, Y., Lin, D.T., Yoshizawa, Y., Higashi, K., Park, A.R., Spiegelman, D., Dobrzeniecka, S., Piton, A. *et al.* (2011). Excess of de novo deleterious mutations in genes associated with glutamatergic systems in nonsyndromic intellectual disability. *American Journal of Human Genetics*, 88(3): 306-316. doi:10.1016/j.ajhg.2011.02.001.

Hamdan, F.F., Srour, M., Capo-Chichi, J.M., Daoud, H., Nassif, C., Patry, L., Massicotte, C., Ambalavanan, A., Spiegelman, D., Diallo, O., *et al.* (2014). De novo mutations in moderate or severe intellectual disability. *PLoS Genetics*, 10(10): e1004772. doi:10.1371/journal.pgen.1004772.

Hamlyn, J., Duhig, M., McGrath, J. and Scott, J. (2013). Modifiable risk factors for schizophrenia and autism--shared risk factors impacting on brain development. *Neurobiology of Disease*, 53: 3-9. doi:10.1016/j.nbd.2012.10.023.

Harding, A.E. and Rosenberg, R.N. A neurologic gene map. In: The Molecular and Genetic Basis of Neurological Disease. Boston: Butterworth-Heinemann, 1993, pp. 21–24.

Hardy, J. and Singleton, A. (2009). Genome-wide association studies and human disease. *New England Journal of Medicine*, 360(17): 1759-1768. doi:10.1056/NEJMra0808700.

Hartong, D.T., Berson, E.L., Dryja, T.P. (2006). Retinitis pigmentosa. *The Lancet*, 368(9549): 1795-809. doi:10.1016/S0140-6736(06)69740-7.

Hastings, P.J., Lupski, J.R., Rosenberg, S.M. and Ira, G. (2009). Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10(8): 551-564. doi:10.1038/nrg2593.

Hayflick, S.J. (2003). Unraveling the Hallervorden-Spatz syndrome: pantothenate kinase–associated neurodegeneration is the name. Current Opinion in *Pediatrics*, 15(6): 572-7.

Hedrick, P.W. (2006). Genetic Polymorphism in Heterogeneous Environments: The Age of Genomics. *Annual Review of Ecology, Evolution and Systematics*, 37: 67-93. doi:10.1146/annurev.ecolsys.37.091305.110132.

Hehir-Kwa, J.Y., Wieskamp, N., Webber, C., Pfundt, R., Brunner, H.G., Gilissen, C., de Vries, B.B.A., Ponting, C. and Veltmann, J.A. (2010). Accurate Distinction of Pathogenic from Benign CNVs in Mental Retardation. *PLoS Computational Biology*, 6(4): e1000752. doi:10.1371/journal.pcbi.1000752.

Helsmoortel, C., Vulto-van Silfhout, A.T., Coe, B.P., Vandeweyer, G., Rooms, L., van den Ende, J., Schuurs-Hoeijmakers, J.H., Marcelis, C.L., Willemsen, M.H., Vissers, L.E., Yntema, H.G., Bakshi, M., Wilson, M., Witherspoon, K.T. *et al.* (2014). A SWI/SNF-related autism syndrome caused by de novo mutations in ADNP. *Nature Genetics*, 46(4): 380-384. doi:10.1038/ng.2899.

Hemani, G., Shakhbazov, K., Westra, H.J., Esko, T., Henders, A.K., McRae, A.F., Yang, J., Gibson, G., Martin, N.G., Metspalu, A. *et al.* (2014). Detection and replication of epistasis influencing transcription in humans. *Nature,* 508(7495): 249-253. doi:10.1038/nature13005.

Herculano-Houzel, S. (2009). The Human Brain in Numbers: A Linearly Scaled-up Primate Brain. *Frontiers in Human Neuroscience*, 3:31. doi:10.3389/neuro.09.031.2009.

He, X., Sanders, S.J., Liu, L., De Rubeis, S., Lim, E.T., Sutcliffe, J.S., Schellenberg, G.D., Gibbs, R.A., Daly, M.J., Buxbaum, J.D. *et al.* (2013). Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genetics*, 9(8): e1003671. doi:10.1371/journal.pgen.1003671.

Hirvikoski, T., Mittendorfer-Rutz, E., Boman, M., Larsson, H., Lichtenstein, P. and Bölte, S. (2016). Premature mortality in autism spectrum disorder. *The British Journal of Psychiatry*, 208(3): 232-238. doi:10.1192/bjp.bp.114.160192.

Hoischen, A., Krumm, N. and Eichler, E.E. (2014). Prioritization of neurodevelopmental disease genes by discovery of new mutations. *Nature Neuroscience*, 17(6): 764-772. doi:10.1038/nn.3703.

Homem, C.C., Repic, M. and Knoblich, J.A. (2015). Proliferation control in neural stem and progenitor cells. *Nature Reviews Neuroscience*, 16(11): 647-659. doi:10.1038/nrn4021.

Hong, S.E., Shugart, Y.Y., Huang, D.T., Shahwan, S.A., Grant, P.E., Hourihane, J.O., Martin, N.D. and Walsh, C.A. (2001). Autosomal recessive lissencephaly with cerebellar hypoplasia is associated with human RELN mutations. *Nature Genetics*, 26(1): 93-96. doi:10.1038/79246.

Honti, F., Meader, S. and Webber, C. (2014). Unbiased functional clustering of gene variants with a phenotypic-linkage network. *PLoS Computational Biology*, 10(8): e1003815. doi:10.1371/journal.pcbi.1003815.

Hoogenraad, C.C. and Akhmanova, A. (2010). Dendritic spine plasticity: new regulatory roles of dynamic microtubules. *Neuroscientist*, 16(6): 650-651. doi:10.1177/1073858410386357.

Hormozdiari, F., Penn, O., Borenstein, E. and Eichler, E.E. (2015). The discovery of integrated gene networks for autism and related disorders. *Genome Research*, 25(1): 142-154. doi:10.1101/gr.178855.114.

Hosking, F.J., Carey, I.M., Shah, S.M., Harris, T., DeWilde, S., Beighton, C. and Cook, D.G. (2016). Mortality Among Adults With Intellectual Disability in England: Comparisons With the General Population. *American Journal of Public Health*, 106(8): 1483-1490. doi:10.2105/AJPH.2016.303240.

Hou, L., Bergen, S.E., Akula, N., Song, J., Hultman, C.M., Landén, M., Adli, M., Alda, M., Ardau, R., Arias, B. *et al.* (2016). Genome-wide association study of 40,000 individuals identifies two novel loci associated with bipolar disorder. *Human Molecular Genetics*, 25(15): 3383-3394. doi:10.1093/hmg/ddw181.

Huang, A.Y., Yu, D., Davis, L.K., Sul, J.-H., Tsetsos, F., Ramensku, V., Zelaya, I., Ramos, E.M., Osiecki, L., Chen, J.A. *et al.* (2016). Rare copy number variants in NRXN1 amd CNTN6 increase risk for Tourette syndrome. *Neuron*, 94(6):1101-111.e7. doi:10.1016/j.neuron.2017.06.010.

Huang, H.T., Kathrein, K.L., Barton, A., Gitlin, Z., Huang, Y.H., Ward, T.P., Hofmann, O., Dibiase, A., Song, A., Tyekucheva, S. *et al.* (2013). A network of epigenetic regulators guides developmental haematopoiesis in vivo. *Nature Cell Biology*, 15: 1516-1525.

Huang, N., Lee, I., Marcotte, E.M. and Hurles, M.E. (2010). Characterising and Predicting Haploinsufficiency in the Human Genome. *PLoS Genetics*, 6(10): e1001154. doi:10.1371/journal.pgen.1001154.

Huang, Y., Yu, S., Wu, Z. and Tang, B. (2014). Genetics of hereditary neurological disorders in children. *Translational Pediatrics*, 3(2): 108-119. doi:10.3978/j.issn.2224-4336.2014.03.04.

Hu, H., Haas, S.A., Chelly, J., Van Esch, H., Raynaud, M., de Brouwer, A.P., Weinert, S., Froyen, G., Frints, S.G., Laumonnier, F. *et al.* (2016). X-exome sequencing of 405 unresolved families identifies seven novel intellectual disability genes. *Molecular Psychiatry*, 21(1): 133-148. doi:10.1038/mp.2014.193.

Hu-Lince, D., Craig, D.W., Huentelman, M.J. and Stephan, D.A. (2005). The Autism Genome Project: goals and strategies. *American Journal of Pharmacogenomics*, 5(4): 233-246.

Hultman, C.M., Sparen, P. and Cnattingius, S. (2002). Perinatal risk factors for infantile autism. *Epidemiology*, 13(4): 417-423.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T.K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S. *et al.* (2012). InterPro in 2011: new developments in the family and protein prediction database. *Nucleic Acids Research*, 40(Database issue): D306-D312. doi:10.1093/nar/gkr948.

Hunt, K.A., Mistry, V., Bockett, N.A., Ahmad, T., Ban, M., Barker, J.N., Barrett, J.C., Blackburn, H., Brand, O., Burren, O. *et al.* (2013). Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature,* 498(7453): 232-235. doi:10.1038/nature12170.

Huttlin, E.L., Ting, L., Bruckner, R.J., Gebreab, F., Gygi, M.P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K. *et al.* (2015). The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell,* 162(2): 425-440. doi:10.1016/j.cell.2015.06.043.

Huynh, L. and Hormozdiari, F. (2017). Ultra-accurate complex disorder prediction: Case study of neurodevelopmental disorders. *BioRxiv*, 129775. doi:10.1101/129775.

Imbrici, P., Camerino, D.C. and Tricarico, D. (2013). Major channels involved in neuropsychiatric disorders and therapeutic perspectives. *Frontiers in Genetics*, 4: 76. doi:10.3389/fgene.2013.00076.

International HapMap Consortium. (2003). The International HapMap Project. *Nature,* 426(6968): 789-796. doi:10.1038/nature02168.

International Schizophrenia Consortium, Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F. and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature,* 460(7256): 748-752. doi:10.1038/nature08185.

Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E. *et al.* (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature,* 515(7526): 216-221. doi:10.1038/nature13908.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.H., Narzisi, G., Leotta, A. *et al.* (2012). De novo gene disruptions in children on the autistic spectrum. *Neuron,* 74(2): 285-299. doi:10.1016/j.neuron.2012.04.009.

Itsara, A., Wu, H., Smith, J.D., Nickerson, D.A., Romieu, I., London, S.J. and Eichler, E.E. (2010). De novo rates and selection of large copy number variation. *Genome Research*, 20(11): 1469-1481. doi:10.1101/gr.107680.110.

Jamra, R.A., Philippe, O., Raas-Rothschild, A., Eck, S.H., Graf, E., Buchert, R., Borck, G., Ekici, A., Brockschmidt, F.F., Nothen, M.M. *et al.* (2011). Adaptor Protein Complex 4 Deficiency Causes Severe Autosomal-Recessive Intellectual Disability, Progressive Spastic Paraplegia, Shy Character, and Short Stature. *American Journal of Human Genetics*, 88(6): 788-795. doi:10.1016/j.ajhg.2011.04.019.

Jansen, A., Dieleman, G.C., Smit, A.B., Verhage, M., Verhulst, F.C., Polderman, T.J. and Posthuma, D. (2017). Gene-set analysis shows an association between FMRP targets and autism spectrum disorder. *European Journal of Human Genetics*, 25(7): 863-868. doi:10.1038/ejhg.2017.55.

Jarick, I., Volckmar, A.L., Pütter, C., Pechlivanis, S., Nguyen, T.T., Dauvermann, M.R., Beck, S., Albayrak, Ö., Scherag, S., Gilsbach, S. *et al.* (2014). Genome-wide analysis of rare copy number variations reveals PARK2 as a candidate gene for attention-deficit/hyperactivity disorder. *Molecular Psychiatry*, 19(1): 115-121. doi:10.1038/mp.2012.161.

Jeong, H., Mason, S.P., Barabasi, A.L. and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. *Nature,* 411(6833): 41-42. doi:10.1038/35075138.

Jeste, S.S. (2011). The Neurology of Autism Spectrum Disorders. Current Opinions in *Neurology,* 24(2): 132-139. doi:10.1097/WCO.0b013e3283446450.

Jeyabalan, N. and Clement, J.P. (2016). SYNGAP1: Mind the Gap. *Frontiers in Cellulalr Neuroscience*, 10: 32. doi:10.3389/fncel.2016.00032.

Jiang, Y.H., Sahoo, T., Michaelis, R.C., Bercovich, D., Bressler, J., Kashork, C.D., Liu, Q., Shaffer, L.G., Schroer, R.J., Stockton, D.W. *et al.* (2004). A mixed epigenetic/genetic model for oligogenic inheritance of autism with a limited role for UBE3A. American *Journal of Medical Genetics* A, 131(1): 1-10. doi:10.1002/ajmg.a.30297.

Ji, J., Lee, H., Argiropoulos, B., Dorrani, N., Mann, J., Martinez-Agosto, J.A., Gomez-Ospina, N., Gallant, N., Bernstein, J.A., Hudgins, L. *et al.* (2015). DYRK1A haploinsufficiency causes a new recognizable syndrome with microcephaly, intellectual disability, speech impairment, and distinct facies. *European Journal of Human Genetics*, 23(11): 1473-1481. doi:10.1038/ejhg.2015.71.

Jones, R.M., Cadby, G., Melton, P.E., Abraham, L.J., Whitehouse, A.J. and Moses, E.K. (2013). Genome-wide association study of autistic-like traits in a general population study of young adults. *Frontiers in Human Neuroscience*, 7: 658. doi:10.3389/fnhum.2013.00658.

Jou, S.H., Chiu, N.Y. and Liu, C.S. (2009). Mitochondrial dysfunction and psychiatric disorders. *Chang Gung Medical Journal*, 32(4): 370-379.

Kanehisa, M. and Goto, S. (2000). Kyoto Encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1): 28-30.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K. (2017). KEGG: a new perspective on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1): D353-36. doi:10.1093/nar/gkw1092.

Kanne, S.M., Gerber, A.J., Quirmbach, L.M., Sparrow, S.S., Cicchetti, D.V. and Saulnier, C.A. (2011). The role of adaptive behaviour in autism spectrum disorders: implications for functional outcome. *Journal of Autism and Neurodevelopmental disorders*, 41(8): 1007-1018. doi:10.1007/s10803-010-1126-4.

Katayama, K., Yamada, K., Ornthanalai, V.G., Inoue, T., Ota, M., Murphy, N.P. and Aruga J. (2010). Slitrk1-deficient mice display elevated anxiety-like behavior and noradrenergic abnormalities. *Molecular Psychiatry*, 15(2): 177-184. doi:10.1038/mp.2008.97.

Katayama, Y., Nishiyama, M., Shoji, H., Ohkawa, Y., Kawamura, A., Sato, T., Suyama, M., Takumi, T., Miyakawa, T. and Nakayama, K.I. (2016). CHD8 haploinsufficiency results in autistic-like phenotypes in mice. *Nature,* 537(7622): 675-679. doi:10.1038/nature19357.

Kaufman, L., Ayub, M. and Vincent, J.B. (2010). The genetic basis of non-syndromic intellectual disability: a review. *Journal of Neurodevelopmental Disorders*, 2(4): 182–209. doi:10.1007/s11689-010-9055-2.

Kaufmann, U., Zuppingerm C., Waibler, Z., Rudiger, M., Urbich, C., Martin, B., Jockusch, B.M., Eppenberger, H. and Starzinski-Powitz, A. (2000). The armadillo repeat region targets ARVCF to cadherin-based cellular junctions. *Journal of Cell Science*, 113(Pt2): 4121-4135.

Kavanagh, D.H., Tansey, K.E., O'Donovan, M.C. and Owen, M.J. (2014). Schizophrenia genetics: emerging themes for a complex disorder. *Molecular Psychiatry*, 20:72-76. doi:10.1038/mp.2014.148.

Keil, A., Daniels, J.L., Forssen, U., Hultman, C., Cnattingius, S., Soderberg, K.C., Feychting, M. and Sparen, P. (2010). Parental autoimmune diseases associated with autism spectrum disorders in offspring. *Epidemiology*, 21(6): 805-808. doi:10.1097/EDE.0b013e3121f26e3f.

Keller, M.C. and Miller, G. (2006). Resolving the paradox of common, harmful, heritable mental disorders: which evolutionary genetic models work best? *Behavioural and Brain Science*s, 29(4): 385-404. doi:10.1017/S0140525X06009095.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, 12(6): 996-1006. doi:10.1101/gr.229102.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U. *et al.* (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, 40(Database issue): D841-D846. doi:10.1093/nar/gkr1088.

Khare, S.P., Habib, F., Sharma, R., Gadewal, N., Gupta, S. and Galande, S. (2011). HIstome: a relational knowledgebase of human histone proteins and histone modifying enzymes. *Nucleic Acids Research*, Database issue. doi:10.1093/nar/gkr1125.

Kim, E., Hwang, S., Kim, H., Shim, H., Kang, B., Yang, S., Shim. J.H., Shin, S.Y., Marcotte, E.M. and Lee, I. (2016). MouseNet v2: a database of gene networks for studying the laboratory mouse and eight other model vertebrates. *Nucleic Acids Research*, 44(Database issue): D848-D854. doi:10.1093/nar/gkv/1155.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M. and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3): 310-315. doi:10.1038/ng.2892.

Kirov, G. (2015). CNVs in neuropsychiatric disorders. *Human Molecular Genetics*, 24(R1): R45-R49. doi:10.1093/hmg/ddv253.

Kirov, G., Pocklington, A.J., Holmans, P., Ivanov, D., Ikeda, M., Ruderfer, D., Moran, J., Chambert, K., Toncheva, D., Georgieva, L. *et al.* (2012). De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Molecular Psychiatry*, 17(2): 142-153. doi:10.1038/mp.2011.154.

Kirschner, M.W. and Gerhart, J.C. (2006). The Plausibility of Life. New Haven: Yale University Press.

Kiztman, J.O., Snyder, M.W., Ventura, M., Lewis, A.P., Qiu, R., Simmons, L.E., Gammill, H.S., Rubens, C.E., Santillan, D.A., Murray, J.C. *et al.* (2012). Noninvasive whole-genome sequencing of a human foetus. *Science Translational Medicine*, 4(137): 137ra76. doi:10.1126/scitranslmed.3004323.

Klei, L., Sanders, S.J., Murtha, M.T., Hus, V., Lowe, J.K., Willsey, A.J., Moreno-De-Luca, D., Yu, T.W., Fombonne, E., Geschwind, D. *et al.* (2012). Common genetic variants, acting additively, are a major source of risk for autism. *Molecular Autism*, 3(1): 9. doi:10.1186/2040-2392-3-9.

Kochinke, K., Zweier, C., Nijhof, B., Fenckova, M., Cizek, P., Honti, F., Keerthikumar, S., Oortveld, M.A., Kleefstra, T., Kramer, J.M. *et al.* (2016). Systematic Phenomics Analysis Deconvolutes Genes Mutated in Intellectual Disability into Biologically Coherent Modules. *American Journal of Human Genetics*, 98(1): 149-164. doi:10.1016/j.ajhg.2015.11.024.

Kodera, H., Nakamura, K., Osaka, H., Maegaki, Y., Haginoya, K., Mizumoto, S., Kato, M., Okamoto, N., Iai, M., Kondo, Y. *et al.* (2013). De novo mutations in SLC35A2 encoding a UDP-galactose transporter cause early-onset epileptic encephalopathy. *Human Mutation*, 34(12): 1708-1714. doi:10.1002/humu.22446.

Kohane, I.S., McMurry, A., Wber, G., MacFadden, D., Rappaport, L., Kunkel, L., Bickelm J., Wattanasin, N., Spence, S., Murphy, S. and Churchill, S. (2012). The Co-

Morbidity Burden of Children and Young Adults with Autism Spectrum Disorders. *PLoS One*, 7(4): e33224. doi:10.1371/journal.pone.0033224.

Köhler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S.M., Boerkoel, C.F., Boycott, K.M. *et al.* (2017). The Human Phenotype Ontology in 2017. *Nucleic Acids Research*, 45(D1): D865-D876. doi:10.1093/nar/gkw1039.

Koolen, D.A., Pfundt, R., de Leeuw, N., Hehir-Kwa, J.Y., Nullesen, W.M., Neefs, I., Scheltinga, I., Sistermans, E., Smeets, D. Brunner, H.G. e*t al.* (2009). Genomic microarrays in mental retardation: a practical workflow for diagnostic applications. *Human mutation*, 30(3): 283-292. doi:10.1002/humu.20883.

Krishnan, A., Zhang, R., Yao, V., Theesfeld, C.L., Wong, A.K., Tadych, A., Volfovsky, N., Packer, A., Lash, A. and Troyanskaya, O.G. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nature Neuroscience*, 19(11): 1454-1462. doi:10.1038/nn.4353.

Krumm, N., Turner, T.N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B.P., Stessman, H.A., He, Z.X. *et al.* (2015). Excess of rare, inherited truncating mutations in autism. *Nature Genetics*, 47(6): 582-588. doi:10.1038/ng.3303.

Kryukov, G.V., Pennacchio, L.A. and Sunyaev, S.R. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *American Journal of Human Genetics*, 80(4): 727-739. doi:10.1086/513473.

Kubota, Y., Ito, C., Sakurai, E., Sakurai, E., Watanabe, T. and Ohtsu, H. (2002) Increased methamphetamine-induced locomotor activity and behavioral sensitization in histamine-deficient mice. *Journal of Neurochemistry*, 83(4): 837-45.

Kumar, R.A., KaraMohamed, S., Sudi, J., Conrad, D.F., Brune, C., Badner, J.A., Gilliam, T.C., Nowak, N.J., Cook, E.H.Jr,, Dobyns, W.B. and Christian, S.L. (2008). Recurrent 16p11.2 microdeletions in autism. *Human Molecular Genetics*, 17(4): 628-538. doi:10.1093/hmg/ddm376.

Kumar, R., Ha, T., Pham, D., Shaw, M., Mangelsdorf, M., Friend, K.L., Hobson, L., Turner, G., Boyle, J., Field, M. *et al.* (2016). A non-coding variant in the 5' UTR of DLG3 attenuates protein translation to cause non-syndromic intellectual disability. *European Journal of Human Genetics*, 24(11): 1612-1616. doi:10.1038/ejhg.2016.46.

Kuo, P.H., Chuang, L.C., Su, M.H., Chen, C.H., Chen, C.H., Wu, J.Y., Yen, C.J., Wu, Y.Y., Liu, S.K., Chou, M.C. iet al. (2015). Genome-Wide Association Study for Autism Spectrum Disorder in Taiwanese Han Population. *PLoS One*, 10(9): e0138695. doi:10.1371/journal.pone.0138695.

Kurbatova, N., Mason, J.C., Morgan, H., Meehan, T.F. and Karp, N.A. (2015). PhenStat: A Tool Kit for Standardized Analysis of High Throughput Phenotypic Data. *PLoS One*, 10(7): e0131274. doi:10.1371/journal.pone.0131274.

Lai, M.C., Lombardo, M.V. and Baron-Cohen, S. (2014). Autism. *The Lancet*, 383(9920): 896-910. doi:10.1016/S0140-6736(13)61539-1.

Lajonchere, C.M. and the AGRE Consortium (2010). Changing the landscape of autism research: the autism genetic resource exchange. *Neuron,* 68(2): 187-191. goi:10.1016/j.neuron.2010.10.009.

Lander, E.S. (1996). The new genomics: global views of biology. *Science*, 274(5287): 536-539.

Lane, J.M., Liang, J., Vlasac, I., Anderson, S.G., Bechtold, D.A., Bowden, J., Emsley, R., Gill, S., Little, M.A., Luik, A.I. *et al.* (2017). Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits. *Nature Genetics*, 49(2): 274-281. doi:10.1038/ng.3749.

Lantieri, F., Glessner, J.T., Hakonarson, H., Elia, J. and Devoto, M. (2010). Analysis of GWAS top hits in ADHD suggests association to two polymorphisms located in genes expressed in the cerebellum. American *Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 153B(6): 1127-1133. doi:10.1002/ajmg.b.31110.

Laplante, M. and Sabatini, D.M. (2012). mTOR signalling in growth control and disease. *Cell*, 149(2): 274-293. doi:10.1016/j.cell.2012.03.017.

LaSalle, J.M. (2013). Autism genes keep turning up chromatin. *OA Autism*, 1(2): 14.

Laumonnier, F., Bonnet-Brilhault, F., Gomot, M., Blanc, R., David, A., Moizard, M.P., Raynaud, M., Ronce, N., Lemonnier, E., Calvas, P. *et al.* (2004). X-linked mental retardation and autism are associated with a mutation in the NLGN4 gene, a member of the neuroligin family. *American Journal of Human Genetics*, 74(3): 552-557. doi:10.1086/382137.

Lauritsen, M.B., Mors, O., Mortensen, P.B. and Ewald, H. (2002). Medical disorders among inpatients with autism in Denmakr according to ICD-8: a nationwide register-based study. *Journal of Autism and Developmental Disorders*, 32(2): 115-119.

Lawson-Yuen, A., Saldivar, J.S., Sommer, S. and Picker, J. (2008). Familial deletion within NLGN4 associated with autism and Tourette syndrome. *European Journal of Human Genetics*, 16(5): 614-618. doi:10.1038/sj.ejhg.5202006.

Leblond, C.S., Heinrich, J., Delorme, R., Proepper, C., Betancur, C., Huguet, G., Konyukh, M., Chaste, P., Ey, E., Rastam, M. *et al.* (2012). Genetic and functional analyses of SHANK2 mutations suggest a multiple hit model of autism spectrum disorders. *PLoS Genetics*, 8(2): e1002521. doi:10.1371/journal.pgen.1002521.

Leckman, J.F., Bloch, M.H., Smith, M.E., Larabi, D. and Hampson, M. (2010). Neurobiological Substrates of Tourette's Disorder. *Journal of Child and Adolescent Psychopharmacology*, 20(4): 237-247. doi:10.1089/cap.2009.0118

Lee, J.M. and Sonnhammer, E.L. (2003). Genomic gene clustering analysis of pathways in eukaryotes. *Genome Research*, 13(5): 875-882. doi: 10.1101/gr.737703.

Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J. and Oavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome Research*, 14(6): 1085-1094. doi:10.1101/gr.1910904.

Lefebvre, J.L., Kostadinov, D., Chen, W.V., Maniatis, T. and Sanes, J.R. (2012). Protocadherins mediate dendritic self-avoidance in the mammalian nervous system. *Nature,* 488(7412): 517-521. doi:10.1038/nature11305.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J. *et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature,* 536(7616): 285-291. doi:10.1038/nature19057.

Leppa, V.M., Kravitz, S.N., Martin, C.L., Andrieux, J., Le Caignec, C., Martin-Coignard, D., DyBuncio, C., Sanders, S.J., Lowe, J.K., Cantor, R.M. and Geschwind, D.H. (2016). Rare Inherited and De Novo CNVs Reveal Complex Contributions to ASD Risk in Multiplex Families. *American Journal of Human Genetics*, 99(3): 540-554. doi:10.1016/j.ajhg.2016.06.036.

Levy, D., Ronemus, M., Yamrom, B., Lee, Y.H., Leotta, A., Kendall, J., Marks, S., Lakshmi, B., Pai, D., Ye, K., Buja, A. *et al.* (2011). Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron,* 70(5): 886-897. doi:10.1016/j.neuron.2011.05.015.

Lewis, B.A. and Thompson, L.A. (1992). A study of developmental speech and language disorders in twins. *Journal of Speech and Hearing Research*, 35(5): 1086-1094.

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A.P., Santonico, E. *et al.* (2012). MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research*, 40(Database issue): D857-D861. doi:10.1093/nar/gkr930.

Lichter, D.G. and Finnegan, S.G. (2015). Influence of gender on Tourette syndrome beyond adolescence. *European Psychiatry*, 30(2): 334-340. doi:10.1016/j.eurpsy.2014.07.003.

Li, J., Cai, T., Jiang, Y., Chen, H., He, X., Chen, C., Li, X., Shao, Q., Ran, X., Li, Z. *et al.*, (2016). Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. *Molecular Psychiatry*, 21(2): 290-297. doi:10.1038/mp.2015.40.

Lima Lde, A., Feio-dos-Santos, A.C., Belangero, S.I., Gadelha, A., Bressan, R.A., Salum, G.A., Pan, P.M., Moriyama, T.S., Graeff-Martins, A.S., Tamanaha, A.C. *et al.* (2016). An integrative approach to investigate the respective roles of single-nucleotide variants and copy-number variants in Attention-Deficit/Hyperactivity Disorder. *Scientific Reports*, 6: 22851. doi:10.1038/srep22851.

Lim, E.T., Raychaudhuri, S., Sanders, S.J., Stevens, C., Sabo, A., MacArthur, D.G., Neale, B.M., Kirby, A., Ruderfer, D.M., Fromer, M. *et al.* (2013). Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. *Neuron,* 77(2): 235-242. doi:10.1016/j.neuron.2012.12.029.

Lin, S., Lin, Y., Nery, J.R., Urich, N., Breschi, A., Davis, C.A., Dobin, A., Zaleski, C., Beer, M.A., Chapman, W.C. *et al.* (2014). Comparison of the transcriptional landscapes between human and mouse tissues. *PNAS*, 111(48): 17224-17229. doi:10.1073/pnas.1413624111.

Lionel, A.C., Vaags, A.K., Sato, D., Gazzellone, M.J., Mitchell, E.B., Chen, H.Y., Costain, G., Walker, S., Egger, G., Thiruvahindrapuram, B. *et al.* (2013). Rare exonic deletions implicate the synaptic organizer Gephyrin (GPHN) in risk for autism, schizophrenia and seizures. *Human Molecular Genetics*, 22(10): 2055-2066. doi:10.1093/hmg/ddt056.

Li, T., Wernersson, R., Hansen, R.B., Horn, H., Mercer, J., Slodkowicz, G., Workman, C.T., Rigina, O., Rapacki, K., Stærfeldt, H.H. *et al.* (2017). A scored human protein-protein interaction network to catalyze genomic interpretation. *Nature Methods*, 14(1): 61-64. doi:10.1038/nmeth.4083.

Liu, L., Lei, J., Sanders, S.J., Willsey, A.J., Kou, Y., Cicek, A.E., Klei, L., Lu, C., He, X., Li, M. *et al.* (2014). DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. *Molecular Autism*, 5(1): 22. doi:10.1186/2040-2392-5-22.

Liu, X-Q., Paterson, A.D., Szatmari, P. and the Autism Genome Project (2008). Genome-wide linkage analyses of quantitative and categorical autism subphenotypes. *Biological Psychiatry*, 64(7): 561-570. Doi:10.1016/j.biopsych.2008.05.023.

Liu, X., Shimada, T., Otowa, T., Wu, Y.Y., Kawamura, Y., Tochigi, M., Iwata, Y., Umekage, T., Toyota, T., Maekawa, M. *et al.* (2016a). Genome-wide Association Study of Autism Spectrum Disorder in the East Asian Populations. *Autism Research*, 9(3): 340-349. doi:10.1002/aur.1536.

Liu, X., Zhou, Z.W. and Wang, Z.Q. (2016b). The DNA damage response molecule MCPH1 in brain development and beyond. *Acta Biochimica et Biophysica Sinica (Shanghai)*, 48(7): 678-685. doi:10.1093/abbs/gmw048.

Loh, P.R., Danecek, P., Palamara, P.F., Fuchsberger, C., Reshef, A.Y., Finucane, K.H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R. *et al.* (2016). Reference-based phasing using the haplotype reference consortium panel. *Nature Genetics*, 48(11): 1443-1448. doi:10.1038/ng.3679.

Lonsdale, J., Thomas, J., Salvatore, M., Philips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N. *et al.* (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45: 580–585. doi:10.1038/ng.2653.

Lubs, H.A., Stevenson, R.E. and Schwartz, C.E. (2012). Fragile X and X-linked intellectual disability: four decades of discovery. *American Journal of Human Genetics*, 90(4): 579-590. doi:10.1016/j.ajhg.2012.02.018.

Luciano, M., Evans, D.M., Hansell, N.K., Medland, S.E., Montgomery, G.W., Martin, N.G., Wright, M.J. and Bates, T.C. (2013). A genome-wide association study for reading and language abilities in two population cohorts. *Genes, Brain and Behaviour*, 12(6): 645-652. doi:10.1111/gbb.12053.

Lukk, M., Kapushesky, M., Nikkila, J., Parkinson, H., Goncalves, A., Huber, W., Ukkonen, E. and Brazma, A. (2010). A global map of human gene expression. *Nature Biotechnology*, 28(4): 322-324. doi:10.1038/nbt0410-322.

Luo, X., Huang, L., Han, L., Luo, Z., Hu, F., Tieu, R. and Gan, L. (2014). Systematic Prioritization and Integrative Analysis of Copy Number Variations in Schizophrenia Reveal Key Schizophrenia Susceptibility Genes. *Schizophrenia Bulletin*, 40(6): 1285-1299. doi:10.1093/schbul/sbu045.

Lupski, J.R. and Stankiewicz, P. (2005). Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genetics*, 1(6): e49.

Lyall, K., Pauls, D.L., Spiegelman, D., Ascherio, A and Santangelo, S.L. (2012). Pregnancy complications of obstetric suboptimality in association with autism spectrum disorders in children of the Nurses' Healthy Study II. *Autism Research*, 5(1): 21-30. doi:10.1002/aur.228.

Lyall, K., Schmidt, R.J. and Hertz-Picciotto, I. (2014). Maternal lifestyle and environmental risk factors for autism spectrum disorders. *International Journal of Epidemiology*, 43(2): 443-464. doi:10.1093/ije/dyt282.

Lyon, G.J., Jiang, T., Van Wijk, R., Wang, W., Bodily, P.M., Xing, J., Tian, L., Robison, R.J., Clement, M., Lin, Y. *et al.* (2011). Exome sequencing and unrelated findings in the context of complex disease research: ethical and clinical implications. *Discovery Medicine*, 12(62): 41-55.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J. *et al.* (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(Database issue): D896–D901. doi:10.1093/nar/gkw1133.

Macdonald, J.R., Ziman, R., Yuen, R.K., Feuk, L. and Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the

human genome. *Nucleic Acids Research*, 42(Database issue): D986-D992. doi:10.1093/nar/gkt958.

Mackay, T.F. and Moore, J.H. (2014). Why epistasis is important for tackling complex human disease genetics. *Genome Medicine*, 6(6):124. doi:10.1186/gm561.

Ma, D., Salyakina, D., Jaworski, J.M., Konidari, I., Whitehead, P.L., Andersen, A.N., Hoffman, J.D., Slifer, S.H., Hedges, D.J., Cukier, H.N. *et al.* (2009). A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. *Annals of Human Genetics*, 73(Pt 3): 263-273. doi:10.1111/j.1469-1809.2009.00523.x.

Maeda, H., Nagata, S., Wolfgang, C.D., Bratthauer, G.L., Bera, T.K. and Pastan, I. (2004) The T cell receptor gamma chain alternate reading frame protein (TARP), a prostate-specific protein localized in mitochondria. *Journal of Biological Chemistry*, 279(23): 24561-24568. doi:10.1074/jbc.M402492200.

Mahfouz, A., Ziats, M.N., Rennert, O.M., Lelieveldt, B.P. and Reinders, M.J. (2015). Shared Pathways Among Autism Candidate Genes Determined by Co-expression Network Analysis of the Developing Human Brain Transcriptome. *Journal of Molecular Neuroscience*, 57(4): 580-594. doi:10.1007/s12031-015-0641-3.

Mühleisen, T.W., Leber, M., Schulze, T.G., Strohmaier, J., Degenhardt, F., Treutlein, J., Mattheisen, M., Forstner, A.J., Schumacher, J., Breuer, R. *et al.* (2014). Genome-wide association study reveals two new risk loci for bipolar disorder. *Nature Communications*, 5:3339. doi:10.1038/ncomms4339.

Makino, T. and McLysaght, A. (2008). Interacting gene clusters and the evolution of the vertebrate immune system. *Molecular Biology and Evolution*, 25(9): 1855-1862. doi:10.1093/molbev/msn137.

Mak, I.W., Evaniew, N. and Ghert, M. (2014). Lost in translation: animal models and clinical trials in cancer treatment. *American Journal of Translational Research*, 6(2): 114-118.

Malhotra, D., McCarthy, S., Michaelson, J.J., Vacic, V., Burdick, K.E., Yoon, S., Cichon, S., Corvin, A., Gary, S., Gershon, E.S. *et al.* (2011). High frequencies of De Novo CNVs in Bipolar Disorder and Schizophrenia. *Neuron,* 7296): 951-963. doi:10.1016/j.neuron.2011.11.007.

Mannion, A. and Leader, G. (2013). Comorbidity in autism spectrum disorder: A literature review. *Research in Autism Spectrum Disorders*, 7(12): 1595-1616. doi:10.1016/j.rasd.2013.09.006.

Mannion, A., Brahm, M. and Leader, G. (2014). Comorbid Psychopathology in Autism Spectrum Disorder. Review *Journal of Autism and Developmental Disorders*, 1:124. doi:10.1007/s40489-014-0012-y.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A. *et al.* (2009). Finding the missing heritability of complex diseases. *Nature,* 461(7265): 747-53. doi:10.1038/nature08494.

Marshall, C.R., Noor, A., Vincent, J.B., Lionel, A.C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y. *et al.* (2008). Structural variation of chromosomes in autism spectrum disorder. *American Journal of Human Genetics*, 82(2): 477-488. doi:10.1016/j.ajhg.2007.12.009.

Marth, G.T., Yu, F., Indap, A.R., Garimella, K., Gravel, S., Leong, W.F., Tyler-Smith, C., Bainbridge, M., Blackwell, T., Zheng-Bradley, X. *et al.* (2011). The functional

spectrum of low-frequency coding variation. *Genome Biology*, 12(9): R84. doi:10.1186/gb-2011-12-9-r84.

Mas, S., Bernardo, M., Parellada, E., Garcia-Rizo, C., Gasso, P., Alvarez, S. and Lafuente, A. (2009). ARVCF single marker and haplotypic association with schizophrenia. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 33(6): 1064-1069. doi:10.1016/j.pnpbp.2009.06.001.

Matson, J.L. and Cervantes, P.E. (2013). Co-morbidity among persons with intellectual disabilities. *Research in Autism Spectrum Disorders*, 7(11): 1318-1322. doi:10.1016/j.rasd.2013.07.018.

Mattheisen, M., Samuels, J.F., Wang, Y., Greenberg, B.D., Fyer, A.J., McCracken, J.T., Geller, D.A., Murphy, D.L., Knowles, J.A., Grados, M.A. *et al.* (2015). Genome-wide association study in obsessive-compulsive disorder: results from the OCGAS. *Molecular Psychiatry*, 20(3): 337-344. doi:10.1038/mp.2014.43.

Maulik, P.K., Mascarenhas, M.N., Mathers, C.D., Dua, T. and Saxena, S. (2011). Prevalence of intellectual disability: a meta-analysis of population-based studies. Research in Developmental Disabilities, 32(2): 419-436. doi:10.1016/j.ridd.2010.12.018.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P. and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5): 356-369. doi:10.1038/nrg2344.

McCarthy, S., Das, S., Kretzscmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K. *et al.* (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, 48(10): 1279-1283. doi:10.1038/ng.3643.

McCarthy, S.E., Gillis, J., Kramer, M., Lihm, J., Yoon, S., Berstein, Y., Mistry, M., Pavlidis, P., Solomon, R., Ghiban, E. *et al.* (2014). De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Molecular Psychiatry*, 19: 652–658; doi:10.1038/mp.2014.29.

McCarthy, S.E., Makarov, V., Kirov, G., Addington, A.M., McClellan, J., Yoon, S., Perkins, D.O., Dickel, D.E., Kusenda, M., Krastoshevsky, O. *et al.* (2009). Microduplications of 16p11.2 are associated with schizophrenia. *Nature Genetics*, 41(11): 1223-1227. doi:10.1038/ng.474.

McClellan, J. and King, M.C. (2010). Genetic heterogeneity in human disease. *Cell,* 141(2): 210-217. doi:10.1016/j.cell.2010.03.032.

McDowall, M.D., Scott, M.S. and Barton, G.J. (2009). PIPs: human protein-protein interaction prediction database. *Nucleic Acids Research*, 37(Database issue): D651-D656. doi:10.1093/nar/gkn870.

McGrath, L.M., Yu, D., Marshall, C., Davis, L.K., Thiruvahindrapuram, B., Li, B., Cappi, C., Gerber, G., Wolf, A., Schroeder, F.A. *et al.* (2014). Copy number variation in obsessive-compulsive disorder and tourette syndrome: a cross-disorder study. *Journal of the American Academy of Child and Adolescent Psychiatry*, 53(8): 910-919. doi:10.1016/j.jaac.2014.04.022.

McKenzie, K., Milton, M., Smith, G. and Ouellette-KuntzEmail, H. (2016). Systematic Review of the Prevalence and Incidence of Intellectual Disabilities: Current Trends and Issues. *Current Developmental Disorders Reports*, 3(2): 104-115.

McKusick, V.A. (1998). Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders (12th ed.) Baltimore: Johns Hopkins University Press.

McPherson, R. and Tybjaerg-Hansen, A. (2016). Genetics of Coronary Artery Disease. *Circulation Research*, 118(4): 564-578. doi:10.1161/CIRCRESAHA.115.306566.

Melom, J.E. and Littleton, J.T. (2011). Synapse development in health and disease. Current *Opinion in Genetics and Development*, 21(3): 2562-61. doi:10.1016/j.gde.2011.01.002.

Merikangas, A.K., Segurado, R., Heron, E.A., Anney, R.J., Paterson, A.D., Cook, E.H., Pinto, D., Scherer, S.W., Szatmari, P., Gill, M. *et al.* (2015). The phenotypic manifestations of rare genic CNVs in autism spectrum disorder. *Molecular Psychiatry*, 20(11): 1366-1372. doi:10.1038/mp.2014.150.

Merner, N.A., Dion, P.A. and Rouleau, G.A. (2015). The mutational spectrum of neurodevelopmental disorders. In: Mitchell, K.J. The genetics of neurodevelopmental disorders. US: Wiley-Blackwell. 49-68.

Mikhaleva, A., Kannan, M., Wagner, C. and Yalcin, B. (2016). Histomorphological phenotyping of the Adult Mouse Brain. *Current Protocols in Mouse Biology*, 6(3): 307-332. doi:10.1002/cpmo.12.

Miller, B.J., Graham, K.L., Chelsea, B.M., Culpepper, N.H.m Waller, J.L. and Buckley, P.F. (2013). A Prevalence Study of Urinary Tract Infections in Acute Relapse of Schizophrenia. *Journal of Clinical Psychiatry*, 74(3):271–277. doi:10.4088/JCP.12m08050.

Miller, D.T., Adam, M.P., Aradhya, S., Biesecker, L.G., Brothman, A.R., Carter, N.P., Church, D.M., Crolla, J.A., Eichler, E.E., Epstein, C.J. *et al.* (2010). Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *American Journal of Human Genetics*, 86(5): 749-764. doi:10.1016/j.ajhg.2010.04.006.

Miller, J.A., Horvath, S. and Geschwind, D.H. (2010). Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *PNAS*, 107(28): 12698-12703. doi:10.1073/pnas.0914257107.

Milligan, G.W. and Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a dataset. *Psychometrika*, 50(2): 159-179. doi:10.1007/BF02294245.

Miranda, D.M., Wigg, K., Kabia, E.M., Feng, Y., Sandor, P. and Barr, C.L. (2009). Association of SLITRK1 to Gilles de la Tourette Syndrome. American *Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 150B(4): 483-486. doi:1002.ajmg.b.30840.

Missler, M., Südhof, T.C. and Biederer, T. Synaptic cell adhesion. *Cold Spring Harbour Perspectives in Biology*, 4(4): a005694. doi:10.1101/cshperspect.a005694.

Mitra, I., Lavillaureix, A., Yeh, E., Traglia, M., Tsang, K., Bearden, C.E., Rauen, K.A. and Weiss, L. (2017). Reverse Pathway Genetic Approach Idenfies Epistasis in Autism Spectrum Disorders. *PLoS Genetics*, 13(1): e1006516. doi:10.1371/journal.pgen.1006516.

Mitra, I., Tsang, K., Ladd-Acosta, C., Croen, L.A., Aldinger, K.A., Hendren, R.L., Traglia, M., Lavillaureix, A., Zaitlen, Oldham, M.C. *et al.* (2016). Pleiotropic mechanisms indicated for sex differences in autism. *PLoS Genetics*, 12(11): e1006425. doi:10.1371/journal.pgen.1006425.

Mlynarski, E.E., Sheridan, M.B., Xie, M., Guo, T., Racedo, S.E., McDonald-McGinn, D.M., Gai, X., Chow, E.W., Vistman, J., Swillen, A., Devriendt, K., Breekpot, J., Digilio, M.C., Marino, B., Dallapiccola, B. *et al.* (2015). Copy-number variation of the glucose transported gene SLC2A3 and congenital heart defects in the 22q11.2 deletion syndrome. *American Journal of Human Genetics*, 96(5): 753-764. doi:10.1016/j.ajhg.2015.03.007.

Mol Debes, N.M. (2013). Co-morbid disorders in Tourette syndrome. *Behavioural Neurology,* 27(1): 7-14. doi: 10.3233/BEN-120275.

Mol Debes, N.M., Hjalgrim, H. and Skov, L. (2008). Validation of the presence of comorbidities in a Danish clinical cohort of children with Tourette syndrome. *Journal of Child Neurology,* 2(9): 1017-1027. doi:10.1177/0883073808316370.

Monaco, G., van Dam, S., Ribeiro, J.L.C.N., Larbi, A. and de Magalhaes, J.P. (2015). A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels. *BMC Evolutionary Biology*, 15:259. doi:10.1186/s12862-015-0534-7.

Moog, U., Bierhals, T., Brand, K., Bautsch, J., Biskup, S., Brune, T., Denecke, J., de Die-Smulders, C.E., Evers, C., Hempel, M. *et al.* (2015). Phenotypic and molecular insights into CASK-related disorders in males. *Orphanet Journal of Rare Diseases*, 10:44. doi:10.1186/s13023-015-0256-3.

Moon, H.M. and Wynshaw-Boris, A. (2013). Cytoskeleton in action: lissencephaly, a neuronal migration disorder. *Wiley Interdisciplinary Reviews: Developmental Biology*, 2(2): 229-245. doi:10.1002/wdev.67.

Moore, J.M., Oliver, P.L., Finelli, M.J., Lee, S., Lickiss, T., Molnar, Z. and Kay, D.E. (2014). Laf4/Aff3, a gene involved in intellectual disability is required for cellular migration in the mouse cerebral cortex. *PLoS One*, 9: e105933. doi:10.1371/journal.pone.0105933.

Moreno-De-Luca, A., Helmers, S., Mao, H., Burns, T.G., Melton, A.M., Schmidt, K.R., Fernhoff, P.M., Ledbetter, D.H. and Martin, C.L. (2011). Adaptor protein complex-4 (AP-4) deficiency causes a novel autosomal recessive cerebral palsy syndrome with microcephaly and intellectual disability. *Journal of Medical Genetics*, 48(2): 141-144. doi:10.1136/hmg.2010.082263.

Moreno-De-Luca, D., Sanders, S.J., Willsey, A.J., Mulle, J.G., Lowe, J.K., Geschwind, D.H., State, M.W., Martin, C.L. and Ledbetter, D.H. (2013). Using large clinical data sets to infer pathogenicity for rare copy number variants in autism cohorts. *Molecular Psychiatry*, 18(10): 1090-1095. doi:10.1038/mp.2012.138.

Moreno-de-Luca, A., Evans, D.W., Boomer, K.B., Hanson, E., Bernier, R., Goin-Kochel, R.P., Myers, S.M., Challman, T.D., Moreno-de-Luca, D., Slane, M.M., Hare, A.E., Chung, W.K., Spiro, J.E., Faucett, W.A., Martin, C.L. and Ledbetter, D.H. (2015). The role of parental cognitive, behavioural, and motor profiles in clinical variability in individuals with chromosome 16p11.2 deletions. *JAMA Psychiatry*, 72(2): 119-126. doi:10.1001/jamapsychiatry.2014.2147.

Mortensen, P.B., Pedersen, C.B., Westergaard, T., Wohlfahrt, J., Ewald, H., Mors, O., Andersen, P.K. and Melbye, M. (1999). Effects of family history and place and season of birth on the risk of schizophrenia. *New England Journal of Medicine*, 340(8): 603-608. doi:10.1056/NEJM199902253400803.

Nag, A., Bochukova, E.G., Kremeyer, B., Campbell, D.D., Muller, H., Valencia-Duarte, A.V., Cardona, J., Rivas, I.C., Mesa, S.C., Cuartas, M. *et al.* (2013). CNV

analysis in Tourette syndrome implicates large genomic rearrangements in COL8A1 and NRXN1. *PLoS One*, 8(3): e59061. doi:10.1371/journal.pone.0059061.

Najmabadi, H., Hu, H., Garshasbi, M., Zemojtel, T., Abedini, S.S., Chen, W., Hosseini, M., Behjati, F., Haas, S., Jamali, P. *et al.* (2011). Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature,* 478(7367): 57-63. doi:10.1038/nature10423.

Nakayama, J. (2009). Progress in searching for the febrile seizure susceptibility genes. *Brain and Development*, 31(5): 359-365. doi:10.1016/j.braindev.2008.11.014.

Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.F., Stevens, C., Wang, L.S., Makarov, V. *et al.* (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature,* 485(7397): 242-245. doi:10.1038/nature11011.

Neale, B.M., Medland, S.E., Ripke, S., Asherson, P., Franke, B., Lesch, K.P., Faraone, S.V., Nguyen, T.T., Schäfer, H., Holmans, P. *et al.* (2010). Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, 49(9): 884-897. doi:10.1016/j.jaac.2010.06.008.

Nei, M. (1967). Modification of linkage intensity by natural selection. *Genetics*, 57(3): 625-641.

Nevado, J., Mergener, R., Palomares-Bralo, M., Souza, K.R., Vallespin, E., Mena, R., Martinez-Glex, V., Mori, M.A., Santos, F., Garcia-Minaur, S. *et al.* (2014). New microdeletion and microduplication syndromes: A comprehensive review. *Genetics and Molecular Biology*, 37(Suppl1): 210-219.

Newbury, D.F., Paracchini, S., Scerri, T.S., Winchester, L., Addis, L., Richardson, A.J., Walter, J., Stein, J.F., Talcott, J.B. and Monaco, A.P. (2011). Investigation of dyslexia and SLI variants in reading- and language-impaired subject. *Behavior Genetics*, 41(1): 90-104. doi:1007/s10519-010-9424-3.

Newschaffer, C.J., Croen, L.A., Daniels, J., Giarelli, E., Grether, J.K., Levy, S.E., Mandell, D.S., Miller, L.A., Pinto-Martin, J., Reaven, J. *et al.* (2007). The epidemiology of autism spectrum disorders. Annual Review of Public Health, 28: 235-58. doi:10.1146/annurev.publhealth.28.021406.144007.

Ng, P.C. and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13): 3812-3814.

Nicholson, T.R.J., Ferdinando, S., Krishnaiah, R.B., Anhoury, S., Lennox, B.R., Mataix-Cols, D., Cleare, A., Veale, D.M., Drummond, M., Fineberg, N.A. *et al.* (2012). Prevalence of anti-basal ganglia antibodies in adult obsessive-compulsive disorder: cross-sectional study. *The British Journal of Psychiartry*, 200(5): 381-386. doi:10.1192/bjp.bp.111.092007.

Nikolaus, S., Antke, C. and Müller, H.W. (2009). In vivo imaging of synaptic function in the central nervous system: II. Mental and affective disorders. *Behavioural Brain Research*, 204(1): 32-66. doi: 10.1016/j.bbr.2009.06.009.

Nikpay, M., Goel, A., Won, H.H., Hall, L.M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C.P., Hopewell, J.C. *et al.* (2015). A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, 47(10): 1121-1130. doi:10.1038/ng.3396.

Noh, H.J., Ponting, C.P., Boulding, H.C., Meader, S., Betancur, C., Buxbaum, J.D., Pinto, D., Marshall, C.R., Lionel, A.C., Scherer, S.W. and Webber, C. (2013).

Network topologies and convergent aetiologies arising from deletions and duplications observed in individuals with autism. *PLoS Genetics*, 9(6): e1003523. doi:10.1371/journal.pgen.1003523.

Norbury, C.F., Gooch, D., Baird, G., Charman, T., Simonoff, E. and Pickles, A. (2016). Younger children experience lower levels of language competence and academic progress in the first year of school: evidence from a population study. *Journal of Child Psychology and Psychiatry*, 57(1): 65-73. doi:10.1111/jcpp.12431.

Nudel, R., Simpson, N.H., Baird, G., O'Hare, A., Conti-Ramsden, G., Bolton, P.F., Hennessy, E.R., SLI Consortium, Ring, S.M., Davey Smith, G. *et al.* (2014). Genome-wide association analyses of child genotype effects and parent-of-origin effects in specific language impairment. *Genes, Brain and Behaviour*, 13(4): 418-429. doi:10.1111/gbb.12127.

O'Connor, R.M., Stone, E.F., Wayne, C.R., Marcinkevicius, E.V., Ulgherait, M., Delventhal, R., Pantalia, M.M., Hill, V.M., Zhou, C.G., McAllister, S. *et al.* (2017). A Drosophila model of Fragile X syndrome exhibits defects in phagocytosis by innate immune cells. *Journal of Cellular Biology*, doi:10.1083/jcb.201607093.

O'Donovan, M.C., Craddock, N., Norton, N., Williams, H., Peirce, T., Moskvina, V., Nikolov, I., Hamshere, M., Carroll, L., Georgieva, L. *et al.* (2008). Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nature Genetics*, 40(9): 1053-1055. doi:10.1038/ng.201.

O'Roak, B.J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J.J., Girirajan, S., Karakoc, E., Mackenzie, A.P., Ng, S.B., Baker, C. *et al.* (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature Genetics*, 43(6): 585-589. doi:10.1038/ng.835.

O'Roak, B.J., Morgan, T.M., Fishman, D.O., Saus, E., Alonso, P., Gratacòs, M., Estivill, X., Teltsh, O., Kohn, Y., Kidd, K.K. *et al.* (2010). Additional support for the association of SLITRK1 var321 and Tourette syndrome. *Molecular Psychiatry*, 15(5): 447-450. doi:10.1038/mp.2009.105.

O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D. *et al.* (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature,* 485(7397): 246-250. doi:10.1038/nature10989.

O'Rourke, J.A., Scharf, J.M., Yu, D., Pauls, D.L. (2009). The genetics of Tourette syndrome: a review. Journal of Psychosomatic Research, 67(6): 533-545. doi:10.1016/j.jpsychores.2009.06.006.

Okbay, A., Baselmans, B.M., De Neve, J.E., Turley, P., Nivard, M.G., Fontana, M.A., Meddens, S.F., Linnér, R.K., Rietveld, C.A., Derringer, J. *et al.* (2016). Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics*, 48(6): 624-33. doi:10.1038/ng.3552.

Oksenberg, N., Haliburton, G.D.E., Eckalbar, W.L., Oren, I., NIshizaki, S., Murphy, K., Pollard, K.S., Birnbaum, R.Y. and Ahituv, N. (2014). Genome-wide distribution of Auts2 binding localizes with active neurodevelopmental genes. *Translational Psychiatry*, 4: e431. doi:10.1038/tp.2014.78.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N. *et al.* (2014). The MIntAct project--IntAct as a common curation platform for 11 molecular interaction

databases. *Nucleic Acids Research*, 42(Database issue): D358-D363. doi:10.1093/nar/gkt1115.

Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., Bridge, A., Briganti, L., Brinkman, F.S., Cesareni, G. *et al.* (2012). Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nature Methods*, 9(4): 345-350. doi:10.1038/nmeth.1931.

Orozco, G., Barrett, J.C. and Zeggini, E. (2010). Synthetic associations in the context of genome-wide association scan signals. *Human Molecular Genetics*, 19(R2): R137-R144. doi:10.1093/hmg/ddq368.

Oti, M. and Brunner, H.G. (2007). The modular nature of genetic diseases. *Clinical Genetics*, 71(1): 1-11. doi:10.1111/j.1399-0004.2006.00708.x.

Ozomaro, U., Cai, G., Kajiwara, Y., Yoon, S., Makarov, V., Delorme, R., Betancur, C., Ruhrmann, S., Falkai, P., Grabe, H.J. *et al.* (2013). Characterization of SLITRK1 variation in obsessive-compulsive disorder. *PLoS One*, 8(8): e70376. doi:10.1371/journal.pone.0070376.

Ozonoff, S., Young, G.S., Carter, A., Messinger, D., Yirmiya, N., Zwaigernbaum, L., Bryson, S., Carver, L.J., Constantino, J.N., Dobkins, K. *et al.* (2011). Recurrence Risk for Autism Spectrum Disorders: A Baby Siblings Research Consortium Study. *Pediatrics*, 128(3): e488–e495. doi:10.1542/peds.2010-2825.

Pajusalu, S., Reimand, T. and Ounap, K. (2015). Novel homozygous mutation in KPTN gene causing a familial intellectual disability-macrocephaly syndrome. American *Journal of Medical Genetics*, 167A(8): 1913-1915. doi:10.1002/ajmg.a.37105.

Parikshak, N.N., Gandal, M.J. and Geschwind, D. (2015). Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nature Reviews Genetics*, 16(8): 441-458. doi:10.1038/nrg3934.

Parikshak, N.N., Luo, R., Zhang, A., Won, H., Lowe, J.K., Chandran, V., Horvath, S. and Geschwind, D.H. (2013). Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell,* 155(5): 1008-1021. doi:10.1016/j.cell.2013.10.031.

Paschou, P. (2013). The genetic basis of Gilles de la Tourette Syndrome. *Neuroscience and Biobehavioral Reviews*, 7(6):1026-39. doi:10.1016/j.neubiorev.2013.01.016.

Paschou, P., Dongmei, Y., Gerber, G., Evans, P., Tsetsos, F., Davis, L.K., Karagiannidis, I., Chaponis, J., Gamazon, E., Mueller-Vahl, K. *et al.* (2014). Genetic association signal near NTN4 in Tourette Syndrome. *Annals of Neurology,* 76(2): 310–315. doi:10.1002/ana.24215.

Patzke, C., Acuna, C., Giam, L.R., Wernig, M. and Südhof, T.C. (2016). Conditional deletion of L1CAM in human neurons impairs both axonal and dendritic arborization and action potential generation. *Journal of Experimental Medicine*, 213(4): 499-515. doi:10.1084/jem.20150951.

Pauls, D.L., Raymond, C.L., Stevenson, J.M. and Leckman, J.F. (1991). A family study of Gilles de la Tourette syndrome. *American Journal of Human Genetics*, 48(1): 154-163.

Peça, J., Feliciano, C., Ting, J.T., Wang, W., Wells, M.F., Venkatraman, T.N., Lascola, C.D., Fu, Z. and Feng, G. (2011). Shank3 mutant mice display autistic-like

behaviours and striatal dysfunction. *Nature,* 472(7344): 437-442. doi:10.1038/nature09965.

Pego-Reigosa, J.M. and Isenberg, D.A. (2009). Psychosis due to systemic lupus erythematosus: characteristics and long-term outcome of this rare manifestation of disease. *Rheumatology (Oxford)*, 47(10): 1498-1502. doi:10.1093/rheumatology/ken260/.

Percy, A., Gillberg, C., Hagberg, B. and Witt-Engerström, I. (1990). Rett syndrome and the autistic disorders. *Neurologic Clinics*, 8(3): 659-676.

Pervouchine, D.D., Djebali, S., Breschi, A., Davis, C.A., Barja, P.P., Dobin, A., Tanzer, A., Lagarde, J., Zaleski, C., See, L.H. *et al.* (2015). Enhanced transcriptome maps from multiple tissues reveal evolutionary constraint in gene expression. *Nature Communications*, 6: 5903. doi:10.1038/ncomms.

Peterson, B.S., Thomas, P., Kane, M.J., Scahill, L., Zhang, H., Bronen, R., King, R.A., Leckman, J.F. and Staib, L. (2003). Basal Ganglia volumes in patients with Gilles de la Tourette syndrome. *Archives of General Psychiatry*, 60(4): 415-24. doi:10.1001/archpsyc.60.4.415

Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. and Goldstein, D.B. (2013). Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genetics*, 9(8): e1003709. doi:10.1371/journal.pgen.1003709.

Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., Thiruvahindrapuram, B., Xu, X., Ziman, R., Wang, Z. *et al.* (2014). Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *American Journal of Human Genetics*, 94(5): 677-694. doi:10.1016/j.ajhg.2014.03.018.

Pinto, S., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S. *et al.* (2010). Functional Impact of global rare copy number variation in autism spectrum disorder. *Nature,* 466(7304): 368-372. doi:10.1038/nature09146.

Piton, A., Michaud, J.L., Peng, H., Aradhya, S., Gauthier, J., Mottron, L., Champagne, N., Lafrenière, R.G., Hamdan, F.F., S2D team *et al.* (2008). Mutations in the calcium-related gene IL1RAPL1 are associated with autism. *Human Molecular Genetics*, 17(24): 3965-3974. doi:10.1093/hmg/ddn300.

Platt, R.J., Zhou, Y., Slaymaker, I.M., Shetty, A.S., Weisbach, N.R., Kim, J.A., Sharma, J., Desai, M., Sood, S., Kempton, H.R. *et al.* (2017). Chd8 Mutations Leads to Autistic-like Behaviors and Impaired Striatal Circuits. *Cell Reports*, 19(2): 335-350. Doi:10.1016/j.cellrep.2017.03.052.

Plummer, J.T., Gordon, A.J. and Levitt, P. (2016). The Genetic Intersection of neurodevelopmental disorders and shared medical comorbidities – Relations that translate from bench to bedside. *Frontiers in Psychiatry*, 7:142. doi:10.3389/fpsyt.2016.00142.

Poot, M. (2015). Connecting the CNTNAP2 Networks with Neurodevelopmental Disorders. *Molecular Syndromology*, 6(1): 7-22. doi:10.1159/000371594.

Power, R.A., Kyaga, S., Uher, R., MacCabe, J.H., Långström, N., Landen, M., McGuffin, P., Lewis, C.M., Lichtenstein, P. and Svensson, A.C. (2013). Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *JAMA Psychiatry*, 70(1): 22-30. doi:10.1001/jamapsychiatry.2013.268.

Prasad, A., Kumar, S.S., Dessimoz, C., Bleuler, S., Laule, O., Hruz, T., Gruissem, W. and Zimmermann, P. (2013). Global regulatory architexture of human, mose and rate tissue transcriptomes. *BMC Genomics*, 14: 716. doi:10.1186/1471-2164-14-716.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8): 904-909. doi:10.1038/ng1847.

Price, R.A., Kidd, K.K., Cohen, D.J., Pauls, D.L. and Leckman, J.F. (1985). A twin study of Tourette syndrome. *JAMA Psychiatry*, 42(8): 815-820.

Purcell, A.E., Jeon, O.H., Zimmerman, A.W., Blue, M.E. and Pevsner, J. (2001). Postmortem brain abnormalities of the glutamate neurotransmitter system in autism. *Neurology,* 57(9): 1618-1628.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Dark, M.J. and Sham, P.C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics*, 81(3): 559–575. doi:10.1086/519795.

Purper-Ouakil, D., Ramoz, N., Lespagnol-Bestel, A.-M., Gorwood, P. and Simonneau, M. (2011). Neurobiology of Attention Deficit/Hyperactivity Disorder. *Pediatric Research*, 69:69R–76R. doi:10.1203/PDR.0b013e318212b40f.

Qiao, Y., Mercier, E., Dastan, J., Hurlburt, J., McGillivray, B., Chudley, A.E., Farrell, S., Bernier, F.P., Lewis, M.S., Pavlidis, P. and Rajcan-Separovic, E. (2014). Copy number variants (CNVs) analysis in a deeply phenotyped cohort of individuals with intellectual disability (ID). BMC Medical Genetics, 15:82. doi:10.1186/1471-2350-15-82.

Ramos-Quiroga, J.-A., Sanchez-Mora, C., Casas, M., Garcia-Martinez, I., Bosch, R., Nogueira, M., Corrales, M., Palomar, G., Vidal, R., Coll-Tane, M. *et al.* (2014). Genome-wide copy number variation analysis in adult attention-deficit and hyperactivity disorder. *Journal of Psychiatric Research*, 49: 60-67. doi:10.1016/j.jpsychires.2013.10.022.

Ramtekkar, U.P., Reiersen, A.M., Todorov, A.A. and Todd, R.D. (2010). Sex and age differences in attention-deficit/hyperactivity disorder symptoms and diagnoses: implications for DSM-V and ICD-11. *Journal of the American Academy of Child and Adolescent Psychiatry*, 49(3): 217-228.

Rashid, A.J., Cole, C.J. and Josselyn, S.A. (2014). Emerging roles for MEF2 transcription factors in memory. *Genes, Brain and Behaviour*, 13(1): 118-125. doi:10.1111/gbb.12058.

Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J., Di Donato, N. *et al.* (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *The Lancet*, 380(9854): 1674-1682. doi:10.1016/S0140-6736(12)61480-9.

Raychaudhuri, S., Korn, J.M., McCarroll, S.A., International Schizophrenia Consortium, Altshuler, D., Sklar, P., Purcell, S. and Daly M.J. (2010). Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. *PLoS Genetics*, 16(9): e1001097. doi:10.1371/journal.pgen.1001097.

Reader, R.H., Covill, L.E., Nudel, R. and Newbury, D.F. (2014). Genome-Wide Studies of Specific Language Impairment. *Current Behavioural Neuroscience Reports*, 1(4): 242–250. doi:10.1007/s40473-014-0024-z.

Reiner, O., Karzbrun, E., Kshirsaga, A. and Kaibuchi, K. (2016). Regulation of neuronal migration, an emerging topic in autism spectrum disorders. *Journal of Neurochemistry*, 136(3): 440-456. doi:10.1111/knc.13403.

Riazuddin, S., Hussain, M., Razzaq, A., Iqbal, Z., Shahzad, M., Polla, D.L., Song, Y., van Beusekom, E., Khan, A.A., Tomas-Roca, L. *et al.* (2016). Exome sequencing of Pakistani consanguineous families identifies 30 novel candidate genes for recessive intellectual disability. *Molecular Psychiatry*. doi:10.1038/mp.2016.109.

Ripke, S., Neale, B.M., Corvin, A., Walters, J.T.R., Farh, K.-H., Holmans, P.A., Lee, P., Bulik-Sullivan, B., Collier, D.A., Huang, H. *et al.* (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature,* 511(7510): 421-427. doi:10.1038/nature13595.

Ripke, S., Sanders, A.S., Kendler, K.S., Levinson, D.F., Sklar, P., Holmans, P.A., Lin, D.-Y., Duan, J., Ophoff, R.A., Andreassen, O.A. *et al.* (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics*, 43(10): 969-976. doi:10.1038/ng.940.

Risch, N. (1990). Linkage strategies for genetically complex traits. I. Multilocus models. *American Journal of Human Genetics*, 46(2): 222-228.

Risch, N. and Merikangas, K.R. (1993). Linkage studies of psychiatric disorders. *European Archives of Psychiatry and Clinical Neuroscience*, 243(3-4):143-149.

Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N. *et al.* (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature Genetics*, 43(11): 1066-1073. doi:10.1038/ng.952.

Robertson, M.M. (2000). Tourette syndrome, associated conditions and the complexities of treatment. *Brain,* 123(3): 425-462.

Robertson, M.M., Eapen, V. and Cavanna, A.E. (2009). The international prevalence, epidemiology, and clinical phenomenology of Tourette syndrome: A cross-cultural perspective. *Journal of Psychosomatic Research*, 67(6): 475-483. DOI:10.1016/j.jpsychores.2009.07.010.

Robinson, E.B., Samocha, K.E., Kosmicki, J.A., McGrath, L., Neale, B.M., Perlis, R.H. and Daly, M.J. (2014). Autism spectrum disorder severity reflects the average contribution of de novo and familial influences. *PNAS*, 111(42): 15161-15165. doi:10.1073/pnas.1409204111.

Robinson, E.B., St Pourcain, B., Anttila, V., Kosmicki, J.A., Bulik-Sullivan, B., Grove, J., Maller, J., Samocha, K.E., Sanders, S.J., Ripke, S. *et al.* (2016). Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nature Genetics*, 48(5): 552-555. doi:10.1038/ng.3529.

Robinson, P.N. and Webber, C. (2014). Phenotype ontologies and cross-species analysis for translational research. *PLoS Genetics*, 10(4): e1004268. doi:10.1371.journal.pgen.1004268.

Robinson, P.N., Kohler, S., Bauer, S., Seelow, D., Horn, D. and Mundlos, S. (2008). The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *American Journal of Human Genetics*, 83(5): 610-615. doi:10.1016/j.ajhg.2008.09.017.

Roelfsema, J.H., White, S.J., Ariyürek, Y., Bartholdi, D., Niedrist, D., Papadia, F., Bacino, C.A., den Dunnen, J.T., van Ommen, G.J., Breuning, M.H. *et al.* (2005). Genetic heterogeneity in Rubinstein-Taybi syndrome: mutations in both the CBP and EP300 genes cause disease. *American Journal of Human Genetics*, 76(4): 572-580. doi:10.1086/429130.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R. *et al.* (2014). A proteome-scale map of the human interactome network. *Cell,* 159(5): 1212-1226. doi:10.1016/j.cell.2014.10.050.

Rommelse, N.N., Arias-Vásquez, A., Altink, M.E., Buschgens, C.J., Fliers, E., Asherson, P., Faraone, S.V., Buitelaar, J.K., Sergeant, J.A., Oosterlaan, J. and Franke, B. (2008). Neuropsychological endophenotype approach to genome-wide linkage analysis identifies susceptibility loci for ADHD on 2q21.1 and 13q12.11. *American Journal of Human Genetics*, 83(1): 99-105. doi:10.1016/j.ajhg.2008.06.006.

Ronan, J.L., Wu, W. and Crabtree, G.R. (2013). From neural development to cognition: unexpected roles for chromatin. *Nature Reviews Genetics*, 14(5): 347-359. doi:10.1038/nrg3413.

Ronemus, M., Iossifov, I., Levy, D and Wigler, M. (2014). The role of de novo mutations in the genetics of autism spectrum disorders. *Nature Reviews Genetics*, 15(2): 133-141. doi:10.1038/nrg3585.

Rosenberg, R.E., Law, J.K., Yenokyan, G., McGready, J., Kaufmann, W.E. and Law, P.A. (2009). Characteristics and concordance of autism spectrum disorders among 277 twin pairs. *Archives of Pediatrics and Adolescent Medicine*, 163(10): 907-914. doi:10.1001/archpediatrics.2009.98.

Rosenfeld, J.A., Coe, B.P., Eichler, E.E., Cuckle, H. and Shaffer, L.G. (2013). Estimates of penetrance for recurrent pathogenic copy-number variations. *Genetics in Medicine*, 15(6): 478-481. doi:10.1038/gim.2012.164.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C. and Mewes, H.W. (2009). Corum: the comprehensive resource of mammalian protein complexes – 2009. *Nucleic Acids Research*, 38(Database issue): D497-D501. doi:10.1083/nar/gkp914.

Rump, P., Jazayeri, O., van Dijk-Bos, K.K., Johansson, L.F., van Essen, A.J., Verheij, J.B., Veenstra-Knol, H.E., Redeker, E.J., Mannens, M.M., Swertz, M.A. *et al.* (2016). Whole-exome sequencing is a powerful approach for establishing the etiological diagnosis in patients with intellectual disability and microcephaly. *BMC Medical Genomics*, 9:7. doi:10.1186/s12920-016-0167-8.

Rung, J., Parkinson, H. and Brazma, A. (2010). Large scale comparison of global gene expression patterns in human and mouse. *Genome Biology*, 11: R124. doi:10.1186/gb-2010-11-12-r124.

Saka, E. and Graybiel, A.M. (2003). Pathophysiology of Tourette's syndrome: striatal pathways revisited. *Brain and Development*, 25(S1): S15-S19. doi:10.1016/S0387-7604(03)90002-7.

Sakai, Y,, Shaw, C.A., Dawson, B.C., Dugas, D.V., Al-Mohtaseb, Z., Hill, D.E. and Zoghbi, H.Y. (2011). Protein interactome reveals converging molecular pathways among autism disorders. *Science Translational Medicine*, 3(86): 86ra49. doi:10.1126/scitranslmed.3002166.

Sala, C., Vicidomini, C., Bigi, I., Mossa, A. and Verpelli, C. (2015). Shank synaptic scaffold proteins: keys to understanding the pathogenesis of autism and other synaptic disorders. *Journal of Neurochemistry*, 135(5): 849-858. doi:10.1111/jnc.13232.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U. and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, 32(Database issue):D449-D451.

Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A. *et al.* (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*, 46(9): 944-950. doi:10.1038/ng.3050.

Sanders, A.R., Rusu, I., Duan, J., Vander Molen, J.E., Hou, C., Schwab, S.G., Wildenauer, D.B., Martinez, M. and Gejman, P.V. (2005). Haplotypic association spanning the 22q11.21 genes COMT and ARVCF with schizophrenia. *Molecular Psychiatry*, 10(4): 353-365. doi:10.1038/sj.mp.4001586.

Sanders, S.J., Ercan-Sencicek, A.G., Hus, V., Luo, R., Murtha, M.T., Moreno-De-Luca, D., Chu, S.H., Moreau, M.P., Gupta, A.R., Thomson, S.A. *et al.* (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron,* 70(5): 863-885. doi:10.1016/j.neuron.2011.05.002.

Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S. *et al.* (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron,* 87(6): 1215-1233. doi:10.1016/j.neuron.2015.09.016.

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parkshak, N.N., Stein, J.L. *et al.* (2012). De novo mutations revealed by whole-exome sequencing was strongly associated with autism. *Nature,* 485(7397): 237-241. doi:10.1038/nature10945.

Sando, R.3rd, Gounko, N., Pieraut, S., Liao, L., Yates, J.3rd and Maximov, A. (2012). HDAC4 governs a transcriptional program essential for synaptic plasticity and memory. *Cell,* 151(4): 821-834. doi:10.1016/j.cell.2012.09.037.

Saunders, E.J., Dadaev, T., Leongamornlert, D.A., Jugurnauth-Little, S., Tymrakiewicz, M., Wiklund, F., Al Olama, A.A., Benlloch, S., Neal, D.E., Hamdy, F.C. *et al.* (2014). Fine-mapping the HOXB region detects common variants tagging a rare coding allele: evidence for synthetic association in prostate cancer. *PLoS Genetics*, 10(2): e1004129. doi:10.1371/journal.pgen.1004129.

Schaaf, C.P., Sabo, A., Sakai, Y., Crosby, J., Muzny, D., Hawes, A., Lewis, L., Akbar, H., Varghese, R., Boerwinkle, E. *et al.* (2011). Oligogenic heterozygosity in individuals with high-functioning autism spectrum disorders. *Human Molecular Genetics*, 20(17): 3366-3375. doi:10.1093/hmg/ddr243.

Schaefer, M.H., Fontaine, J.F., Vinayagam, A., Porras, P., Wanker, E.E. and Andrade-Navarro, M.A. (2012). HIPPIE: Integrating protein interaction networks with experiment based quality scores. *PLoS One*, 7(2): e31826. doi:10.1371/journal.pone.0031826.

Scharf, J.M., Moorjani, P., Fagerness, J., Platko, J.V., Illmann, C., Galloway, B., Jenike, E., Stewart, S.E., Pauls, D.L. and the Tourette Syndrome International Consortium for Genetics. (2008). Lack of association between SLITRK1var321 and

Tourette syndrome in a large family-based sample. *Neurology,* 70(16 Pt 2): 1495-1496. doi:10.1212/01.wnl.0000296833.25484.bb.

Scharf, J.M., Yu, D., Mathews, C.A., Neale, B.M., Stewart, S.E., Fagerness, J.A., Evans, P., Gamazon, E., Edlund, C.K., Service, S.K. *et al.* (2013). Genome-wide association study of Tourette's syndrome. *Molecular Psychiatry*, 18(6): 721-728. doi:10.1038/mp.2012.69.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J. *et al.* (2007). Strong Association of De Novo Copy Number Mutations with Autism. *Science*, 316(5823): 445-449. doi:10.1126/science.1138659.

Seth, R.B., Sun, L., Ea CK and Chen, Z.J. (2005). Identification and characterization of MAVS, a mitochondrial antiviral signalling protein that activates NF-kappaB and IRF 3. *Cell*, 122(5): 669-682. doi:10.1016/j.cell.2005.08.012.

Shabalin, A.A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10): 1353-1358. doi:10.1093/bioinformatics/bts163.

Shalev, S.A. and Hall, J.G. (2003). Another adult with Meier-Gorlin syndrome--insights into the natural history. *Clinical Dysmorphology*, 12(3): 167-169. doi:10.1097/01.mcd.0000065052.36236.32.

Sharp, S.I., McQuillin, A. and Gurling, H.M. (2009). Genetics of attention-deficit hyperactivity disorder (ADHD). *Neuropharmacology*, 57(7-8): 590-600. doi:10.1016/j.neuropharm.2009.08.011.

Shen, J., Gilmore, E.C., Marshall, C.A., Haddadin, M., Reynolds, J.J., Eyaid, W., Bodell, A., Barry, B., Gleason, D., Allen, K. *et al.* (2010). Mutations in PNKP cause microcephaly, seizures and defects in DNA repair. *Nature Genetics*, 42(3): 245-249. doi:10.1038/ng.526.

Sheppard, D.M., Bradshaw, J.L., Purcell, R. and Pantelis, C. (1999). Tourette's and comorbid syndromes: obsessive compulsive and attention deficit hyperactivity disorder. A common etiology? *Clinical Psychology Review*, 19(5): 531-52.

Shevell, M., Ashwal, S., Donley, D., Flint, J., Gingold, M., Hirtz, D., Majnemer, A.m Noetzel, M. and Sheth, R.D. (2003). Practice parameter: Evaluation of the child with global developmental delay. *Neurology,* 60(3): 367-380. doi:http://dx.doi.org/10.1212/01.WNL.0000031431.81555.16.

Shim, K.S. and Lubec, G. (2002). Drebrin, a dendritic spine protei, is manifold decreased in brains of patients with Alzheimer's disease and Down syndrome. *Neuroscience Letters*, 324(3): 209-212.

Shishido, E., Aleksic, B. and Ozaki, N. (2013). Copy-number variation in the pathogenesis of autism spectrum disorder. *Psychiatry and Clinical Neurosciences*, 68(2): 85-95. doi:10.1111/pcn.12128.

Shohat, S., Ben-David, E. and Shifman, S. (2017). Varying intolerance of gene pathways to mutational classes explain genetic convergence across neuropsychiatric disorders. *Cell Reports*, 18(9): 2217-2227. doi:10.1016/j.cellrep.2017.02.007.

Shyamsundar, R., Kim, Y.H., Higgins, J.P., Montgomery, K., Jorden, M., Sethuraman, A., van de Rijn, M., Botstein, S., Brown, P.O. and Pollack, J.R. (2005). A DNA microarray survey of gene expression in normal human tissues. *Genome Biology*, 6(3): R22.

Silverman, J.L., Yang, M., Lord, C. and Crawley, J.N. (2010). Behavioural phenotyping assays for mouse models of autism. *Nature Reviews Neuroscience*, 11(7): 490-502. doi:10.1038/nrn2851.

Sim, K., Chan, W.Y., Woon, P.S., Low, H.Q., Lim, L., Yang, G.L., Lee, J., Chong, S.A., Sitoh, Y.Y., Chan, Y.H. *et al.* (2012). ARVCF genetic influences on neurocognitive and neuroanatomical intermediate phenotypes in Chinese patients with schizophrenia. *Journal of Clinical Psychiatry*, 73(3): 320-326. doi:10.4088/JCP.10m06491.

Simonoff, E., Pickles, A., Charman, T., Chandler, S., Loucas, T. and Baird, G. (2008). Psychiatric disorders in children with autism spectrum disorders: prevalence, comorbidity, and associated factors in a population-derived sample. *Journal of the American Academy of Child and Adolescent Psychiatry*, 47(8): 921-929. doi:10.1097/CHI.0b013e318179964f.

Singer, H.S., Morris, C., Gause, C., Pollard, M., Zimmerman, A.W. and Pletnikov, M. (2009). Prenatal exposure to antibodies from mothers of children with autism produces neurobehavioral alterations: a pregnant dam mouse model. *Journal of Neuroimmunology*, 211: 39–48.

Sirotkin, H., O'Donnell, H., DasGuta, R., Halford, S., St Jore, B., Puech, A., Parimoo, S., Morrow, B., Skoultchi, A., Weissman, S.M. *et al.* (1997). Identification of a new human catenin gene family member (ARVCF) from the region deleted in velo-cardio-facial syndrome. *Genomics*, 41(1): 75-83.

Skarnes, W.C., Rosen, B., West, A.P., Koutsourakis, M., Bushell, W., Iyer, V., Mujica, A.O., Thomas, M., Harrow, J., Cox, T. *et al.* (2011). A conditional knockout resource for the genome-wide study of mouse gene function. *Nature*, 474(7351): 337-342. doi:10.1038/nature10163.

Sklar, P., Ripke, S., Scott, L.J., Andreassen, O.A., Cichon, S., Craddock, N., Edenber, H.J., Nurnberger, J.I., Riestschel, M., Blackwood, D. *et al.* (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature Genetics*, 43(10): 977-983. doi:10.1038/ng.943.

SLI Consortium (SLIC) (2004). Highly significant linkage to the SLI1 locus in an expanded sample of individuals affected by specific language impairment. *American Journal of Human Genetics*, 74(6):1225-1238. doi:10.1086/421529.

Smith, C.L. and Eppig, J.T. (2009). The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 1(3): 390-399. doi:10.1002/wsbm.44.

Smith, D.J., Escott-Price, V., Davies, G., Bailey, M.E., Colodro-Conde, L., Ward, J., Vedernikov, A., Marioni, R., Cullen, B., Lyall, D. *et al.* (2016). Genome-wide analysis of over 106 000 individuals identifies 9 neuroticism-associated loci. *Molecular Psychiatry*, 21(6): 749-757. doi:10.1038/mp.2016.49.

Smith, D.J., Stevens, M.E., Sudanagunta, S.P., Bronson, R.T., Makhinson, M., Watabe, A.M., O'Dell, T.J., Fung, J., Weier, H.U., Cheng, J.F. and Rubin, E.M. (1997). Functional screening of 2 Mb of human chromosome 21q22.2 in transgenic mice implicates minibrain in learning defects associated with Down syndrome. *Nature Genetics*, 16(1):28-36.

Solovieff, N., Hartley, S.W., Baldwin, C.T., Perls, T.T., Steinberg, M.H. and Sebastiani, P. (2010). Clustering by genetic ancestry using genome-wide SNP data. *BMC Genetics*, 11: 108. doi:10.1186/1471-2156-11-108.

Soto, D., Altafaj, X., Sindreu, C. and Bayés, A. (2014). Glutamate receptor mutations in psychiatric and neurodevelopmental disorders. *Communicative and Integrative Biology*, 7(1): e27887. doi:10.4161/cib.27887.

Spillane, J., Kullmann, D.M. and Hanna, M.G. (2016). Genetic neurological channelopathies: molecular genetics and clinical phenotypes. *Journal of Neurology, Neurosurgery and Psychiatry*, 87(1): 37-48. doi:10.1136/jnnp-2015-311233.

Srour, M. and Shevell, M. (2014). Genetics and the investigation of developmental delay/intellectual disability. *Archives of Disease in Childhood*, 99:386-389. doi:10.1136/archdischild-2013-304063.

Stankiewicz, P. and Lupski, J.R. (2010). Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, 61:437-455. doi:10.1146/annurev-med-100708-204735.

Stark, C., Breitkreutz, B.-J., Chatr-aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X. *et al.* (2011). The BioGRID Interaction Database: 2011 update. *Nucleic Acids Research*, 39(Database issue): D698-D704. doi:10.1093/nar/gkq1116.

Stefansson, H., Meyer-Lindenberg, A., Steinberg, S., Magnusdottir, B., Morgen, K., Arnarsdottir, S., Bjornsdottir, G., Walters, G.B., Jonsdottir, G.A., Doyle, O.M. *et al.* (2014). CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature,* 505(7483): 361-366. doi:10.1038/nature12818.

Stefansson, H., Ophoff, R.A., Steinberg, S., Andreassen, O.A., Cichon, S., Rujescu, D., Werge, T., Pietilainen, O.P., Mors, O., Mortensen, P.B. *et al.* (2009). Common variants conferring risk of schizophrenia. *Nature,* 460(7256): 744-747. doi:10.1038/nature08186.

Stefansson, H., Rujescu, D., Cichon, S., Pietilainen, O.P., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J.E. *et al.* (2008). Large recurrent microdeletions associated with schizophrenia. *Nature,* 455(7210): 232-236. doi:10.1038/nature07229.

Stegle, O., Parts, L., Durbin, R. and Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Computational Biology*, 6(5): e1000770. doi:10.1371/journal.pcbi.1000770.

Steinberg, J. and Webber, C. (2013). The roles of FMRP-regulated genes in autism spectrum disorder: single- and multiple-hit genetic etiologies. *American Journal of Human Genetics*, 93(5): 825-839. doi:10.1016/j.ajhg.2013.09.013.

Steinberg, J., Honti, F. Meader, S. and Webber, C. (2015). Haploinsufficiency predictions without study bias. *Nucleic Acids Research*, 43(15): e101. doi:10.1093/nar/gkv474.

Steinhausen, H.C., Novik, T.S., Baldursson, G., Curatolo, P., Lorenzo, M.J., Rodrigues Pereira, R., Ralston, S.J., Rothenberger, A. and ADORE Study Group (2006). Co-existing psychiatric problems in ADHD in the ADORE cohort. *European Child and Adolescent Psychiatry*, 15 Suppl 1:I25-9.

Stergiakouli, E., Hamshere, M., Holmans, P., Langley, K., Zaharieva, I., deCODE Genetics, Psychiatric GWAS Consortium, Hawi, Z., Kent, L., Gill, M. *et al.* (2012). Investigating the contribution of common genetic variants to the risk and pathogenesis of ADHD. *American Journal of Psychiatry*, 169(2): 186-194. doi:10.1176/appi.ajp.2011.11040551.

Stern, E., Silbersweig, D.A., Chee, K.Y., Holmes, A., Robertson, M.M., Trimble, M., Frith, C.D., Frackowiak, R.S. and Dolan, R.J. (2000). A functional neuroanatomy of tics in Tourette syndrome. *Archives of General Psychiatry*, 57(8): 741-748.

Stessman, H.A.F., Bernier, R. and Eichler, E.E. (2014). A Genotype-First Approach to Definitng the Subtypes of a Complex Disease. *Cell,* 156: 872-877. doi:10.1016/j.cell.2014.02.002.

Stessman, H.A.F., Turner, T. and Eichler, E.E. (2016). Molecular subtyping and improved treatment of neurodevelopmental disease. *Genome Medicine*, 8: 22. doi:10.1186/s13073-016-0278-z.

Stessman, H.O.F., Xiong, B., Coe, B.P., Wang, T., Hoekzema, K., Fenckova, M., Kvarnung, M., Gerdts, J., Trinh, S., Cosemans, N. *et al.* (2017). Targeted resequencing identifies 91 neurodevelopmental-dosrderr risk genes with autism and developmental-disability biases. *Nature Genetics*, 49(4): 515-526. doi:10.1038/ng.3792.

Stewart, S.E., Yu, D., Scharf, J.M., Neale, B.M., Fagerness, J.A., Mathews, C.A., Arnold, P.D., Evans, P.D., Gamazon, E.R., Davis, L.K. *et al.* (2013). Genome-wide association study of obsessive-compulsive disorder. *Molecular Psychiatry*, 18(7): 788-798. doi:10.1038/mp.2012.85.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G. *et al.* (2004). A gene atlase of the mouse and human protein-encoding transcriptomes. *PNAS*, 101(16): 6062-6067.

Subramanian, S. (2012). Quantifying harmful mutations in human populations. *European Journal of Human Genetics*, 20(12): 1320-1322. doi:10.1038/ejhg.2012.68.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M. *et al.* (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3): e1001779. doi:10.1371/journal.pmed.1001779.

Sun, J., Zhu, G., Liu, Y., Standley, S., Ji, A., Tunuguntla, R., Wang, Y., Claus, C., Luo, Y., Baudry, M. and Bi, X. (2015). UBE3A Regulates Synaptic Plasticity and Learning and Memory by Controlling SK2 Channel Endocytosis. *Cell Reports*, 12(3): 449-461. doi:10.1016/j.celrep.2015.06.023.

Sun, X., Namkung, J., Zhu, X. and Elston, R.C. (2011). Capability of common SNPs to tag rare variants. *BMC Proceedings*, 5 Suppl 9:S88. doi:10.1186/1753-6561-5-S9-S88.

Sundaram, S.K., Huq, A.M. and Chugani, H.T. (2010). Tourette syndrome is associated with recurrent exonic copy number variants. *Neurology,* 74(20): 1583-1590. doi:10.1212/WNL.0b013e3181e0f147.

Sunkin, S.M., Ng, L., Lau, C., Dolbeare, T., Gilbert, T.L., Thompson, C.L., Hawrylycz, M. and Dang, C. (2013). Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Research*, 41(Database issue): D996-D1008. doi:10.1093/nar/gks1042.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P. *et al.* (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(Database issue): D447-D452. doi:10.1093/nar/gku1003.

Tao, Y., Gao, H., Ackerman, B., Guo, W., Saffen, D and Shugart, Y.Y. (2016). Evidence for contribution of common genetic variants within chromosome 8p21.2-

8p21.1 to restricted and repetitive behaviors in autism spectrum disorders. *BMC Genomics*, 17:163. doi:10.1186/s12864-016-2475-7.

Tarpey, P., Parnau, J., Blow, M., Woffendin, H., Bignell, G., Cox, C., Cox, J., Davies, H., Edkins, S., Holden, S. *et al.* (2004). Mutations in the DLG3 gene cause nonsyndromic X-linked mental retardation. *American Journal of Human Genetics*, 75(2): 318-324. doi:10.1086/422703.

Tarpey, P.S., Smith, R., Pleasance, E., Whibley, A., Edkins, S., Hardy, C., O'Meara, S., Latimer, C., Dicks, E., Menzies, A. *et al.* (2009). A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nature Genetics*, 41(5): 535-543. doi:10.1038/ng.367.

Taylor, A., Steinberg, J. and Webber, C. (2015). Duplications in ADHD patients harbour neurobehavioural genes that are co-expressed with genes associated with hyperactivity in the mouse. *American Journal of Medical Genetics* Part *B: Neuropsychiatrics Genetics*, 168B(2): 97-107. doi:10.1002/ajmg.b.32285.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature,* 526(7571): 68-74. doi:10.1038/nature15393.

The Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature,* 409(6822):860-921. doi:10.1038/35057062.

The International HapMap Consortium (2003). The International HapMap Project. *Nature,* 426(6968): 789-796. doi:10.1038/nature02168.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature,* 437(7063): 1299-1320. doi:10.1038/nature04226.

Thomas, R., Sanders, S., Doust, J., Beller, E. and Glasziou, P. (2015). Prevalence of Attention-Deficit/Hyperactivity Disorder: A Systematic Review and Meta-analysis. *Pediatrics*, 135(4): e994-e100. doi:10.1542/peds.2014-348.

Tischfield, M.A. and Engle, E.C. (2010). Distinct alpha- and beta-tubulin isotypes are required for the positioning, differentiation and survival of neurons: new support for the 'multi-tubulin' hypothesis. *Bioscience Reports*, 30(5): 319-330. doi:10.1042/BSR20100025.

Toma, C., Torrico, B., Hervás, A., Valdés-Mas, R., Tristán-Noguero, A., Padillo, V., Maristany, M., Salgado, M., Arenas, C., Puente, X.S. *et al.* (2014). Exome sequencing in multiplex autism families suggests a major role for heterozygous truncating mutations. *Molecular Psychiatry*, 19(7): 784-790. doi:10.1038/mp.2013.106.

Tomblin, J.B. and Buckwalter, P.R. (1998). Heritability of poor language achievement among twins. *Journal of Speech, Language and Hearing Research*, 41(1):188-199.

Tomblin, J.B., Records, N.L., Buckwalter, P., Zhang, X., Smith, E. and O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language and Hearing Research*, 40(6): 1245-1260.

Torres, F., Barbosa, M. and Maciel, P. (2016). Recurrent copy number variations as risk factors for neurodevelopmental disorders: critical overview and analysis of clinical implications. *Journal of Medical Genetics*, 53(2): 73-90. doi:10.1136/jmedgenet-2015-103366.

Torrey, E.F., Miller, J., Rawlings, R. and Yolken, R.H. (1997). Seasonality of births in schizophrenia and bipolar disorder: a review of the literature. *Schizophrenia Research*, 28(1):1-38.

Torrey, E.F., Rawlings, R.R., Ennis, J.M., Merrill, D.D. and Flores, D.S. (1996). Birth seasonality in bipolar disorder, schizophrenia, schizoaffective disorder and stillbirths. *Schizophrenia Research*, 21(3): 141-149.

Traylor, M., Farrall, M., Holliday, E.G., Sudlow, C., Hopewell, J.C., Cheng, Y.C., Fornage, M., Ikram, M.A., Malik, R., Bevan, S. *et al.* (2012). Genetic risk factors for ischaemic stroke and its subtypes (the METASTROKE collaboration): a meta-analysis of genome-wide association studies. *Lancet Neurology,* 11(11): 951-962. doi:10.1016/S1474-4422(12)70234-X.

Traylor, M., Markus, H. and Lewis, C.M. (2015). Homogeneous case subgroups increase power in genetic association studies. *European Journal of Human Genetics*, 23: 863–869; doi:10.1038/ejhg.2014.194.

Trottier, G., Srivastava, L. and Walker, C.D. (1999) aetiology of infantile autism: a review of recent advances in genetic and neurobiological research. *Journal of Psychiatry and Neuroscience*, 24(2): 103–115.

Tryka, K.A., Hao, L., Sturcke, A., Jin, Y., Wang, Z.Y., Ziyabari, L., Lee, M., Popova, N., Sharopova, N., Kimura, M. and Feolo, M. (2014). NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Research*, 42(Database issue): D975-D979. doi:10.1093/nar/gkt1211.

Turner, T.N., Hormozdiari, F., Duyzend, M.H., McClymont, S.A., Hook, P.W., Iossifov, I., Raja, A., Baker, C., Hoekzema, K., Stessman, H.A. *et al.* (2016). Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *American Journal of Human Genetics*, 98(1): 58-74. doi:10.1016/j.ajhg.2015.11.023.

Uddin, M., Thiruvahindrapuram, B., Walker, S., Wang, Z., Hu, P., Lamoureux, S., Wei, J., MacDonald, J.R., Pellecchia, G., Lu, C. *et al.* (2015). A high-resolution copy-number variation resource for clinical and population genetics. *Genetics In Medecine*, 17(9): 747-752. doi:10.1038/gim.2014.178.

Uhlen, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A. *et al.* (2015). Tissue-based map of the human proteome. *Science*, 347(6220). doi: 10.1126/science.1260419.

Vaags, A.K., Lionel, A.C., Sato, D., Goodenberger, M., Stein, Q.P., Curran, S., Ogilvie, C., Ahn, J.W., Drmic, I., Senman, L. *et al.* (2012). Rare deletions at the neurexin 3 locus in autism spectrum disorder. *American Journal of Human Genetics*, 90(1): 133-141. doi:10.1016/j.ajhg.2011.11.025.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3). doi:10.18637/jss.v045.i03.

van Harssel, J.J., Weckhuysen, S., van Kempen, M.J., Hardies, K., Verbeek, N.E., de Kovel, C.G., Gunning, W.B., van Daalen, E., de Jonge, M.V., Jansen, A.C. *et al.* (2013). Clinical and genetic aspects of PCDH19-related epilepsy syndromes and the possible role of PCDH19 mutations in males with autism spectrum disorders. *Neurogenetics*, 14(1): 23-34. doi:10.1007/s10048-013-0353-1.

Van Loo, K.M.J and Martens, G.J.M. (2007). Genetic and Environmental Factors in Complex Neurodevelopmental Disorders. *Current Genomics*, 8(7): 429–444. doi:10.2174/138920207783591717.

Van Naarden Braun, K., Christensen, D., Doernberg, N., Schieve, L., Rice, C., Wiggins, L., Schendel, D. and Yeargin-Allsopp, M. (2015). Trends in the prevalence

of autism spectrum disorder, cerebral palsy, hearing loss, intellectual disability, and vision impairment, metropolitan atlanta, 1991-2010. *PLoS One*, 10(4): e0124120. doi:10.1371/journal.pone.0124120.

Vardarajan, B.N., Eran, A., Jung, J.Y, Kunkel, L.M. and Wall, D.P. (2013). Haplotype structure enables prioritization of common markers and candidate genes in autism spectrum disrorder. *Translational Psychiatry*, 3: e262. doi:10.1038/tp.2013.38.

Vassos, E., Collier, D.A., Holden, S., Patch, C., Rujescu, D., St Clair, D. and Lewis, C.M. (2010). Penetrance for copy number variants associated with schizophrenia. *Human Molecular Genetics*, 19(17): 3477-3481. doi:10.1093/hmg/ddq259.

Veatch, O.J., Veenstra-Vanderweele, J., Potter, M., Pericak-Vance, M.A. and Haines, J.L. (2014). Genetically meaningful phenotypic subgroups in autism spectrum disorders. *Genes, Brain and Behaviour,* 13(3):276-85. doi:10.1111/gbb.12117.

Veltman, J.A. and Brunner, H.G. (2012). De novo mutations in human genetic disease. *Nature Reviews Genetics*, 13(8): 565-575. doi:10.1038/nrg3241.

Vermeer, S., Hoischen, A., Meijer, R.P., Gilissen, C., Neveling, K., Wieskamp, N., de Brouwer, A., Koenig, M., Anheim, M., Assoum, M. *et al.* (2010). Targeted next-generation sequencing of a 12.5 Mb homozygous region reveals ANO10 mutations in patients with autosomal-recessive cerebellar ataxia. *American Journal of Human Genetics*, 87(6): 813-819. doi:10.1016/j.ajhg.2010.10.015.

Vernes, S.C., Newbury, D.F., Abarhams, B.S., Winchester, L., Nicod, J., Groszer, M., Alarcon, M., Oliver, P., Davies, K.E., Geschwind, D.H. *et al.* (2008). A functional genetic link between distinct developmental language disorders. *New England Journal of Medicine*, 359: 2337-2345. doi:10.1056/NEIJMoa0802828.

Vissers, L.E.L.M., Gilissen, C. and Veltman, J.A. (2016) Genetic studies in intellectual disability and related disorders. *Nature Reviews Genetics*, 17: 9–18. doi:10.1038/nrg3999.

Vogel-Ciernia, A. and Wood, M.A. (2014). Neuron-specific chromatin remodelling: a missing link in epigenetic mechanisms underlying synaptic plasticity, memory, and intellectual disability disorders. *Neuropharmacology*, 80: 18-27. doi:10.1016/j.neuropharm.2013.10.002.

Voineagu, I., Wang, X., Johnston, P., Lowe, J.K., Tian, Y., Horvath, S., Mill, J., Cantor, R.M., Blencowe, B.J. and Geschwind, D.H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*, 474(7351): 380-384. doi:10.1038/nature10110.

von Gontard, A., Pirrung, M., Niemczyk, J. and Equit, M. (2015). Incontinence in children with autism spectrum disorder. *Journal of Pediatric Urology*, 11(5):264.e1-7. doi:10.1016/j.jpurol.2015.04.015.

Vorstman, J.A., Parr, J.R., Moreno-De-Luca, D., Anney, R.J.L., Nurnberger, J.I.Jr. and Hallmayer, J.F. (2017). Autism genetics: opportunities and challenges for clinical translation. *Nature Reviews Genetics*,18(6): 362-276. doi:10.1038/nrg.2017.4.

Vulto-van Silfhout, A.T., Hehir-Kwa, J.Y., van Bon, B.W., Schuurs-Hoeijmakers, J.H., Meader, S., Hellebrekers, C.J., Thoonen, I.J., de Brouwer, A.P., Brunner, H.G., Webber, C. *et al.* (2013a). Clinical significance of de novo and inherited copy-number variation. *Human Mutation*, 34(12): 1679-1687. doi:10.1002/humu.22442.

Vulto-van Silfhout, A.T., van Ravenswaaij, C.M., Hehir-Kwa, J.Y., Verwiel, E.T., Dirks, R., van Vooren, S., Schinzel, A., de Vries, B.B. and de Leeuw, N. (2013b). An

update on ECARUCA, the European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations. *European Journal of Medical Genetics*, 56(9): 471-474. doi:10.1016/j.ejmg.2013.06.010.

Wall, D.P., Esteban, F.J., Deluca, T.F., Huyck, M., Monaghan, T., Velez de Mendizabal, N., Goñí, J. and Kohane, I.S. (2009). Comparative analysis of neurological disorders focuses genome-wide search for autism genes. *Genomics*, 93(2): 120-129. doi:10.1016/j.ygeno.2008.09.015.

Walsh, C.A. and Engle, E.C. (2010). Allelic diversity in human developmental neurogenetics: insights into biology and disease. *Neuron,* 68(2): 245-253. doi:10.1016/j.neuron.2010.09.042.

Wang, K., Dickson, S.P., Stolle, C.A., Krantz, I.D., Goldstein, D.B. and Hakonarson, H. (2010). Interpretation of association signals and identification of causal variants from genome-wide association studies. *American Journal of Human Genetics*, 86(5): 730-742. doi:10.1016/j.ajhg.2010.04.003.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F., Hakonarson, H. and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, 17(11): 1665-1674. doi:10.1101/gr.686190.

Wang, K., Zhang, H., Ma, D., Bucan, M., Glessner, J.T., Abrahams, B.S., Salyakina, D., Imielinski, M., Bradfield, J.P., Sleiman, P.M. *et al.* (2009). Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature,* 459(7246): 528-533. doi:10.1038/nature07999.

Wang, T., Bray, S.M. and Warren, S.T. (2012). New perspectives onto the biology of fragile X syndrome. *Current Opinion in Genetics and Development*, 22(3): 256-263. doi:10.1016/j,gde.2012.02.002.

Wang, X., McCoy, P.A., Rodriguiz, R.M., Pan, Y., Je, H.S., Roberts, A.C., Kim, C.J., Berrios, J., Colvin, J.S., Bousquet-Moore, D. *et al.* (2011). Synaptic dysfunction and abnormal behaviors in mice lacking major isoforms of Shank3. *Human Molecular Genetics*, 20(15): 3093-3108. doi:10.1093/hmg/ddr212.

Warrier, V., Chee, V., Smith, P., Chakrabarti, B. and Baron-Cohen, S. (2015). A comprehensive meta-analysis of common genetic variants in autism spectrum conditions. *Molecular Autism*, 6: 49. doi:10.1186/d13229-015-0041-0.

Webber, C. (2017). Epistasis in Neuropsychiatric Disorders. *Trends in Genetics*, 33(4): 256-265. doi:10.1016/j.tig.2017.01.009.

Weiss, L.A., Arking, D.E., Gene Discovery Project of Johns Hopkins & the Autism Consortium, Daly, M.J. and Chakravarti, A. (2009). A genome-wide linkage and association scan reveals novel loci for autism. *Nature,* 461(7265): 802-808. doi:10.1038/nature08490.

Weiss, L.A., Shen, Y., Korn, J.M., Arking, D.E., Miller, D.T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M.A., Green, T. *et al.* (2008). Association between microdeletion and microduplication at 16p11.2 and autism. *New England Journal of Medicine*, 358(7): 667-675. doi:10.1056/NEJMoa075974.

Weiner, D.J., Wigdor, E.M., Ripke, S., Walters, R.K., Kosmicki, J.A., Grove, J., Samocha, K., Goldstein, J.I., Okbay, A., Byjerg-Grayholm, J. *et al.* (2017). Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nature Genetics*, 49(7): 978-985. doi:10.1038/ng.3836.

White, J.K., Gerdin, A.K., Karp, N.A., Ryder, E., Buljan, M., Bussell, J.N., Salisbury, J., Clare, S., Ingham, N.J., Podrini, C. *et al.* (2013). Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. *Cell,* 154(2): 452-464. doi:10.1016/j.cell.2013.06.022.

Wier, M.L., Yoshida, C.K., Odouli, R., Grether, J.K. and Croen, L.A. (2006). Congenital anomalies associated with autism spectrum disorders. Developmental Medicine and Child *Neurology,* 48(6): 500-507. doi:10.1017/S001216220600106X.

Willcutt, E.G. (2012). The prevalence of DSM-IV attention-deficit/hyperactivity disorder: a meta-analytic review. *Neurotherapeutics*, 9(3):490-499. doi:10.1007/s13311-012-0135-8.

Williams, N.M., Franke, B., Mick, E., Anney, R.J., Freitag, C.M., Gill, M., Thapar, A., O'Donovan, M.C., Owen, M.J., Holmans, P. *et al.* (2012). Genome-wide analysis of copy number variants in attention deficit hyperactivity disorder: the role of rare variants and duplications at 15q13.3. *American Journal of Psychiatry*, 169(2): 195-204. doi:10.1176/appi.ajp.2011.11060822.

Williams, N.M., Zaharieva, I., Martin, A., Langley, K., Mantripragada, K., Fossdal, R., Stefansson, H., Stefansson, K., Magnusson, P., Gudmundsson, O.O. *et al.* (2010). Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: a genome-wide analysis. *The Lancet*, 376(9750): 1401-1408. doi:10.1016/S0140-6736(10)61109-9.

Willsey, A.J., Fernandez, T.V., Yu, D., King, R.A., Dietrich, A., Xing, J., Sanders, S.J., Mandell, J.D., Huang, A.Y., Richer, P. *et al.* (2017). De novo coding variants are stronglu associated with Tourette disorder. *Neuron*, 94(3): 486-499: e9. doi:10.1016/j.neuron.2017.04.024.

Willsey, A.J., Sanders, S.J., Li, M., Dong, S., Tebbenkamp, A.T., Muhle, R.A., Reilly, S.K., Lin, L., Fertuzinhos, S., Miller, J.A. *et al.* (2013). Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell,* 155(5): 997-1007. doi:10.1016/j.cell.2013.10.020.

Wolfe, C.J., Kohane, I.S. and Butte, A.J. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, 6: 227. doi:10.1186/1471-2105-6-227.

Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z. *et al.* (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, 46(11): 1173-1186. doi:10.1038/ng.3097.

Wray, N.R., Purcell, S.M. and Visscher, P.M. (2011). Synthetic associations created by rare variants do not explain most GWAS results. *PLoS Biology*, 9(1): e1000579. doi:10.1371/journal.pbio.1000579.

Wu, S., Ding, Y., Wu, F., Li, R., Xie, G., Hou, J., Mao, P. (2015). Family history of autoimmune diseases is associated with an increased risk of autism in children: A systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews*, 55: 322-332. doi:10.1016/j.neubiorev.2015.05.004.

Wu, Y., Bolduc, F.V., Bell, K., Tully, T., Fang, Y., Sehgal, A. and Fischer, J.A. (2008). A Drosophila model for Angelman syndrome. *PNAS*, 105(34): 12399-12404. doi:10.1073/pnas.0805291105.

Yamagata, A., Yoshida, T., Sato, Y., Goto-Ito, S., Uemura, T., Maeda, A., Shiroshima, T., Iwasawa-Okamoto, S., Mori, H., Mishina, M. and Fukai, S. (2015).

Mechanisms of splicing-dependent trans-synaptic adhesion by PTPδ-IL1RAPL1/IL-1RAcP for synaptic differentiation. *Nature Communications*, 6:6926. doi:10.1038/ncomms7926.

Yang, S.Z., Lin, F.T. and Lin, W.C. (2008). MCPH1/BRIT1 cooperates with E2F1 in the activation of checkpoint, DNA repair and apoptosis. *EMBO Reports*, 9(9): 907-915. doi:10.1038/embor.2008.128.

Yang, Y.J., Baltus, A.E., Mathew, R.S., Murphy, E.A., Evrony, G.D., Gonzalez, D.M., Wang, E.P., Marshall-Walker, C.A., Barry, B.J., Murn, J., Tatarakis, A. *et al.* (2012). Microcephaly gene links trithorax and REST/NRSF to control neural stem cell proliferation and differentiation. *Cell,* 151(5): 1097-1112. doi:10.1016/j.cell.2012.10.043.

Yoon, D.Y., Gause, C.D., Leckman, J.F. and Singer H.S. (2007). Frontal dopaminergic abnormality in Tourette syndrome: a postmortem analysis. *Journal of the Neurological Science*s, 255(1-2): 50-6. doi:10.1016/j.jns.2007.01.069.

Yousefichaijan, P., Sharafkhah, M., Rafiei, M. and Salehi, B. (2016). Attention-deficit/hyperactivity disorder in children with overactive bladder; a case-control study. *Journal of Renal Injury Prevention*, 5(4):193-199. doi:10.15171/jrip.2016.41

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ruba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B.D. *et al.* (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature,* 515(7527): 355-364. doi:10.1038/nature.13992.

Yuen, R.K.C., Merico, D., Bookman, M., Howe, J.L., Thiruvahinhindrapuram, B., Patel, R.V., Whitney, J., Deflaux, N., Bingham, J., Wang, Z. *et al.* (2017). Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nature Neuroscience*, 20(4): 602-611. doi:10.1038/nn.4524.

Yu, T.W., Chahrour, M.H., Coulter, M.E., Jiralerspong, S., Okamura-Ikeda, K., Ataman, B., Schmitz-Abe, K., Harmin, D.A., Adli, M., Malik, A.N. *et al.* (2013). Using whole-exome sequencing to identify inherited causes of autism. *Neuron,* 77(2): 259-273. doi:10.1016/j.neuron.2012.11.002.

Zapala, M.A., Hovatta, I., Ellison, J.A., Wodicka, L., Del Rio, J.A., Tennant, R., Tynan, W., Broide, R.S., Helton, R., Stoveken, B.S. *et al.* (2005). Adult mouse brain gene expression patterns bear an embryologic imprint. *PNAS*, 102(29): 10357-10362.

Zayats, T., Athanasiu, L., Sonderby, I., Djurovic, S., Westlye, L.T., Tamnes, C.K., Fladby, T., Aase, H., Zeiner, P., Reichborn-Kjennerud, T. *et al.* (2015). Genome-wide analysis of attention deficit hyperactivity disorder in Norway. *PLoS One*, 10(4): e0122501. doi:10.1371/journal.pone.0122501.

Zhang, L., Yang, S., Wennmann, D.O., Chen, Y., Kremerskothen, J. and Dong, J. (2017). KIBRA: In the brain and beyong. *Cell Signalling*, 26(7): 1329-1399. doi:10.1016/j.cellsig.2014.02.023

Zhang, Q.C., Petrey, D., Garzón, J.I., Deng, L. and Honig, B. (2013). PrePPI: a structure-informed database of protein-protein interactions. *Nucleic Acids Research*, 41(Database issue): D828-D833. doi:10.1093/nar/gks1231.

Zheng-Bradley, X., Nehrt, N.L., Clark, W.T., Radivojac, P. and Hahn, M.W. (2011). Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Computational Biology*, 7(6):e1002073. doi:10.1371/journal.pcbi.1002073.

Zhu, Q., Ge, D., Maia, J.M., Zhu, M., Petrovski, S., Dickson, S.P., Heinzen, E.L., Shianna, K.V. and Goldstein, D.B. (2011). A genome-wide comparison of the functional properties of rare and common genetic variants in humans. *American Journal of Human Genetics*, 88(4): 458-468. doi:10.1016/j.ajhg.2011.03.008.

Zilhao, N.R., Padmanabhuni, S.S., Pagliaroli, L., Barta, C., Smit, D.J.A., Cath, D., Nivard, M.G., Baselmans, B.M.L., van Dongen, J., Paschou, P. and Boosma, D.I. (2015). Epigenome-Wide Association Study of Tic Disorders. *Twin Research and Human Genetics*, 18(6): 699-709. doi:http://dx.doi.org/10.1017/thg.2015.72.

SUPPLEMENTARY INFORMATION

S1 Overview of genes implicated in rare neurological syndromes and NDDs

S1.1 Genes involved in brain malformations and NDDs

The disruption of neurodevelopmental processes, and more specifically of the proliferation, migration and differentiation of neural progenitors, is associated with a variety of cognitive, behavioural and unsurprisingly morphological abnormalities. This section reviews several classes of genes associated with two brain malformations (i.e. microcephaly and lissencephaly) (for all cortical malformations see Barkovich *et al.*, 2012) and generalises their implication across NeuroDevelopmental Disorders (NDDs).

S1.1.1 Microcephaly

Microcephaly is a birth defect characterised by a disproportionately smaller head compared to the body, which presumably results from abnormalities, and more specifically reduction, in neuronal proliferation (compared to the increase in proliferation or decrease in apoptosis observed in macrocephaly – abnormally large head). Many microcephaly-associated genes have roles in neurogenesis and particularly in mitotic processes, such as DNA repair, spindle formation and regulation of cell cycle length.

The first microcephaly gene discovered, microcephalin (*MCPH1*; causes autosomal primary recessive microcephaly, called MCPH), encodes a DNA

damage response protein. Dysfunction of microcephalin compromises cell cycle arrest during response to DNA damage and induces early cell cycle exit followed by differentiation, hence depleting the pool of neuroprogenitors (Liu *et al.*, 2016b). Subsequently, other DNA repair genes have been linked to microcephaly, including *PNKP*, encoding proteins involved in single- and double-strand DNA repairs by facilitating DNA ligation (Shen *et al.*, 2010).

Many of the genes implicated in MCPH encode centrosomal components (i.e. *CENPJ*, *CEP152*, *CEP135*, *STIL* and *CDK5RAP2*) or proteins involved in mitotic spindle formation (i.e. *WDR62* and *ASPM*) (Chavali *et al.*, 2012). Centrosomes have key roles during mitosis as they segregate with chromosomes to form the mitotic spindle, duplicate and control DNA migration to daughter cells. Dysfunction of centrosomal proteins is likely to deplete the neuroprogenitor pool by (1) improperly orienting the mitotic spindle and switching symmetric division to asymmetric division (i.e. WDR62, CENPJ and STIL). Alternatively, (2) it can induce the construction of an abnormal number of centrosomes, which delays spindle formation and subsequently mitosis, thus indirectly inducing early cell cycle exit of daughter cells followed by differentiation (i.e. WDR62, ASPM, CDK5RAP2 and STIL) (Chavali *et al.*, 2012). Additionally, (3) an abnormal number of centrosomes can lead to the formation of mono- or multi-polar spindles, impairing chromosomal distribution and causing aneuploidy and cell death.

Genes encoding Transcription Factors (TFs) have also been implicated in MCPH, such as *MCPH1*, which encodes a protein regulating the expression

of other DNA damage response and apoptosis genes (Yang *et al.*, 2008; Barkovich *et al.*, 2012). Furthermore, the identification of a mutation in the *ZNF335* gene in a consanguineous family with severe microcephaly (Yang *et al.*, 2012) supports a role for chromatin modifiers in disease pathogenesis. Indeed, *ZNF335* encodes a component of the trithorax H3K4-methylation complex, which regulates neural gene expression and cell fate.

S1.1.2 Lissencephaly

Lissencephaly is a disorder associated with simplified gyral brain patterns, stemming from the improper development of brain folds and grooves. This brain abnormality involves a disorganised neuronal network, resulting from neuronal migration defects.

Neuronal migration relies on the cytoskeletal remodelling of microtubules and actin. Briefly, the centrosome orientation drives microtubules to pull the nucleus towards that direction (enabling migration), while actin filaments from the growth cone adhere to new substrates based on the extracellular guidance cues (defining the migration path) (Moon and Wynshaw-Boris, 2013). Dysfunction of such cytoskeletal processes causes lissencephaly. For example, heterozygous mutations and deletions of the *LIS1* and *DCX* genes, encoding microtubule stabilisers, which control cytoskeletal transport and thus migration, have been implicated in lissencephaly. Specifically, the *LIS1*-encoded enzyme binds the cytosplasmic dynein complex, which regulates the cytoskeletal changes generating microtubule-pulling forces, while the

doublecortin protein, encoded by *DCX*, forms a complex with the LIS1 and dynein (Moon and Wynshaw-Boris, 2013). Furthermore, mutations in genes encoding tubulins, collectively forming microtubules, also cause neuronal migration disorders by restricting microtubule polymerisation. Mutations in many tubulin genes have been implicated in lissencephalies including *TUBA1A*, responsible for up to 4% and 30% of cases of classic lissencephaly and lissencephaly with cerebellar hypoplasia, respectively; *TUBB2B* causing type II lissencephaly and polymicrogyria; *TUBA8* associated with polymicrogyria; and *TUBB3* leading to syndromes characterised by eye movement restrictions (Tischfield and Engle, 2010; Barkovich *et al.*, 2012).

The ExtraCellular Matrix (ECM) provides a microenvironment influencing various neuronal processes such as migration, during which the ECM acts as a scaffold. Accordingly, many ECM constituents have been implicated in cortical malformations, including genes encoding dystroglycans and collagen. For example, reelin is an ECM-associated glycoprotein, encoded by the *RELN* gene, which binds lipoprotein receptors expressed on migrating neurons, generating a signalling cascade, which controls migration by recruiting various proteins including LIS1 (Barros *et al.*, 2011). Mutations and deletions of *RELN* have been implicated in a severe form of autosomal recessive lissencephaly with cerebellar hypoplasia (Hong *et al.*, 2001).

S1.1.3 Broader NDD categories

Many of the abovementioned genes are pleiotropic. For example, *WDR62*, the second most common microcephaly gene, is also associated with other cortical malformations (i.e. lissencephaly, pachygryia and schizencephaly) and more generally Intellectual Disability (ID), suggesting it confers susceptibility to broader NDD categories (Bilgüvar *et al.*, 2010; Chen *et al.*, 2014). Similarly, while deletions of *DYRK1A* are associated with microcephaly and ID, duplications cause Down's syndrome, characterised by growth delays, facial abnormalities and ID (Smith *et al.*, 1997; Ji *et al.*, 2015), and *de novo* (*dn*) mutations have recently been implicated in Autism Spectrum Disorder (ASD) (De Rubeis *et al.*, 2014; Dang *et al.*, 2017). Furthermore, NDD-associated genes are also able to cause brain malformations. For example, an appreciable number of individuals with *dn* Loss of Function mutations (LoF; include nonsense, frameshift and splice-site variants) in the ASD-associated *CHD8* gene exhibit macrocephaly (Barnard *et al.*, 2015). *CHD8* encodes a chromodomain helicase binding protein, which acts as a master transcriptional regulator during neurodevelopment and modulates the expression of a large number of ASD-implicated genes (Barnard *et al.*, 2015). Although its exact function is unclear, it appears to promote cellular proliferation, which could be a pathogenic mechanism contributing to macrocephaly.

Overall, these examples indicate that genes participating in neurogenesis and migration affect multiple gene classes, including cell cycle control and mitosis,

transcriptional regulators, cytoskeletal remodellers and ECM components. We highlight the implication of these genes in the aetiology of brain malformations, as well as broader NDD categories.

S1.2 Synaptic genes in NDDs

Copy Number Variant (CNV) and exome studies have implicated many functional classes of genes involved in processes of synaptic formation and plasticity in the pathogenesis of neurological syndromes and NDDs. Evidence for the involvement of cell adhesion molecules, scaffolding proteins, ion channels and NMDA receptors, signal transduction components, TFs and chromatin modifiers is presented here.

S1.2.1 Cell adhesion molecules

Several groups of Cell Adhesion Molecules (CAMs) have been linked to NDDs, including neurexins and neuroligins, which encode pre- and post-synaptic CAMs, respectively (Betancur *et al.*, 2009). For example, while *NRNX1* and *CNTNAP2* are implicated in the pathogenesis of ASD, schizophrenia, epilepsy and ID (*CNTNAP2* is also associated with OCD, ADHD and SLI) (Dabell *et al.*, 2013; Poot, 2015), *NLGN3* and *NLGN4X* have been linked to ID and ASD (Laumonnier *et al.*, 2004) and variants affecting *NRXN2* and *NRXN3* have been observed in ASD (Gauthier *et al.*, 2011; Vaags *et al.*, 2012). Furthermore, cadherins, encoded by the *CHD13* gene, and protocadherins, encoded by the *PCDH9*, *PCDH10* and *PCDH19* genes,

are involved in dendritic self-avoidance (Lefebvre *et al.*, 2012) and disrupted in ASD. For example, *PCDH19* causes epilepsy and mental retardation restricted to females (Betancur *et al.*, 2009; Sanders *et al.*, 2011; van Harssel *et al.*, 2013). Another prominent type of CAMs are immunoglobulins, such as those encoded by *L1CAM* for which over 350 deletions and LoF mutations have been uncovered and cause many X-linked recessive neurological disorders, including severe ID (Patzke *et al.*, 2016). Although the precise function of L1CAM is unknown, its dysfunction is likely to affect the development of axons and dendrites, thus impairing synaptic transmission (Patzke *et al.*, 2016). Similarly, deletions and LoF mutations in *IL1RAPL1*, encoding a protein, which upon receptor binding induces differentiation of pre- and post-synaptic structures at excitatory synapses, have been implicated in X-linked mental retardation and ASD (Carrié *et al.*, 1999; Piton *et al.*, 2008; Yamagata *et al.*, 2015). Finally, mutations in *SLITRK1*, a gene encoding a protein with leucine-rich repeats, have been observed in Obsessive Compulsive Disorder (OCD) and Gilles de la Tourette Syndrome (GTS) (Scharf *et al.*, 2008; O'Roak *et al.*, 20010; Ozomaro *et al.*, 2013).

S1.2.2 Scaffolding proteins

Scaffolding proteins contribute to the construction of specialised pre- and post-synaptic structures, such that approximately 500 PostSynaptic Density (PSD) proteins have been described in humans (Collins *et al.*, 2006). These occur in diverse combinations across different types of synapse. The MAGUK, SHANK and gephyrin families of scaffolding proteins have been related to

NDDs. DLG proteins from the MAGUK super-family facilitate the formation and maintenance of PSDs by recruiting components such as the NMDA receptor subunits. Of the four *DLG* homologues, three have been implicated in NDD susceptibility. Specifically, LoF and non-synonymous (collectively termed Likely Gene-Disrupting; LGD) variants have been identified in *DLG3* among patients with X-linked mental retardation (Tarpey *et al.*, 2004). Furthermore, CNV deletions overlapping *DLG1* and *DLG2* are present in individuals with schizophrenia and ASD, with schizophrenia patients additionally exhibiting *dn* LoF mutations in both genes (Kirov *et al.*, 2012; Egger *et al.*, 2014; Fromer *et al.*, 2014; Kumar *et al.*, 2016). *CASK*, another gene from the MAGUK family, has been linked to several neurological disorders with accompanying ID (Moog *et al.*, 2015). Similarly, to MAGUK proteins, gephyrins recruit GABA and glycine receptor subunits to inhibitory synapses. Microdeletions mapping to the gephyrin encoded by the *GPHN* gene have been identified in NDDs although their implication in disease remains to be validated (Lionel *et al.*, 2013). Finally, SHANK proteins encoded by the *SHANK1*, *SHANK2* and *SHANK3* genes are master scaffold proteins localised at the PSD of glutaminergic synapses and have been linked to multiple clinical presentations of NDDs. *SHANK3* haploinsufficiency was initially identified as the cause of the 22q13 microdeletion syndrome (i.e. Phelan-McDermin syndrome), following which mutations in all three genes were uncovered among ASD patients (Sala *et al.*, 2015).

S1.2.3 Ion channels and receptors

Dysfunction of ion channels and neurotransmitter receptors has consistently been associated to NDDs. Specifically, many forms of syndromic epilepsies arise as a result of mutations in genes encoding subunits of GABA-ergic receptors as well as calcium, sodium and potassium channels, such as the *GABRG2*, *KCNMA1*, *SCN1A* and *KCNT1* genes, respectively (Spillane *et al.*, 2016). Furthermore, genes encoding calcium (i.e. *CACNA1C*), sodium (i.e. *SCN1A* and *SCN2A*) and potassium (i.e. *KCNT1*) channels have been more broadly linked to ASD, ID and schizophrenia (Imbrici *et al.*, 2013). GABA-ergic receptor subunit genes have collectively been implicated in ASD and related syndromes, such that 15q11-q13 microdeletions and microduplication, overlapping *GABRB3*, *GABRA5* and *GABRG3*, provide susceptibility to ASD and Prader-Willi or Angelman's syndromes, respectively (Coghlan *et al.*, 2012). Glutamate receptors, which mediate transmission and plasticity at excitatory synapses, have also been implicated in the pathogenesis of NDDs. Indeed, penetrant gene mutations or CNVs in genes encoding many of the α-Amino-3-hydroxy-5-Methyl-4-isoxazolePropionic Acid (AMPA) (i.e. *GRIA2* and *GRIA3*) and *N*-Methyl-D-Aspartate (NMDA) (i.e. *GRIN2A* and *GRIN2B*) receptor subunits have been observed in ASD and in particular in ID (Soto *et al.*, 2014).

S1.2.4 Signal transduction pathways

A major mechanism of synaptic plasticity involves the activity-dependent translational regulation of dendritic messenger RNAs to drive long-term potentiation and depression. Accordingly, it is unsurprising that loss of function of key translational regulators, such as the Fragile X Mental Retardation Protein (FMRP), leads to aberrant protein synthesis, ultimately resulting in a variety of synaptic and behavioural abnormalities observed in the Fragile X syndrome. Misregulation of signalling pathways controlling activity-dependent biological processes has also been observed in NDDs. A commonly misregulated pathway in ASD is the PI3K/Akt/mTOR pathway, which regulates various processes including protein synthesis, transcription, actin dynamics and neuronal morphology. Moreover, mutations in genes encoding the tuberous sclerosis complex (i.e. *TSC1* and *TSC2*) and *PTEN*, which act as inhibitors of mTOR signalling, confer strong susceptibility to ASD and ID and are associated with rare syndrome such as tuberous sclerosis and macrocephaly (Enriquez-Barreto and Morales, 2016). Other signalling pathways, with high degrees of cross-talk have also been implicated in NDDs. For example, mutations in the *NF1* gene, encoding a protein, which indirectly upregulates PI3K signalling by enhancing RAS/MAPK pathway activity, cause neurofibrotomis, a condition characterised by tumour growths in the nervous system, as well as ASD and epilepsy (Enriquez-Barreto and Morales, 2016). Similarly, mutations in *MECP2*, which are associated with Rett syndrome, indirectly induce PI3K over-activation through a poorly understood mechanism (Enriquez-Barreto and Morales, 2016). The ubiquitin-proteasome pathway has

also been linked to NDDs through the *UBE3A* gene, which encodes a ligase crucial in enabling activity-dependent postsynaptic protein degradation. *UBE3A* duplications confer susceptibility to ASD while mutations and deletions cause Angelman syndrome (Sun *et al.*, 2015). Mutations in genes encoding Ras GTPases modulators of synaptic plasticity, their regulators and their downstream targets can also impair signalling cascades. For example, *SYNGAP1* encodes a calcium-regulated Ras GTPase activating protein, which controls the strength of excitatory synapses by managing the expression of postsynaptic AMPA receptors; *SYNGAP1* mutations represent common causes of ASD, ID and epilepsy (Jeyabalan and Clement, 2016).

S1.2.5 Transcription factors and chromatin regulators

Genetic variants in regulators of gene expression, such as TFs and chromatin regulators have been increasingly uncovered in NDDs and rare syndromes. Although these often exert their effect during neurogenesis and neuronal migration as seen in **section S1.1**, their disruption also impacts activity-dependent gene expression, crucial for synaptic plasticity.

For example, TFs from the Myocite-Enhancer Factor (MEF) family modulate long-term memory formation in the adult brain, such that activity-dependent activation of MEF proteins leads to gene expression programs aiming to control the number of excitatory synapses formed (Rashid *et al.*, 2014). Accordingly, disruptions of *MEF* genes have been implicated in ASD, ID and epilepsy. Specifically, penetrant mutations in *MEF2C* are associated with

severe ID while mutations in MEF2C-regulated genes lead to epilepsy or other syndromes, such as Angelman syndrome in the case of *UBE3A* (Rashid *et al.*, 2014).

Chromatin regulators modulate gene expression indirectly, by controlling the access of the transcriptional machinery to genomic DNA. Various types of modifiers have been implicated in NDDs, such as (1) genes encoding histone acetyltransferases (e.g. mutations in *CREBBP* and its paralog *EP300* cause Rubinstein-Taybi syndrome, characterised by mental retardation and have been liked to ASD) (Roelfsema *et al.*, 2005). A second type of modifiers involves (2) histone deacetylases (e.g. *HDAC4* encodes a protein, which shuttles between the nucleus and cytoplasm as a result of NMDA receptor signalling and regulates the expression of synaptic genes; *HDAC4* mutations are associated with a range of cognitive and developmental abnormalities seen in the brachydactyly with mental retardation syndrome; Sando *et al.*, 2012). Furthermore (3) mutations in genes encoding subunits of nucleosome remodelling BAF complexes (i.e. *SMARCA2, SMARCA4*, *SMARCB1*, *ARID1A* and *ARID1B*) have been linked to ID, ASD and many rare clinical syndromes (Vogel-Ciernia and Wood, 2014). Finally (4) mutations of genes encoding other chromatin remodelling enzymes, such as CHD proteins (i.e. *CHD2*, *CHD7* and *CHD8*) have been linked to ASD, epilepsy and CHARGE syndrome (Ronan *et al.*, 2013). Nonetheless the study of *CHD* genes focuses predominantly on neurodevelopment and their role in activity-dependent synaptic plasticity remains to be proven.

S2 Construction of the phenotypic linkage network (PLN)

In an effort to investigate gene function and interactions, datasets that provide physical Protein-Protein Interactions (PPIs), gene expression data and gene annotations have been produced. Specifically, annotation datasets include (1) functional gene annotations from Gene Ontology (GO) (Gene Ontology Consortium, 2000), the Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000; Kanehisa *et al.,* 2017), Reactome (Croft *et al.,* 2014*;* Fabrecat *et al.,* 2016) and String (Szklarczyk *et al*., 2015). Additionally, (2) data from the Mouse Genome Informatics (MGI) (Blake *et al.,* 2017) provides annotations regarding the mouse phenotypes associated with the disruption of unique gene orthologues. As the function of a large number of genes remains unknown, such datasets can be used to infer functional associations. This relies on the assumption that genes with similar functions participate in the same biological pathways, interact and/or are co-expressed.

The use of the aforementioned datasets, individually, may be inadequate for the study of large numbers of genes, due to confounding factors such as incompleteness, biases and false-positives. Therefore, we sought to build a comprehensive gene network, in which pairs of human genes are assigned a functional similarity score, based on the combination of such genomics datasets following the method described by Honti *et al*., 2014 (**Fig 2.1**).

Individual genomics datasets were processed in different ways in order to score gene pairs (**Table 2.4**). Using PPI datasets, gene pairs were scored

relative to the number and types of experiments suggesting physical interaction (i.e. low throughput evidence had higher scores than high throughput experiments) (Honti *et al.*, 2014). Expression datasets were used to derive co-expression coefficients between pairwise genes combinations (**section 2.4.1**). Annotation data (i.e. GO, MGI, KEGG, Reactome and String) was processed using semantic similarity, such that the distance between two annotations was weighted by their information content (i.e. in a hierarchically-organised resource, general and specific annotations have low and high information contents, respectively). Gene pairs were assigned the maximum semantic similarity score across all pairwise combinations of their annotations. We refer to the MGI semantic similarity dataset as the benchmark.

Based on the assumption that genes involved in the same biological pathways influence similar phenotypes when disrupted in mouse models (Robinson and Webber, 2014), we assessed and re-scored each dataset on the MGI benchmark (Honti *et al.*, 2014). For this, we ordered gene pairs from a given dataset and evaluated the median score of bins of 1,000 pairs against their median MGI similarity. Informative dataset scores correspond to MGI similarities above the upper 45^{th} percentile of all MGI similarities for the given dataset (**Fig 2.1b**). We regressed informative dataset scores on the MGI benchmark using linear regression and re-scored all gene pairs – even if they did not have MGI scores – according to this function. Ultimately, we re-scored all datasets on the same MGI scale. We then combined these re-scored datasets, such that when many datasets suggested functional linkage

between two genes, we summed their scores, weighting them according to the lines of evidence for functional associations (**Fig 2.1c**).

S.3 Phenotypic SSC dataset

The Simons Simplex Collection (SSC) provides extensive clinical data for individuals with ASD from 2,644 simplex families, recruited as part of the Simons Foundation Autism Research Initiative (SFARI). Briefly, patients were evaluated for diagnostic and core autism features, adaptive behaviour, motor skills and cognitive abilities using various instruments. Additionally, detailed clinical descriptions regarding patients and family maladies, abnormalities, intolerances as well as life and personal information are available. Phenotypic data has been integrated and can be downloaded from the SFARI Base database.

S.3.1 Construction of the phenotypic dataset

Exploiting the rigorous clinical characterisation of SSC patients, we constructed a binary phenotypic dataset to enable subsequent comprehensive patient comparisons across a wide range of non-core ASD symptoms. We undertook an exhaustive literature search to define clinically relevant phenotypic categories (**Fig S4.1**). Aiming to cover the full autistic spectrum, we selected 63 SSC phenotypes encompassing comorbid disorders, clinical and perinatal features. The categories and phenotypes considered are described below and summarised in **Table 4.1**.

Allergies: Environmental and food intolerances are present in up to 90% of ASD patients and have been shown to influence autism features, such as

social affect, restricted and repetitive behaviours as well as aggressive character. The SSC provides data on four environmental (i.e. animals, dust, latex and pollen) and eight food (i.e. casein, egg, fish or shellfish, gluten, milk, nuts, dyes and wheat) allergies, such that 27.9% and 15.9% of patients display at least one environmental or food allergy, respectively. Of note, while no gender bias has previously been reported (Mrożek-Budzyn *et al*., 2013), an excess of SSC males displays allergies ($p = 3.96 \times 10\text{-}6$, Odds Ratio, OR = 1.82; two-sided Fisher test).

Autoimmune disorders: A strong link between immune system abnormalities, such as decreased lymphocyte number, increased α-interferon levels and reduced TGF-β levels (Trottier *et al*., 1999), and autism has been described, suggesting autoimmunity could be a possible mechanism underlying ASD. Moreover, ASD patients display increased levels of circulating antibodies reactive against the brain, which can occur in autoimmune disorders, such as Systemic Lupus Erythematosus (SLE) (Trottier *et al*., 1999; Ashwood and Van de Water, 2004; Goines and Van der Water, 2010). We, therefore, considered ten autoimmune disorders provided by the SSC, many of which occur at rates of less than 1% (i.e. diabetes mellitus, SLE, hypothyroidism, hashimotos tyroiditis, rheumatoid arthritis, multiple sclerosis and celiac disease).

Birth defects: Congenital abnormalities have been linked to ASD, such that birth defects, including lip palate, heart and brain malformations, are 1.6- to 2.4-fold more common in ASD than in the general population (Hultman *et al.*, 2002; Lauritsen *et al.*, 2002; Wier *et al.*, 2006; Dawson *et al.*, 2009). Here, we

considered three organ defects (i.e. congenital heart defect, open spine and kidney defect), three facial or physical abnormalities (i.e. dysmorphism, cleft lip palate and polydactyly) and the presence of microcephaly (i.e. disproportionately small head compared to body). While 14.5% and 2.8% of SSC patients are dysmorphic or microcephalic, respectively, the remaining abnormalities occur in less than 1.1% of individuals. Of note, following imputation of missing data (**section S3.2**), the incidence of microcephaly is increased 1.6-fold (2.8% to 4.3%; **Table 4.1**). This is because the presence of microcephaly is most strongly correlated to the presence of epilepsy and seizures (Pearson's correlation coefficients $r > 0.3$). Accordingly, during imputation of missing values, a significant proportion of individuals with epilepsy or seizures are annotated with microcephaly, as well.

Epilepsy and seizures: While epilepsy generally affects between 5% and 38% of ASD patients (Jeste, 2011; Mannion and Leader, 2013), it has a lower prevalence in the SSC. Indeed, while 5.3% of individuals have experienced at least one seizure episode (febrile and non-febrile seizures are considered together), only 2% have an epilepsy diagnosis (from the Aberrant Checklist Behaviour test). Following imputation of missing data (**section S3.2**), the incidence of epilepsy or seizures increases to 11.33%, in agreement with previous reports. Contrary to previous findings (Giarelli *et al*., 2010), we found no excess of affected females ($p = 0.11$; $OR = 1.35$; two-sided Fisher test).

Gastro-intestinal problems: In line with previous reports, 43.3% of SSC patients present some kind of Gastro-Intestinal (GI) problem, such as

bloating, constipation, Crohn's disease, diarrhoea, or Gastro-oEsophageal Reflux disease (GER). Excluding under-represented GI diseases, we focused, specifically, on problems paediatric gastroenterologists recommend ASD patients be evaluated for, due to high odds ratios > 2.4 in ASD (McElhanon *et al*., 2014; Buie *et al*., 2010). These include abdominal pain, constipation, diarrhoea and GER. Among SSC patients, constipation is the most common problem (i.e. 26.7%), followed by diarrhoea and GER (i.e. 11.9% and 10.3%, respectively) while vomiting and severe abdominal pain are rare, with an incidence of 5.5% each. Of note, the presence of GI problems reflects the parents' or caretakers' opinion, lacking a formal medical investigation, and is therefore biased across families.

Gestational age: Gestational age represents a perinatal risk factor as premature (i.e. < 37 weeks) or post-term (i.e. > 41 weeks) births have been linked to ASD susceptibility and severity (Sugie *et al*., 2005; Schendel and Bhasin, 2008; Williams *et al*., 2008; Guinchat *et al*., 2012; Movsas and Paneth, 2012; Leavey *et al*., 2013; Atladóttir *et al*., 2016; Guy *et al*., 2015). The SSC provides data regarding the gestational age at birth, enabling the characterisation of 8.5% pre- and 5.3% post-mature SSC patients.

More generally, we considered the overall optimality of the pregnancy, based on the observation that pregnancy and perinatal difficulties and complications are associated with ASD in the offspring (Lyall *et al.*, 2012). Specifically, the SSC provides a pregnancy optimality code, such that we defined sub-optimal

pregnancies using a code of 0 (i.e. below the median optimality score) (i.e. 32.8% of patients).

Language delay: Delays in spoken language are observed in about 50% of ASD patients, who on average produce their first words around 38 months compared to 12 months in normally developing children (Alarcón *et al*., 2008; Howlin, 2003). While delays in first spoken words and phrases are evident after the ages of 24 and 36 months, respectively, we shifted these thresholds to 36 and 42 months in the aim of examining severer phenotypes. The ages, at which patients have spoken their first words and put together words in a sentence, are provided within the Autism Diagnostic Interview Revised (ADI-R). Patients, who never reached the respective milestone (ADI-R codes 994 and 997), were also defined as having word or phrase delay. While most patients with language delay have phrase delay (i.e. 98.06%), only half (the more severely impaired half) also have a word delay (i.e. 51.90%).

Low functioning ASD: Up to 70% of ASD patients display comorbid ID (Van Naarden Braun *et al.*, 2015; Christensen *et al.*, 2016), although official ID diagnoses are not available for SSC patients. Therefore, under the assumption that patients with lower Intellectual Quotients (IQs) are more likely to share a common aetiopathology with ID than individuals with higher IQs, we considered low functioning ASD. Considering Differential Ability Scale II (DAS-II), Wechsler Adult Intelligence Scale (WASI) and Wechsler Intelligence Scale for Children IV (WISC-IV) measures, as they provide comparable IQ scores, we defined patients with a non-verbal IQ < 60 as low functioning.

Individuals assessed on the Mullen scale were assigned missing values, which were further imputed. Following imputation (**section S3.2**), 13.01% of patients were defined as low functioning. Of note, females are over-represented ($p = 1.93 \times 10\text{-}8$, OR = 2.34; two-sided Fisher test), consistent with previous reports revealing the male:female autism ratio decreases with more severe cognitive impairments (Volkmar *et al*., 1993; Yeargin-Allsopp *et al*., 2003; Giarelli *et al*., 2010).

Mother autoimmune disorders: Many epidemiologic studies found an association between a maternal or family history of autoimmune disorders, such as rheumatoid arthritis (Comi *et al*., 1999; Atladóttir *et al*., 2009; Brimberg *et al*., 2013), psoriasis (Croen *et al*., 2005) and SLE (Brimberg *et al*., 2013), and ASD risk in the offspring. This link stems from the observation that maternal auto-antibodies are able to interact with foetal brain antigens and rodent offsprings, gestationally exposed to brain-reactive antibodies, present increased ASD behaviours as compared to those exposed to control immunoglobulins (Dalton *et al*., 2003; Singer *et al*., 2009; Braunschweig *et al*., 2012). Moreover, SSC mothers have higher levels of anti-brain antibodies and are significantly more commonly diagnosed with rheumatoid arthritis or SLE compared to mothers of healthy children (Brimberg *et al*., 2013).

As we were unable to retrieve the list of SSC mothers with brain-reactive antibodies, we reasoned mothers with autoimmune disorders are likely to display increased levels of brain-reactive auto-antibodies able to cross to the embryo during gestation. We, therefore, considered the presence of eleven

autoimmune disorders among mothers of SSC patients. The most common disorders are asthma (i.e. 11% of patients), bowel disorder (i.e. 3.2% of patients) and psoriasis (i.e. 2.7% of patients). Conversely, other conditions, such as rheumatoid arthritis, diabetes mellitus, multiple sclerosis, adrenal insufficiency, thyroid disorders, SLE and celiac disease, are rare and occur in < 2% of mothers.

Motor disease and neurologic problems: Motor problems are common in ASD, such that 80.2% of SSC patients are diagnosed with a Development Coordination Disorder (DCD), based on the DCD questionnaire, which assesses control during movement, fine and gross motor skills and general coordination. This is consistent with previous reports suggesting 75% of ASD patients have some type of motor dysfunction (Duchan and Patel, 2012). We, additionally, considered three variously debilitating neurologic problems affecting movement and coordination, namely cerebral palsy (i.e. 0.3% of patients), movement abnormalities (i.e. 3.2% of patients) and excessive uncoordination and clumsiness (i.e. 20.2% of patients).

Psychiatric disorders: In addition to ID and epilepsy, we considered eight comorbid psychiatric disorders, including (1) Attention Deficit Hyperactivity Disorder (ADHD) (i.e. 14.3% of patients), (2) OCD (i.e. 3.5% of patients), (3) anxiety disorder (i.e. 3.4% of patients), (4) behavioural disorder (i.e. 2.1% of patients), (5) depression (i.e. 0.9% of patients), (6) bipolar disorder (i.e. 0.5% of patients), (7) Gilles de la Tourette Syndrome (GTS) (i.e. 1.4% of patients) or the presence of chronic vocal or motor tics (i.e. 4.9% of patients) and (8)

schizophrenia. Schizophrenia was subsequently removed from the study because of a null incidence, likely due to the discrepancy between the age of the cohort (mean age = 9.03 years) and disease onset (i.e. > 16 years). More generally, with the exception of GTS, psychiatric comorbidities are less frequent in the SSC than previously reported, possibly owing to the high functioning nature of the SSC cohort (Simonoff *et al*., 2008; Van Steensel *et al*., 2011; Russell *et al*., 2013).

Regression: Up to 50% of ASD patients experience regression, during which they lose previously acquired developmental skills. While most researchers argue that language is central to regression, there is no consensus as to which skills are lost, prompting us to consider the loss of both language and non-linguistic social or motor skills. Regression measures are provided by the ADI-R test. Language regression, involving the loss of use of 3 meaningful words, occurs in 17.7% of patients. We defined the loss of social skills as the mild or severe loss of constructive play (i.e. 6.4% of patients) or social engagement (i.e. 23.3% of patients). Lastly, individuals displaying mild or severe loss of hand movements (i.e. 3.3% of patients), motor skills (i.e. 4.3% patients) or self-help (i.e. 4.1% of patients) were qualified as having motor regression.

Severe ASD: The Autism Diagnostic Observation Schedule (ADOS) provides a standardised continuous measure of overall ASD symptom severity, termed Calibrated Severity Score (CSS) (Gotham *et al.*, 2009). This score should not be confounded by age, language skills or IQ (Gotham *et al.*, 2009). We

constructed a phenotypic variable relating to the presence of severe ASD, which was defined as a maximal CSS score of 10 / 10, qualifying 14.5% of SSC patients as severely autistic.

Sleep problems: A wide range of sleep disturbances occur in individuals with ASD (Cohen *et al*., 2014), with up to 90% of patients experiencing difficulties, depending on the cohort and types of problems evaluated. The SSC provides data regarding many sleep abnormalities, including reduced total sleep time, resistance for bedtime, awakenings through the night, daytime sleepiness amongst many others, such that 68% of patients exhibited at least one of these common sleep problems at the time assessment (the associated prevalence across lifetime is higher). Due to the lack of consensus definition of sleep problems and the existence of many confounding effects (e.g. gastro-intestinal problems, allergies, ASD severity, medication and diets), we excluded sleep problems from our phenotypic study.

S3.2 Imputation of the phenotypic dataset

While the SSC cohort consists of 2,644 simplex families, our analysis was restricted to 2,508 families, for which genetic data was available. We further excluded 10 patients due to missing phenotypic information on more than 48/63 (i.e. 75%) variables, leaving 2,498 patients for analysis (herein the SSC cohort refers to these 2,498 patients). Phenotypic variables were missing 5.81 ± 6.51% of values, which were imputed using the Multivariate Imputation by Chained Equations approach (MICE), implemented in R (van Buuren and

Groothuis-Oudshoorn, 2011). Briefly, under the assumption that missing data is *missing at random*, MICE applies logistic regression models to each variable. Accordingly, plausible values are predicted based on the observed data distribution and its relationship to the remaining variables from the dataset. Multiple imputed data sets are created and confidence estimates are generated across imputations using Ruben's rule. As the variance across imputations was low, we calculated the average imputed values and assigned presence and absence of phenotype to mean values ≥ 0.5 and < 0.5, respectively.

To ensure the imputation method was chosen appropriately, we compared phenotypic rates between the pre- and post-imputed datasets and found that most phenotypes occur at similar frequencies (**Table 4.1**). Exceptions include seizures, microcephaly and some psychiatric, autoimmune and mother autoimmune disorders, which were associated with higher incidence in the imputed dataset. Nevertheless, as prevalence rates remain relatively consistent with our literature search, we did not investigate the underlying causes.

SUPPLEMENTARY FIGURES

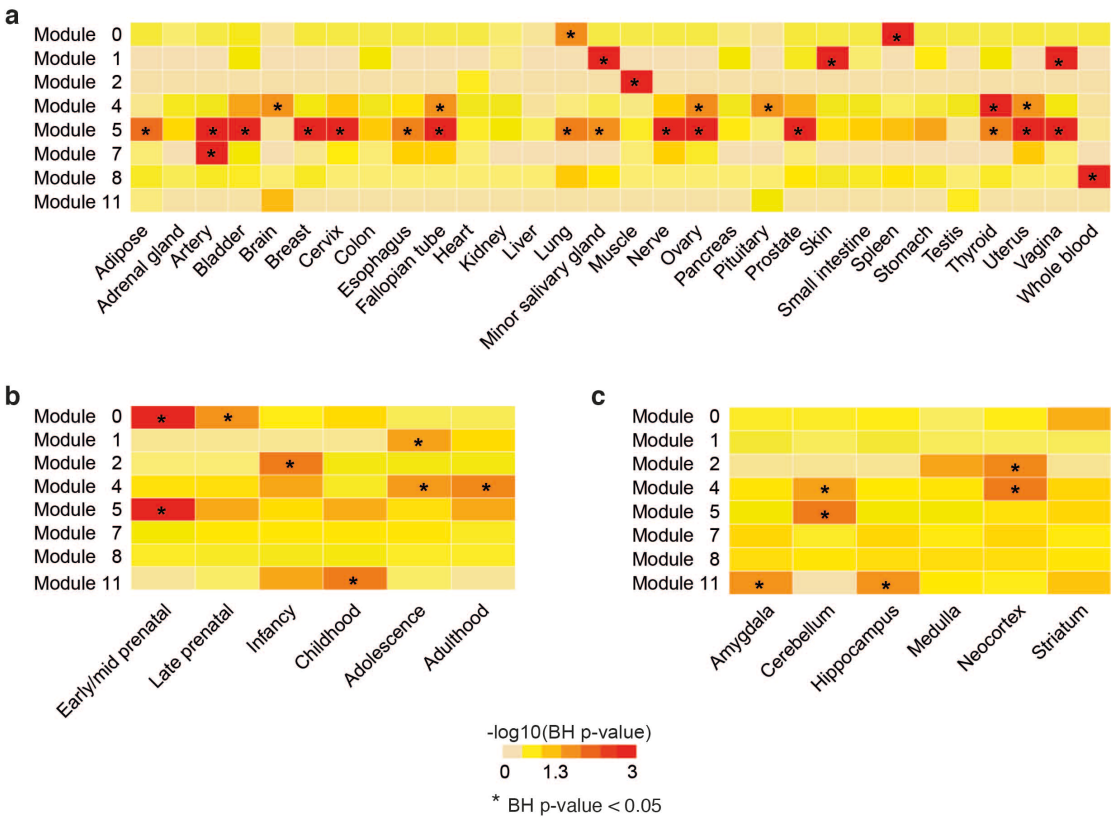


Fig S3.1 | Expression dynamics of module compared to genome genes. Expression patterns of module genes compared to random genome genes (matched for coding DNA sequence length to module genes) in (**a**) 30 body tissues (GTEx dataset) and in the brain at (**b**) six developmental points and in (**c**) six regions (Brainspan datasets). P-values were derived using a permutation-based approach (**section 2.7.2**) and Benjamini-Hochberg (BH) corrections for multiple testing were applied to control for the numbers of tissues, stages or regions tested.

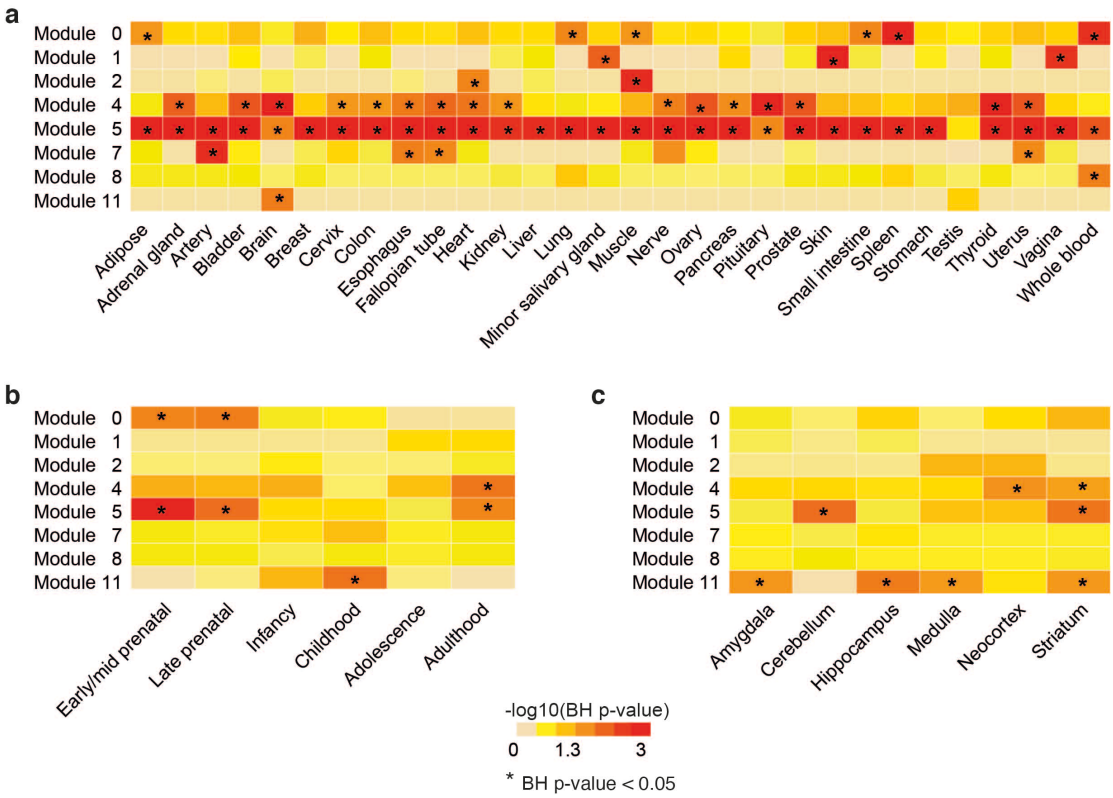


Fig S3.2 | Expression dynamics of module compared to CT genes. Expression patterns of module genes compared to control (CT) genes in (**a**) 30 body tissues (GTEx dataset) and in the brain at (**b**) six developmental points and in (**c**) six regions (Brainspan datasets). P-values were derived using right-tailed Mann-Whitney U tests and a Benjamini-Hochberg (BH) correction for multiple testing was applied to control for the numbers of tissues, stages or regions tested.

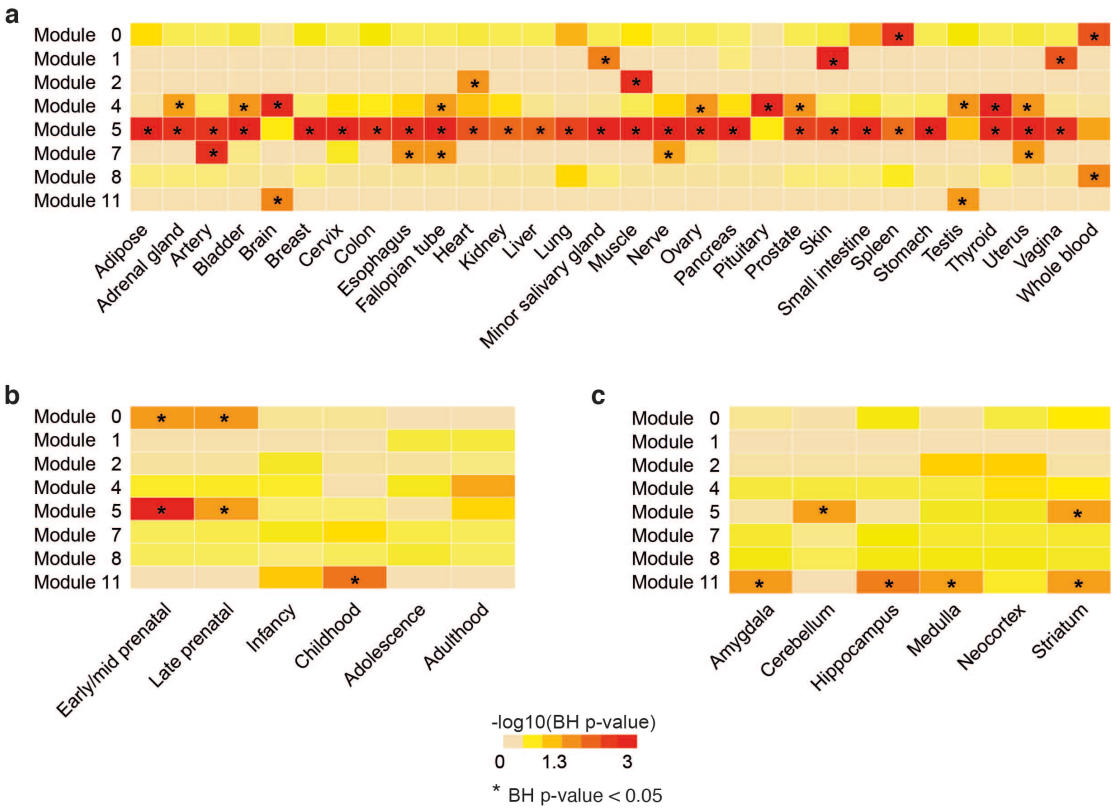


Fig S3.3 | Expression dynamics of module compared to co-expressed CT genes. Expression patterns of module genes compared to co-expressed control (CT) genes in (**a**) 30 body tissues (GTEx dataset) and in the brain at (**b**) six developmental points and in (**c**) six regions (Brainspan datasets). P-values were derived using right-tailed Mann-Whitney U tests and a Benjamini-Hochberg (BH) correction for multiple testing was applied to control for the numbers of tissues, stages or regions tested.

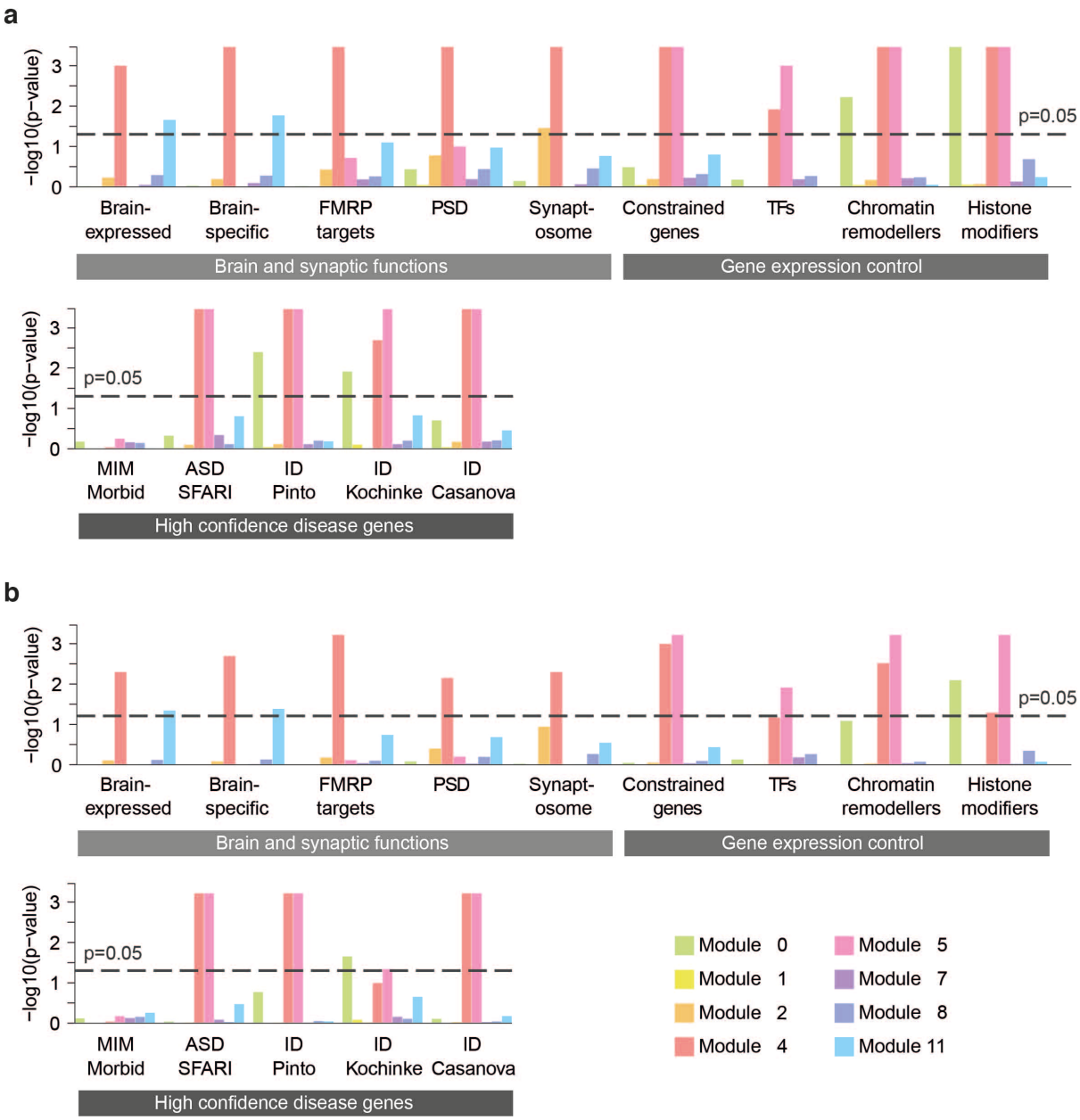


Fig S3.4 | Co-expression patterns of module genes and various gene sets. Using the GTEx gene co-expression network with r > 0.6, we assessed whether module genes were more closely expressed with various gene sets (**Table 2.1**) than (**a**) all control (CT) genes and (**b**) co-expressed CT genes (**section 2.5.3**).

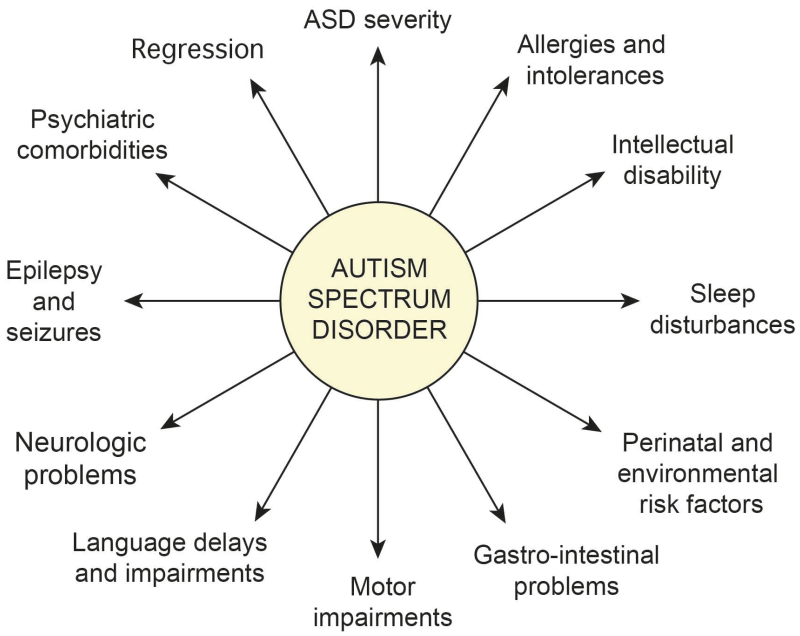


Fig S4.1 | Clinical correlates in ASD.

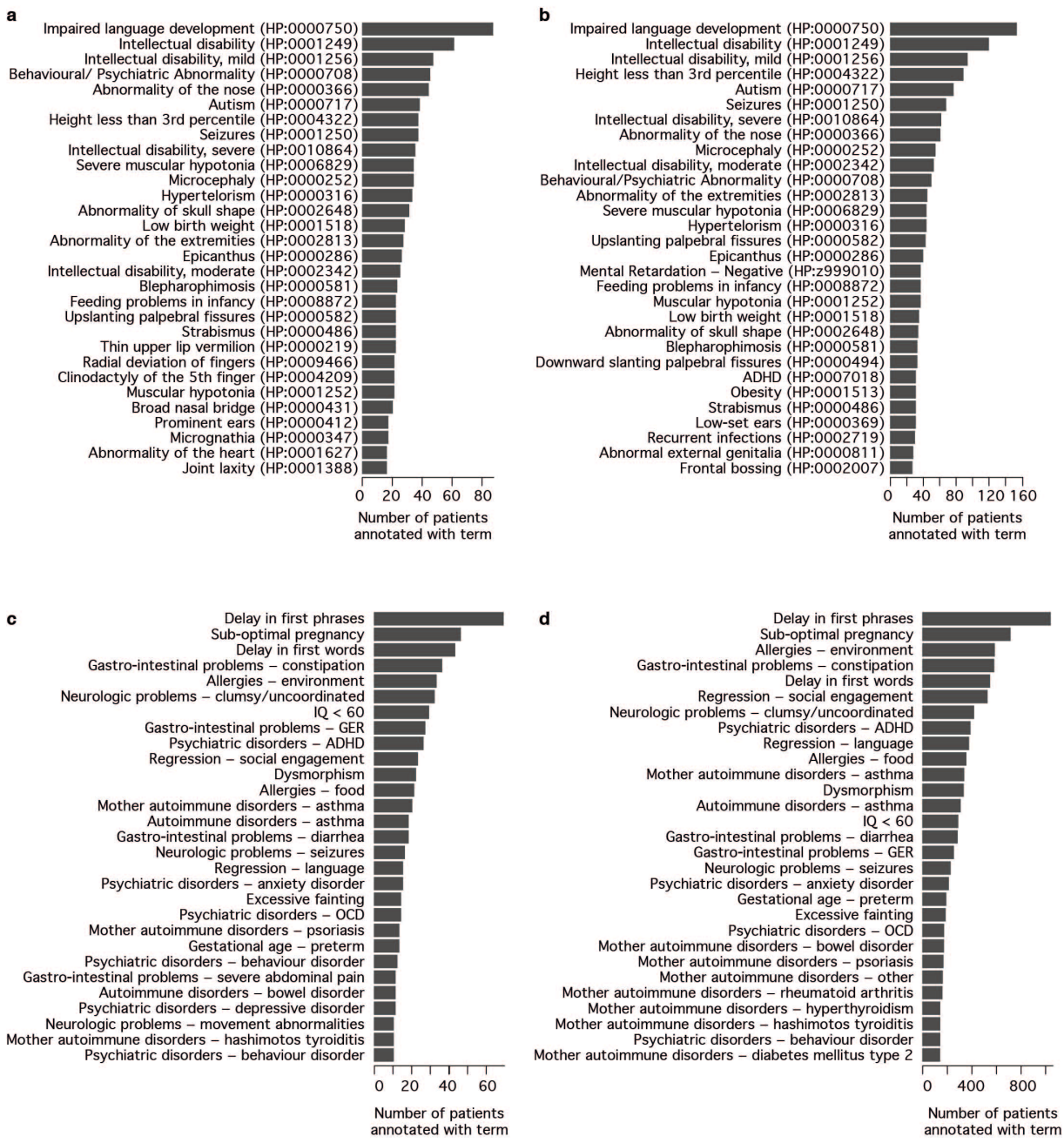


Fig S4.2 | Frequency of phenotypes in the Nijmegen and SSC cohorts. Frequency of the 30 most common non-imputed HPO terms displayed by Nijmegen patients with (**a**) *de novo* and (**b**) inherited CNVs and the most common phenotypes from SSC patients with (**c**) *de novo* and (**d**) inherited CNVs. GER refers to gastro-intestinal reflux disease.

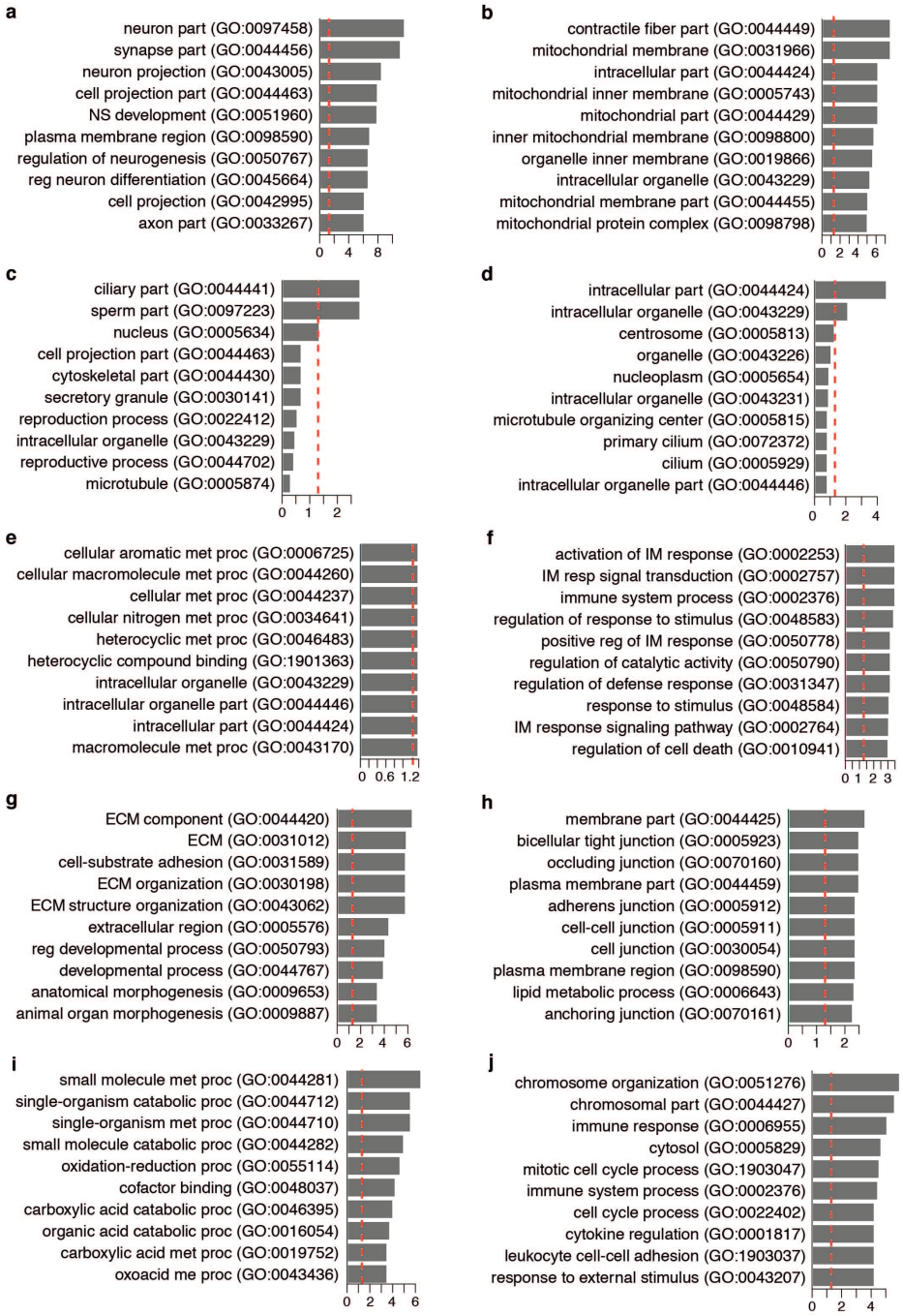


Fig S4.3 | GO enrichments of GTEx modules identified from genes mapping to *de novo* CNVs from Nijmegen patients. Top 10 GO terms most strongly enriched among genes from GTEx modules (**a**) 0, (**b**) 1, (**c**) 2, (**d**) 3, (**e**) 4, (**f**) 5, (**g**) 9, (**h**) 10, (**i**) 12 and (**j**) 13. The over-representation of GO terms was assessed using right-tailed hypergeometric tests, using all genes annotated in GO as a background. A correction for multiple testing was applied using the BH method to correct for number of terms tested. The x-axes represent -log(Benjamini-Hochberg-adjusted p-values). We note GO terms descriptions were shorted such that *NS, met*, *proc*, *IM*, *reg* and *cytokine regulation* stand for regulation of nervous system, metabolic, process, immune, regulation and regulation of cytokine, respectively.

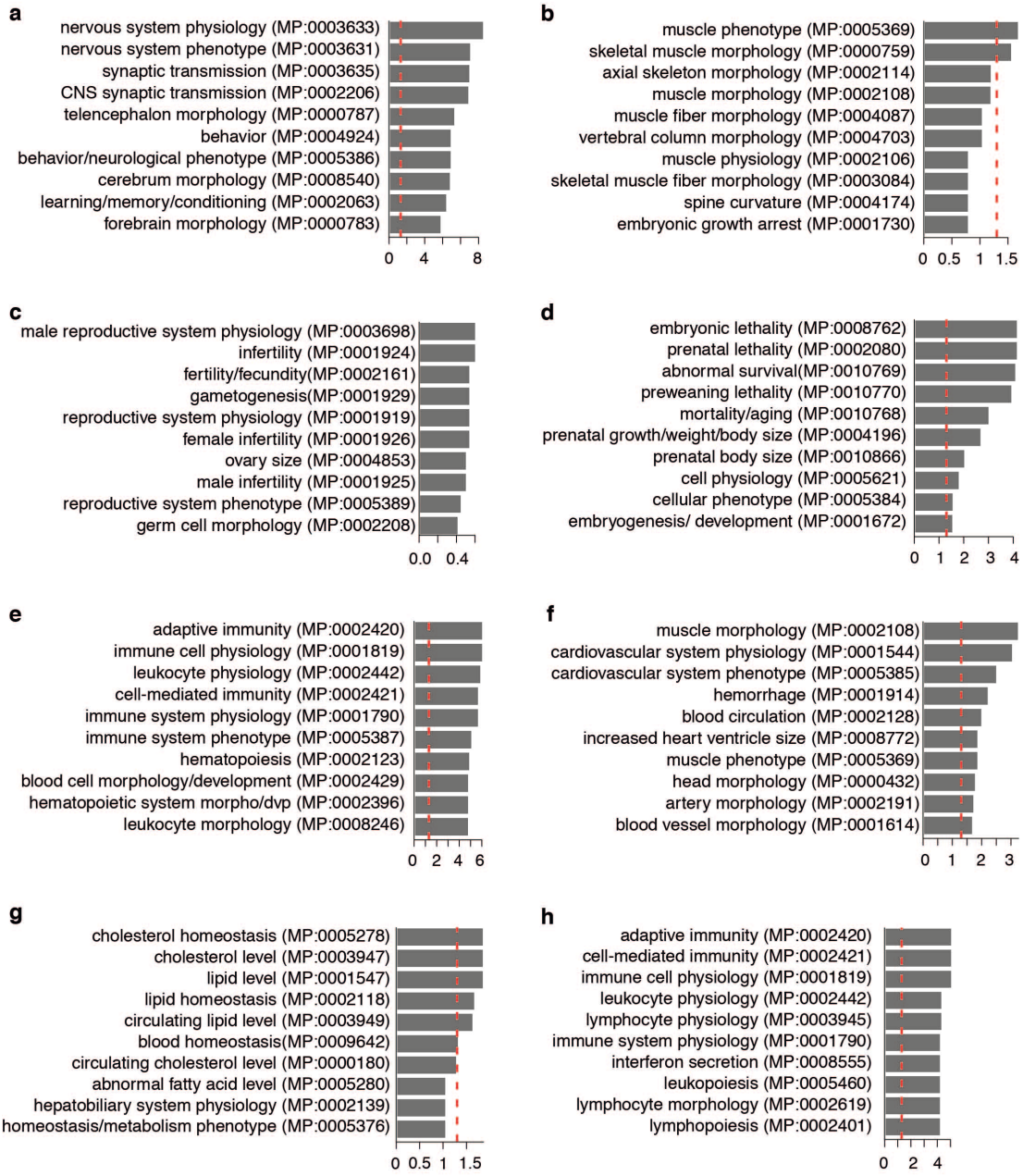


Fig S4.4 | MGI enrichments of GTEx modules identified from genes mapping to *de novo* CNVs from Nijmegen patients. Top 10 MGI terms most strongly enriched among genes from GTEx modules (**a**) 0, (**b**) 1, (**c**) 2, (**d**) 4, (**e**) 5, (**f**) 9, (**g**) 12 and (**h**) 13. Modules 3 and 10 were excluded, as no phenotypes reached nominal significance. The over-representation of mouse phenotypes (MPs) was assessed using right-tailed hypergeometric tests, using all genes annotated in MGI as a background. A correction for multiple testing was applied using the BH method to correct for number of terms tested. The x-axes represent -log(Benjamini-Hochberg-adjusted p-values). We note that MP term descriptions were shortened such that all terms refer to abnormalities of the respective phenotypes and *CNS* and *morpho/dvp* stand for central nervous system and morphology/development, respectively.

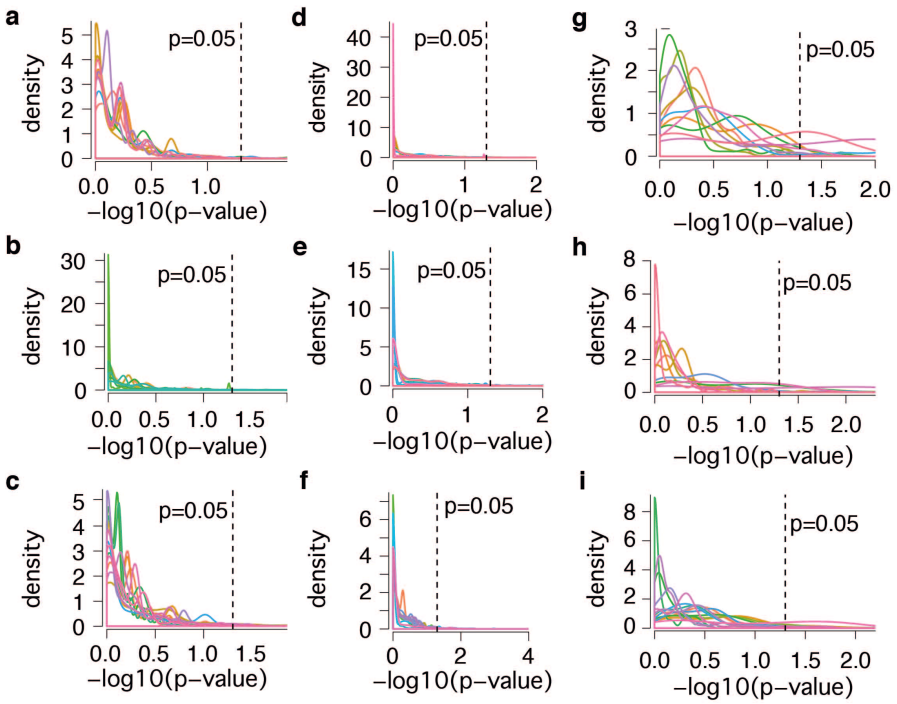


Fig S4.5 | Phenotype enrichments of patient modules. We compared the frequency of all phenotypes present in the examined cohort between module patients and remaining cohort patients using right-sided Fisher tests. Displayed here is the distribution of nominal p-values associated with the over-representation of each phenotype among patients from (**a, d, g**) GTEx, (**b, e, h**) Brainspan and (**c, f, i**) PLN (Phenotypic Linkage Network) modules. Modules were identified using genes mapping from Nijmegen patients with (**a, b, c**) *de novo* and (**d, e, f**) inherited CNVs and (**g, h, i**) SSC patients with inherited CNVs.

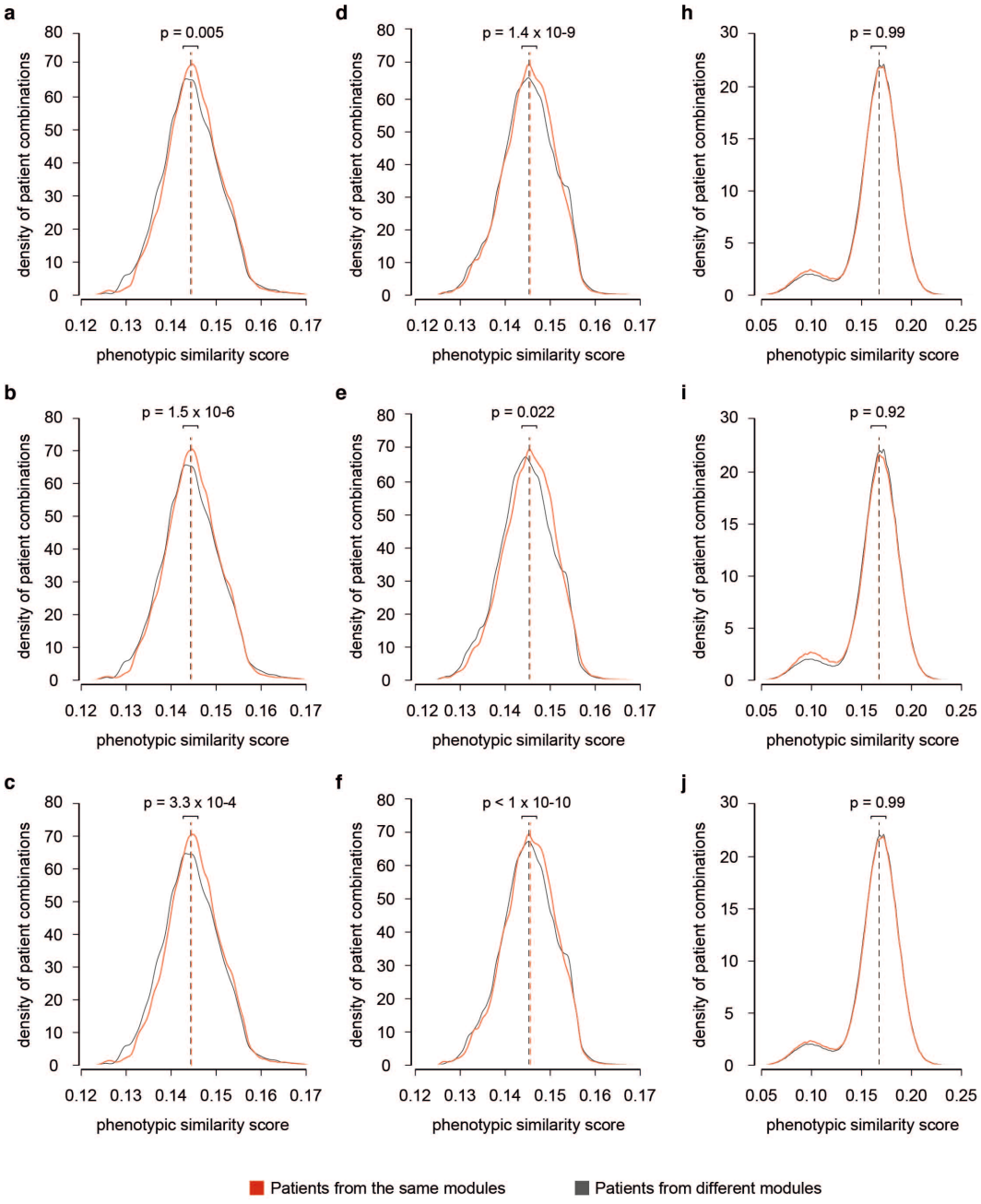


Fig S4.6 | Distribution of patient phenotypic similarity scores. The distribution of phenotypic similarity scores was compared between patients from the same (**a, d, g**) GTEx, (**b, e, h**) Brainspan and (**c, f, i**) PLN (Phenotypic Linkage Network) modules and those from different modules using right-tailed Mann-Whitney U tests. The individuals examined consist of Nijmegen patients with (**a, b, c**) *de novo* or (**d, e, f**) inherited CNVs < 5 mb, and (**g, h, i**) SSC patients with inherited CNVs < 5 mb.

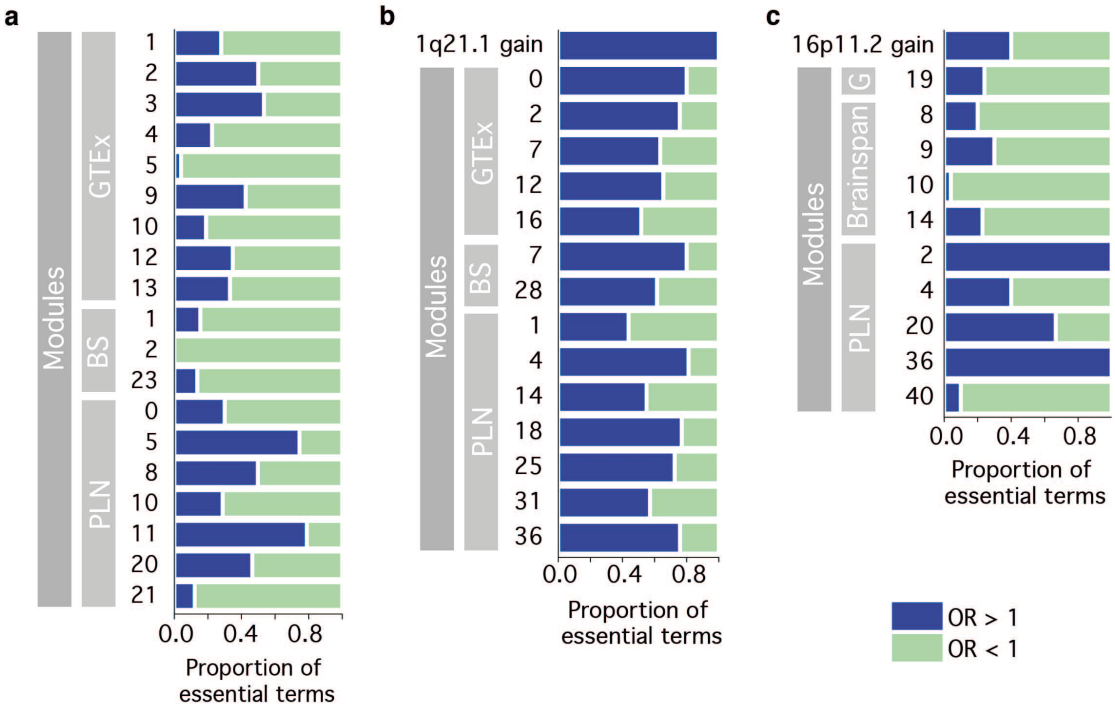


Fig S4.7 | Odds ratios of essential phenotypes. Proportion of essential phenotypes, which are more (odds ratio, OR > 1) or less (OR < 1) common among patients from phenotypically coherent groups compared to remaining patients from the cohort. Over- and under-representations were assessed using two-sided Fisher tests (they are insignificant) and are displayed for Nijmegen patients with (**a**) *de novo* and (**b**) inherited CNVs and (**c**) SSC patients with inherited CNVs.

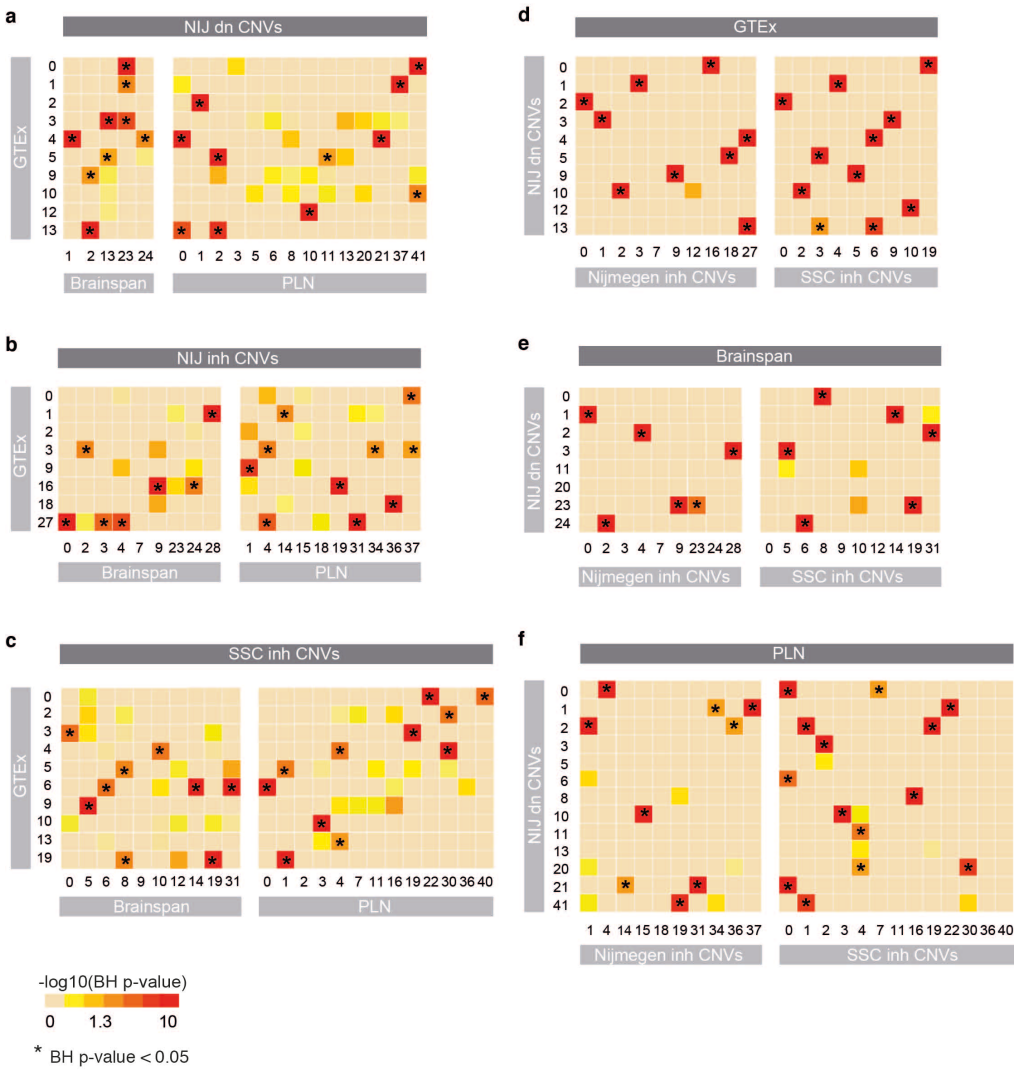


Fig S4.8 | Module overlaps across networks and cohorts. (**a to c**) Gene overlap between GTEx, Brainspan and PLN (Phenotypic Linkage Network) gene modules uncovered using Nijmegen (NIJ) patients with (**a**) *de novo* (*dn*) or (**b**) inherited (inh) CNVs and (**c**) SSC patients with inherited CNVs. (**d, e, f**) Functional overlap between gene modules uncovered using Nijmegen patients with *dn* or inh CNVs and those identified using SSC patients with inh CNVs using (**d**) GTEx, (**e**) Brainspan and (**f**) PLN networks. P-values were derived by evaluating the over-representation of shared nominally enriched GO terms between pairs of modules, given the total number of distinct nominal enrichments. Nominal GO enrichments refer to terms over-represented among module genes at $p < 0.05$, based on right-tailed hypergeometric tests using all genes annotated in GO as a background. (**a to f**) P-values were derived using right-tailed Fisher tests and a Benjamini-Hochberg (BH) correction was applied to control for the number of pairwise comparisons performed.

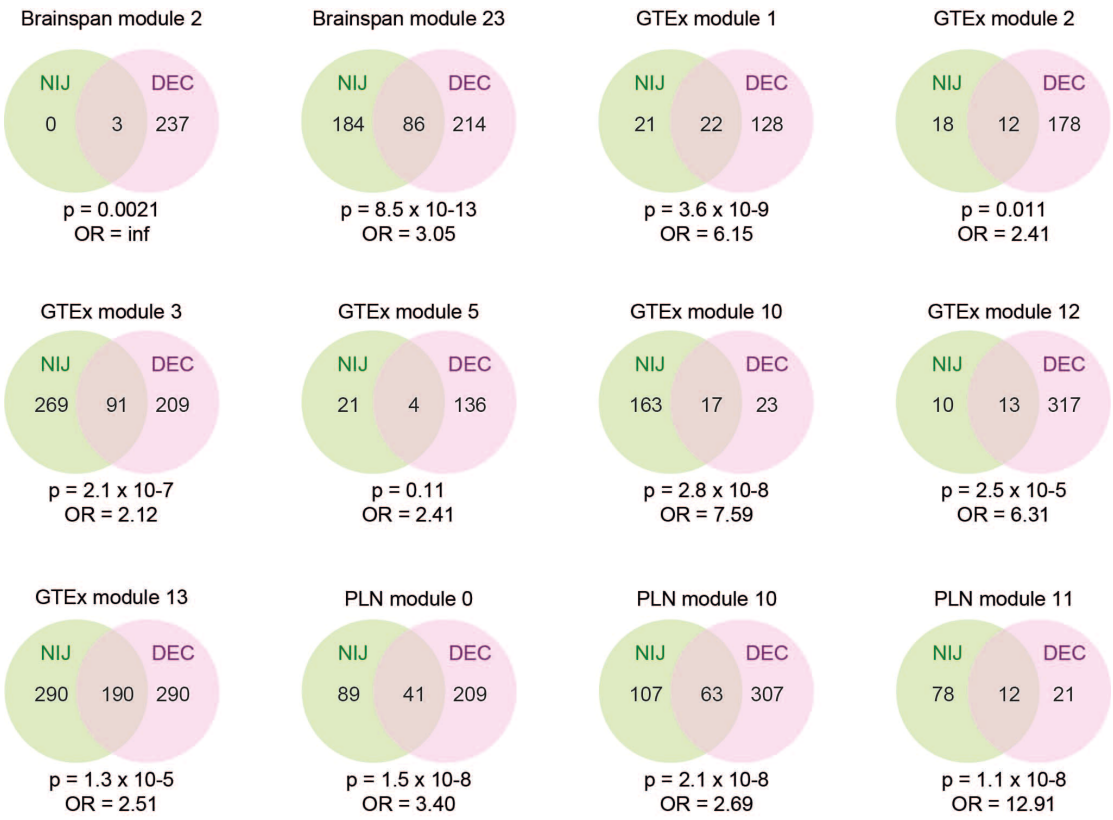


Fig S4.9 | Overlap of essential HPO term between Nijmegen and DECIPHER modules. For each phenotypically coherent module uncovered using Nijmegen (NIJ; pink) patients with *de novo* CNVs and validated using DECIPHER patients (DEC; green), we evaluated the overlap between the corresponding essential HPO (Human Phnotype Ontology) terms using right-tailed Fisher tests.

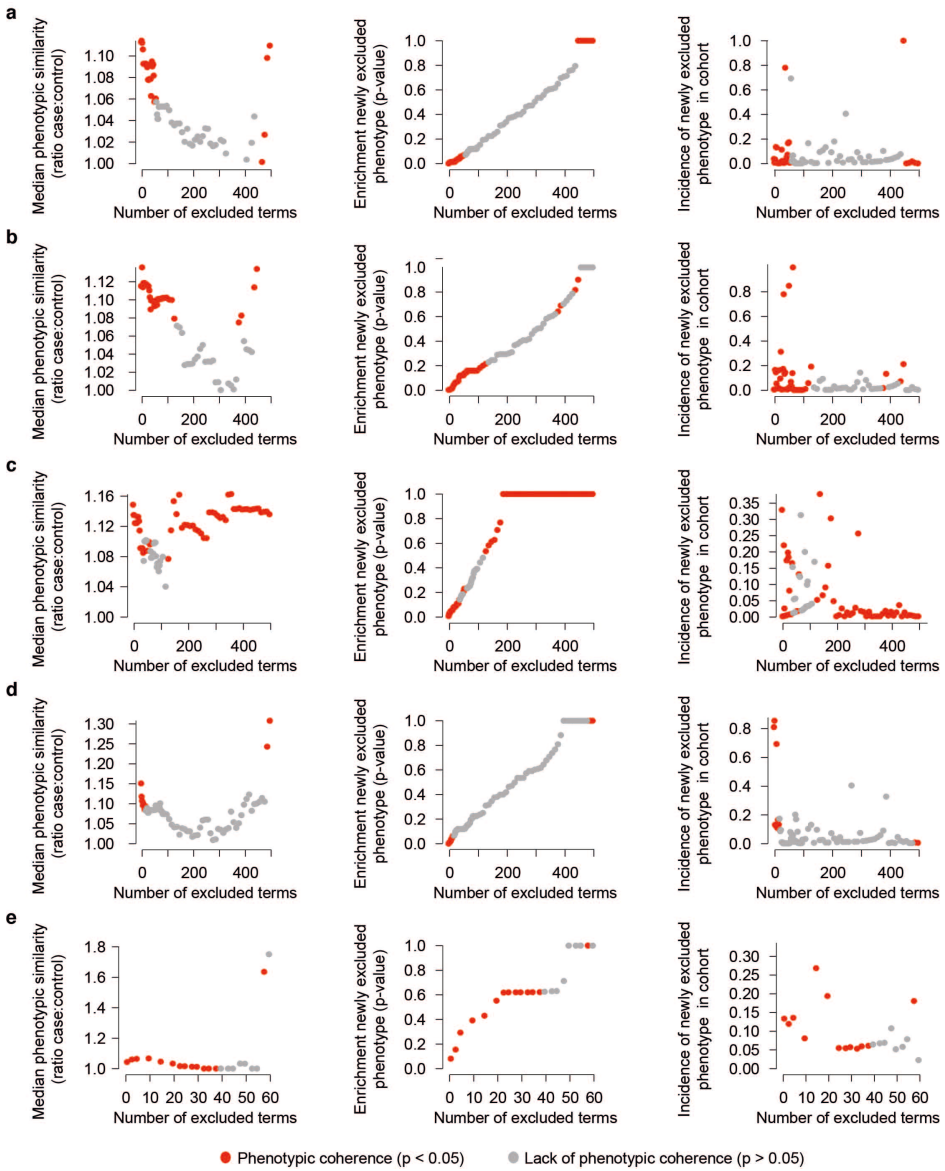


Fig S4.10 | Power and essential phenotypes for module of Nijmegen and SSC patients with inherited CNVs. (**Left panels**) Following the removal of large numbers of phenotypes (beyond essential terms), strong phenotypic coherence is observed among Nijmegen patients from GTEx modules (**a**) 0 and (**b**) 2, (**c**) Brainspan module 7 and (**d**) PLN (Phenotypic Linkage Network) module 14, and (**e**) SSC patients from Brainspan module 9. (**Right panels**) The phenotypes left to evaluate phenotypic similarity are present in an extremely (**a**) large or (**a to e**) small fraction of patients, suggesting the similarity signal derives from their segregation outside or within the module, respectively. (**Middle panels**) As these phenotypes are weakly associated with patients from the examined modules, we conclude this segregation is random and the similarity signal is artificially inflated. The y-axes represent (**left panels**) the ratio between the median phenotypic similarity of module patients and non-module patients, (**middle panels**) the p-values associated with the over- or under-representation of the newly excluded phenotypes, based on two-sided Fisher tests and (**right panels**) the fraction of all patients annotated with it. Red and grey data points refer to significant and insignificant phenotypic coherence, respectively. (**Left panels**) Of note, data points associated with ratios < 1 are not shown in the figures.

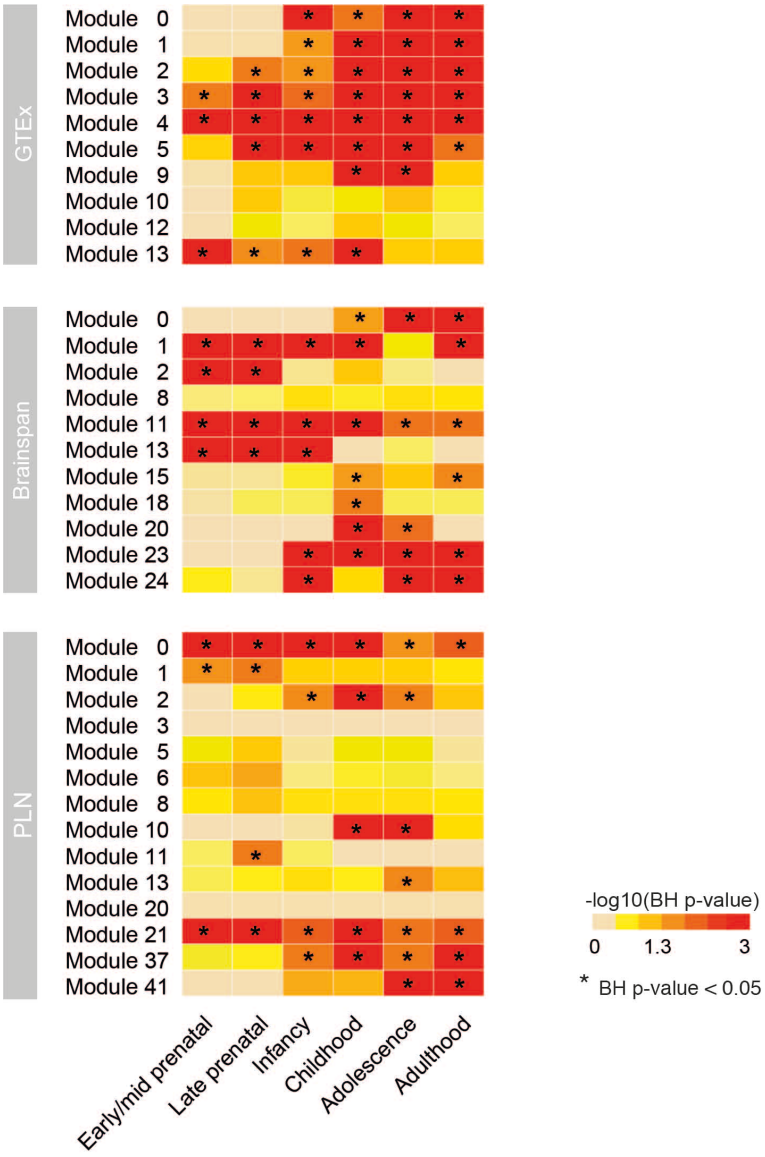


Fig S4.11 | Temporal brain expression patterns of genes from modules identified using *de novo* CNVs from Nijmegen patients. Brain expression dynamics of module genes, identified using *de novo* CNVs from Nijmegen patients, compared to random genome genes (matched for coding DNA sequence length to module genes) in the brain at six developmental points (Brainspan dataset). P-values were derived using a permutation-based approach (**section 2.7.2**) and Benjamini-Hochberg (BH) corrections for multiple testing were applied to control for the number of developmental stages tested.

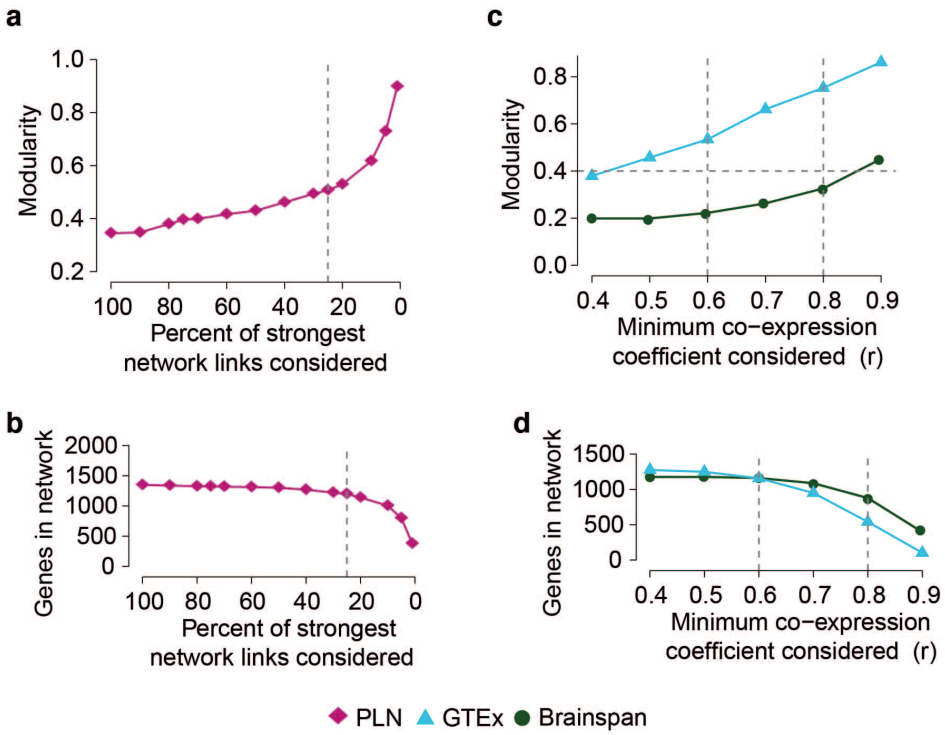


Fig S5.1 | Choice of gene network thresholds. Modularity of networks consisting of genes affected by *de novo* likely gene-disrupting (LGD) variants in SSC patients, at various thresholds. We restricted PLN (Phenotypic Linkage Network; pink diamond) sub-networks to the top k = [100, 90, 80 .. 10, 1] % strongest links and GTEx (blue triangle) and Brainspan (green circle) co-expression sub-networks to links with r > [0.4, 0.5 .. 0.9]. We examined (**b, d**) the number of genes present in each network and (**a, c**) the modularity associated with their partitioning using the Louvain algorithm. To maximise modularity and the number of genes present in the network, we considered (**a, b**) the top 25% strongest PLN links and (**c, d**) GTEx and Brainspan co-expression networks with r > 0.6 and r > 0.8, respectively, for further analysis.

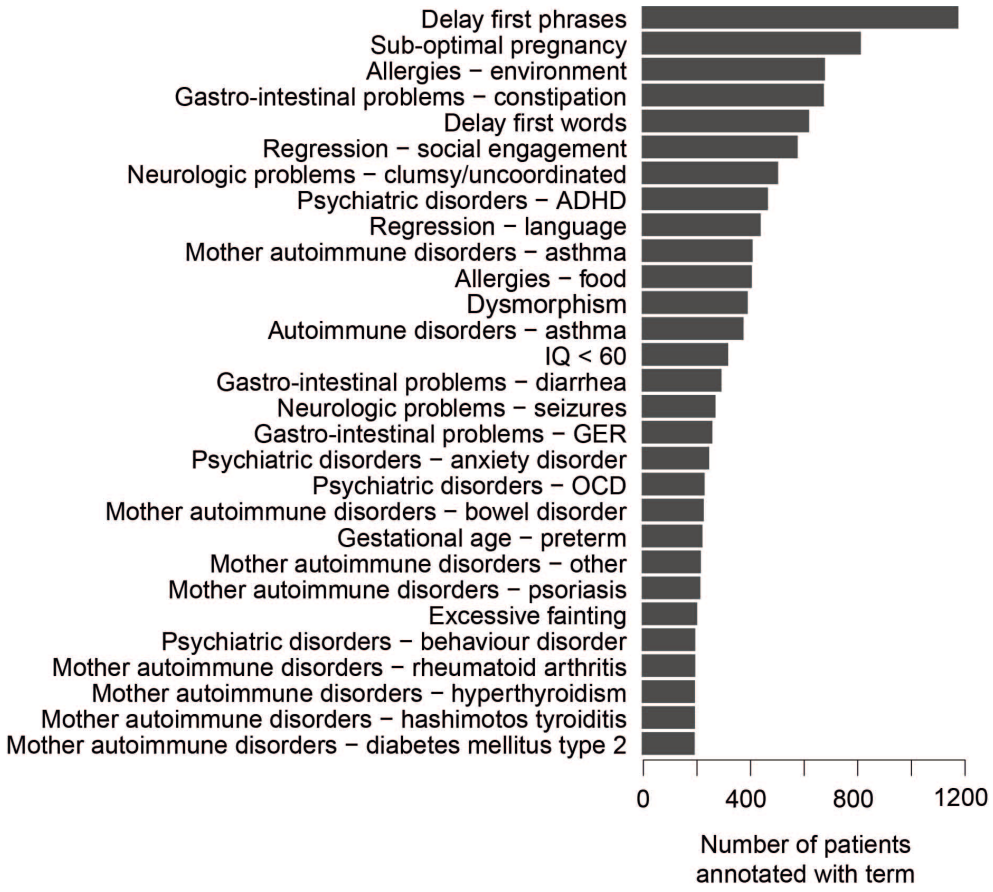


Fig S5.2 | Frequency of the 30 most common clinical phenotypes among sequenced SSC patients (n = 2,498). GER refers to gastro-intestinal reflux disease.

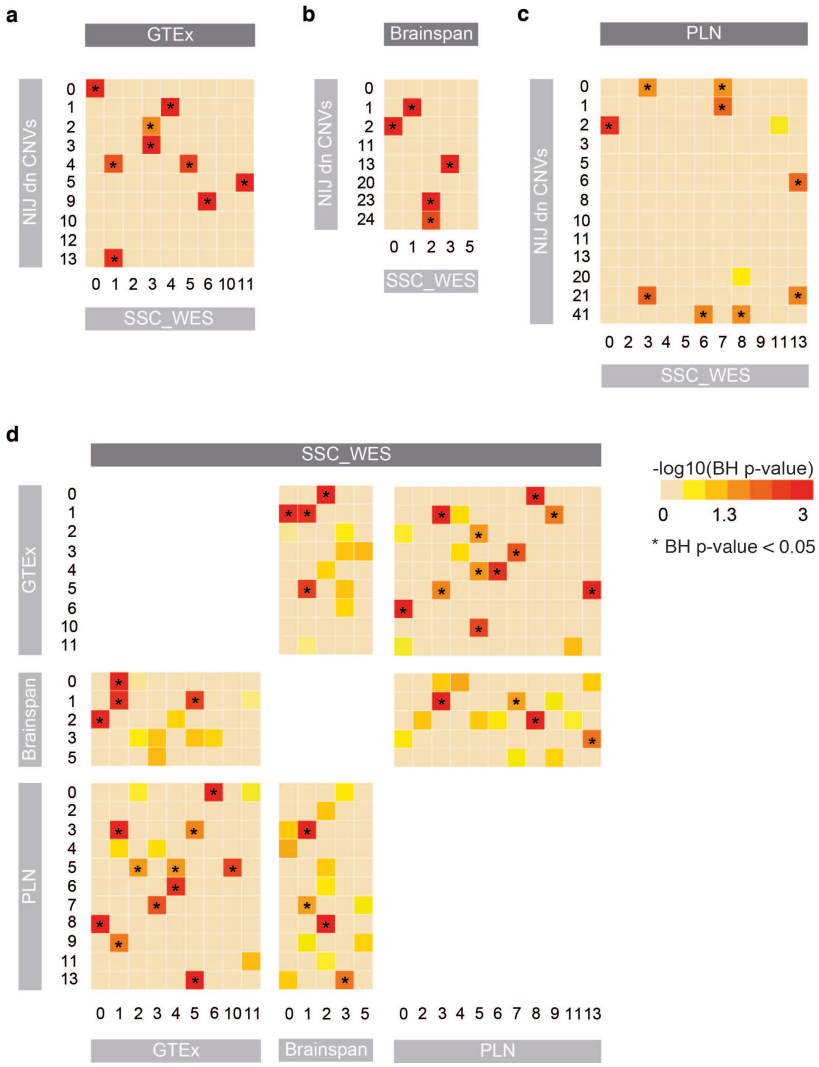


Fig S5.3 | Module overlaps across networks and studies. (**a to c**) Functional overlap between modules from this study and those uncovered using Nijmegen patients with *de novo* (*dn*) CNVs (**chapter 4**) using (**a**) GTEx, (**b**) Brainspan and (**c**) PLN (Phenotypic Linkage Network) networks. P-values were derived by evaluating the over-representation of shared nominally enriched GO terms between pairs of modules, given the total number of distinct nominal enrichments. Nominal GO enrichments refer to terms over-represented among module genes at $p < 0.05$, based on right-tailed hypergeometric tests using all genes annotated in GO as a background. (**d**) Gene overlap between GTEx, Brainspan and PLN modules from this study was assessed by examining the number of common genes between pairwise module combinations, given the total number of genes affected by *dn* likely gene-disrupting variants in SSC patients. (**a to d**) P-values were derived using right-tailed Fisher tests and a Benjamini-Hochberg (BH) correction was applied to control for the number of pairwise comparisons performed. SSC_WES, subset of sequenced SSC patients (n = 2,498).

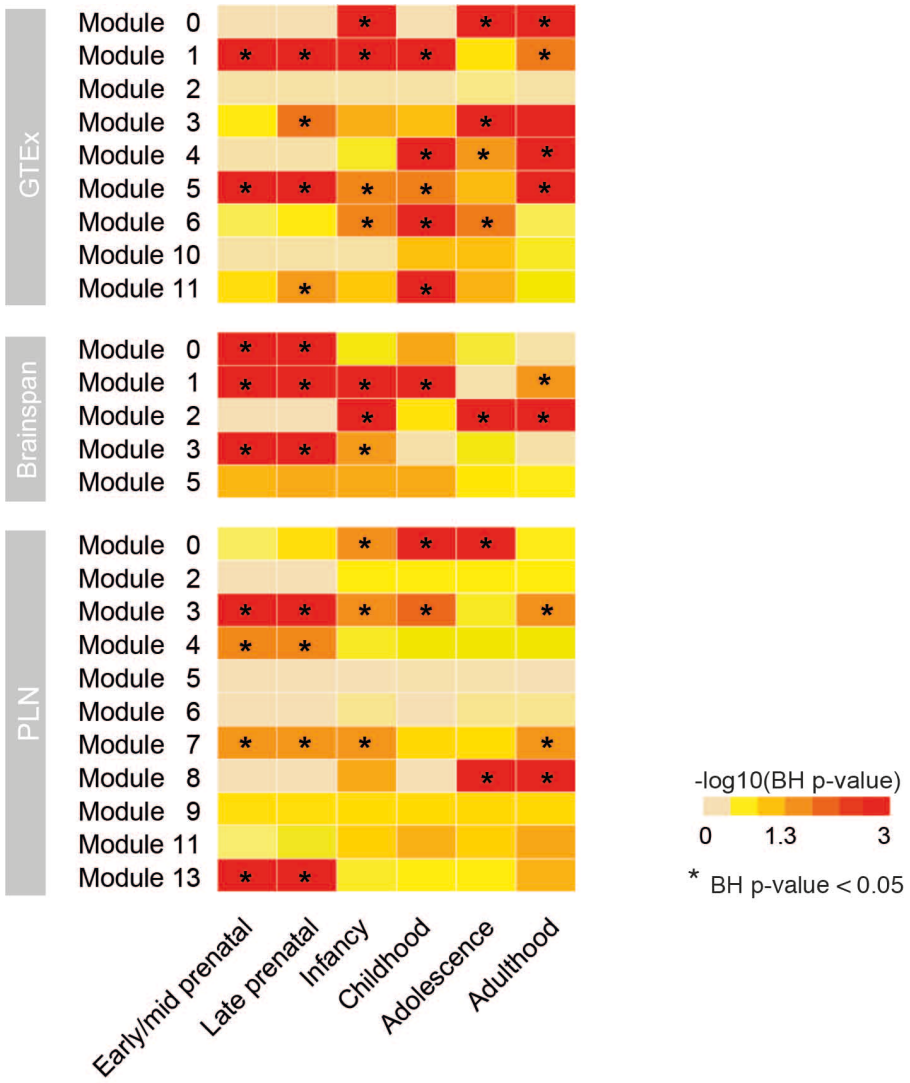


Fig S5.4 | Temporal brain expression patterns module genes. Brain expression dynamics of genes from GTEx, Brainspan and PLN (Phenotypic Linkage Network) modules, identified using SSC patients with *de novo* likely gene-disrupting variants. Expression was compared between module genes and random genome genes (matched for CDS to module genes) in the brain at six developmental points (Brainspan dataset). P-values were derived using a permutation-based approach (**section 2.7.2**) and Benjamini-Hochberg (BH) corrections for multiple testing were applied to control for the number of developmental stages tested.

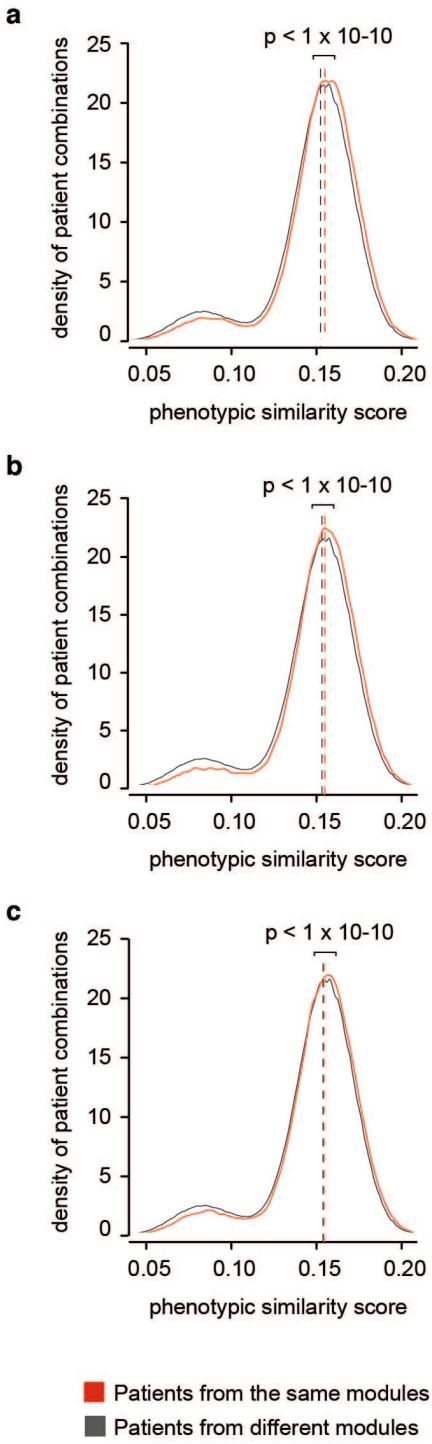


Fig S5.5 | Distribution of patient phenotypic similarity scores. Comparison of phenotypic similarity scores between SSC patients from the same (**a**) GTEx, (**b**) Brainspan and (**c**) PLN (Phenotypic Linkage Network) modules and those from different modules using right-tailed Mann-Whitney U tests.

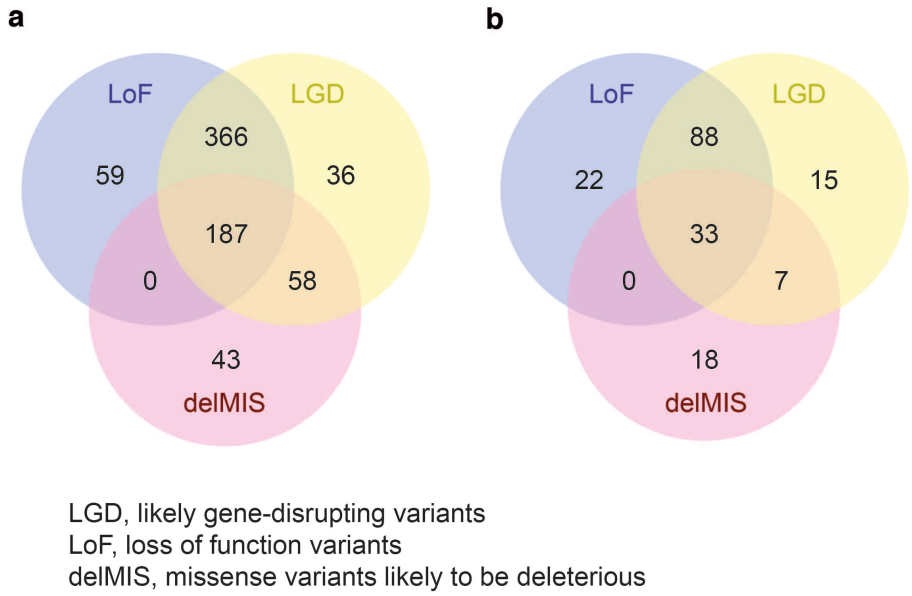


Fig S5.6 | Genes affected by LoF and delMIS variants affect common processes. Venn diagram of nominally enriched (**a**) GO and (**b**) MGI terms among genes affected by *de novo* likely gene-disrupting LGD (yellow), LoF (*de novo* loss of function defined as nonsense, splice-site or frameshift variants) (blue) and delMIS (*de novo* missense variants with CADD scores ≥ 15) (pink) SNVs in SSC patients. Over-representations were tested using right-tailed Fisher tests.

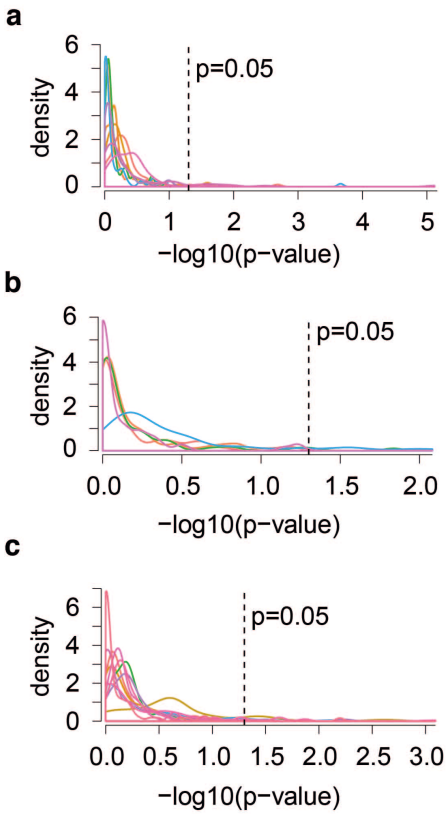


Fig S5.7 | Phenotype enrichments across patient modules. We compared the frequency of all phenotypes present in the SSC cohort between patients from each module and the remaining patients from the cohort using right-sided Fisher tests. Displayed here are the distributions of nominal p-values associated with the over-representation of each phenotype among patients from (**a**) GTEx, (**b**) Brainspan and (**c**) PLN (Phenotypic Linkage Network) modules.

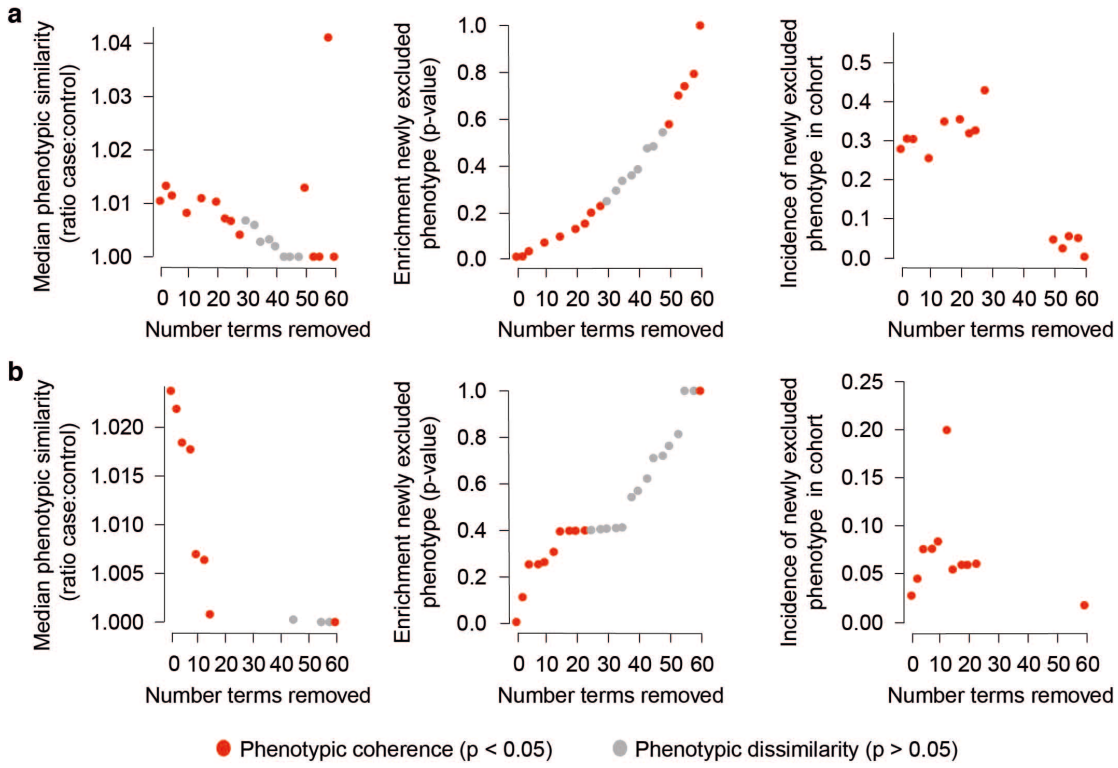


Fig S5.8 | Statistical power and essential phenotypes. (**Left panels**) While 30 and 25 essential phenotypes were uncovered for GTEx module 11 and (**b**) PLN (Phenotypic Linkage Network) module 9, strong phenotypic coherence is observed among module patients, following the removal of ≥ 48 or 60 phenotypic terms, most strongly associated with the module, respectively. (**Right panels**) The few phenotypes left to evaluate phenotypic similarity are present in a small fraction of patients, suggesting the similarity signal derives from their segregation within or outside the module. (**Middle panels**) As these phenotypes are not (or only merely) associated with the comprised patients (large p-values), we assume this segregation is random and the similarity signal is artificially inflated. The y-axes represent (**left panels**) the ratio between the median phenotypic similarity of module patients and non-module patients (with ≥ 1 *de novo* likely gene-disrupting (LGD) SNV(s)), (**middle panels**) the p-values associated with the over- or under-representation of the newly excluded phenotypes, based on two-sided Fisher tests and (**right panels**) the fraction of all patients with ≥ 1 *de novo* LGD SNV(s) annotated with the phenotype. Red and grey data points refer to significant and insignificant phenotypic coherence, respectively.

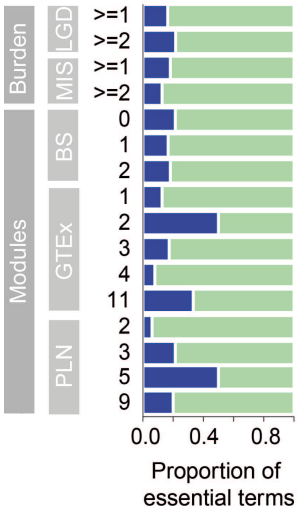


Fig S5.9 | Odds ratios of essential phenotypes. Proportion of essential phenotypes, which are more (odds ratio, OR > 1) or less (OR < 1) common among patients from phenotypically coherent groups, compared to remaining patients from the cohort. Over- and under-representations were assessed using two-sided Fisher tests (they are insignificant).

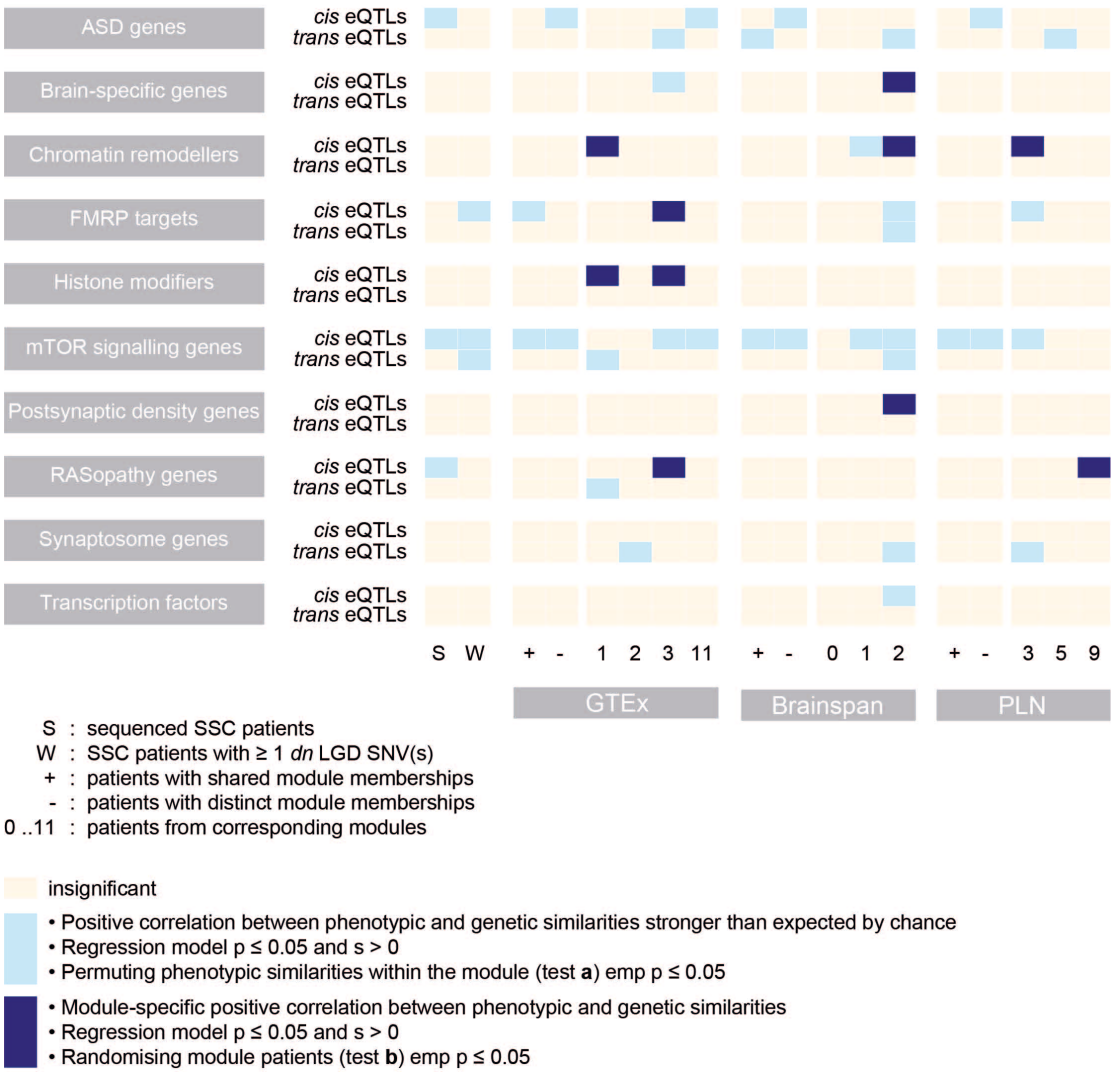
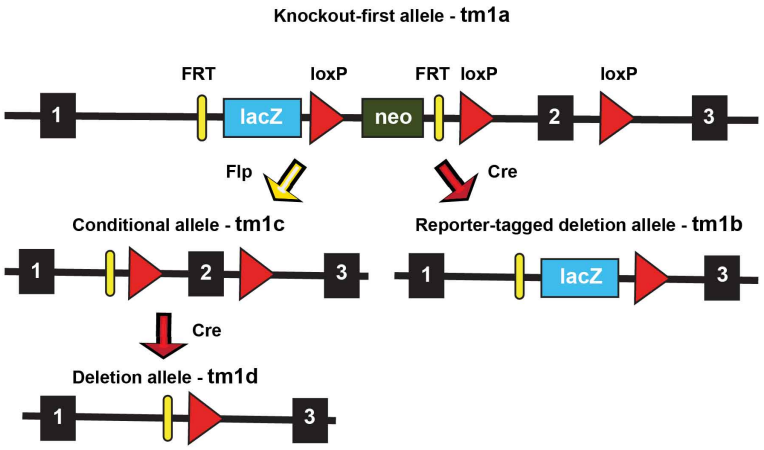


Fig S5.10 | Correlation between phenotypic and genetic similarities across eQTLs affecting NDD-associated genes. The relationship between phenotypic similarity and genetic similarity was examined using linear regression across local (*cis*) and distal (*trans*) eQTLs influencing the expression of genes from NDD-associated functional classes (**section 5.3.1; Table 5.3**). This relationship was tested across sequenced patients (S), patients with *de novo* (*dn*) likely gene-disrupting (LGD) SNVs (W), patients with shared module memberships (+), patients that never occur within the same module (-) and module patients (depicted using the module number). Beige depicts a lack of correlation between genetic and phenotypic similarities. Light blue depicts a positive correlation between phenotypic and genetic similarities, which is stronger than expected by chance (i.e. common variants (CVs) act as genetic modifiers of rare variants). Dark blue depicts a significant positive relationship between phenotypic and genetic similarities, which is specific to the module examined (i.e. CVs specifically modulate the phenotypes associated with the disruption of module genes). The way each level of significance was derived is described in **section 5.3.8** (light blue is associated with test **a** and dark blue with test **b**).

PROMOTER-DRIVEN CASSETTE



PROMOTER-LESS CASSETTE

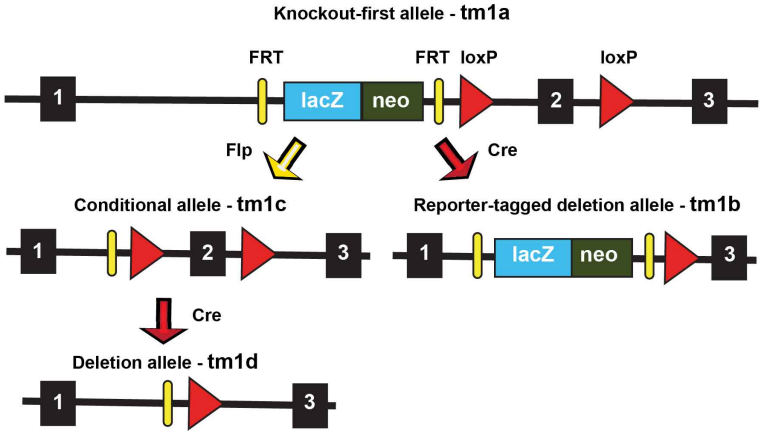


Fig S6.1 | Strategy used by the IMPC to generate knockout mice. (**Top panel**) Promoter-driven targeting cassettes or (**bottom panel**) promoter-less cassettes were used for the generation of the targeting mutation in embryonic stem cells. In both cases, the targeting mutation strategy relies on the identification of a common exon for all transcript variants that, when deleted, creates a frame-shift mutation. The *knockout-first allele* (tm1a) is flexible and can produce reporter knockouts (tm1b), conditional knockouts (tm1c), and null alleles (tm1d) following exposure to site-specific recombinases Cre and Flp. In the alleles name (tm1a, tm1b, tm1c and tm1d), the number *1* refers to the number of trials to produce a mutant.

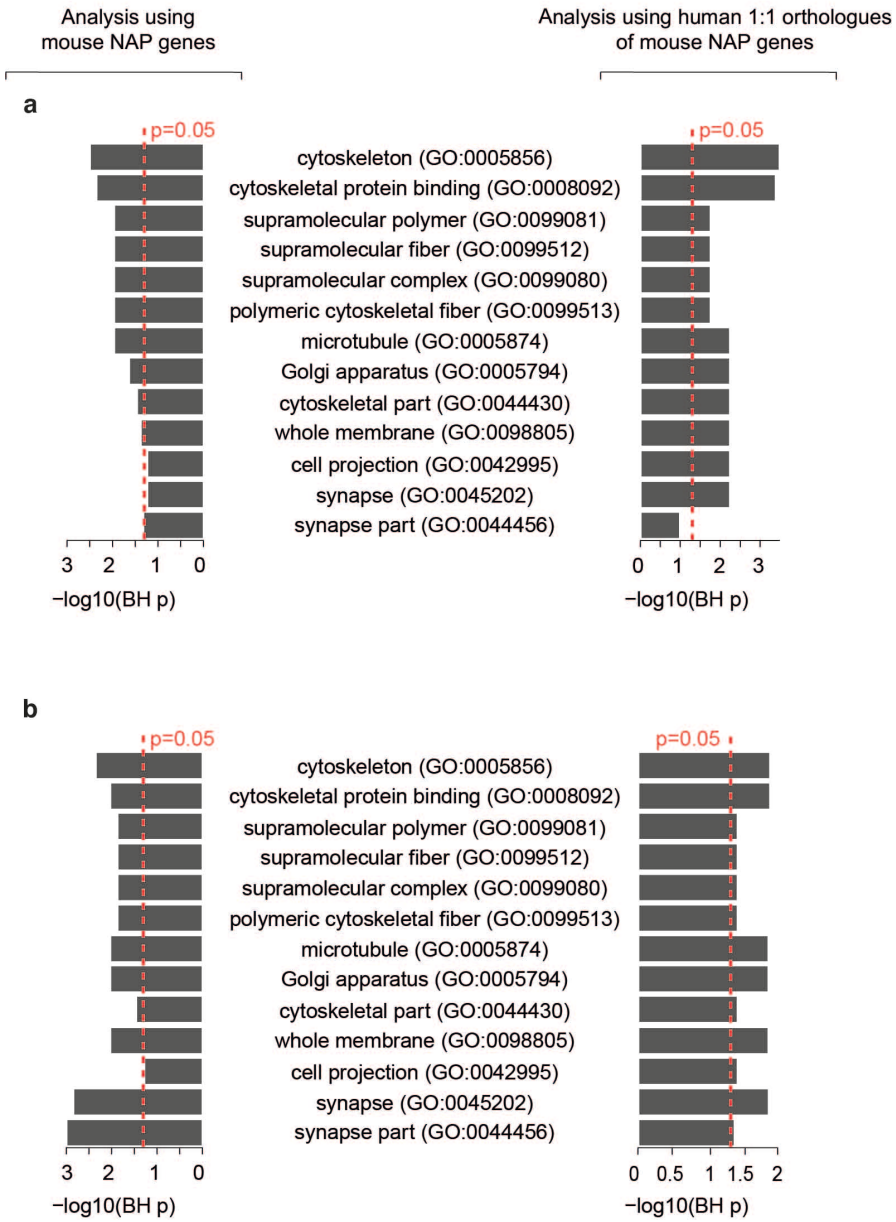


Fig S6.2 | GO enrichment analysis of NAP genes determined at 5% FDR. (**Left panel**) Mouse NAP genes determined at 5% FDR recapitulate the synaptic and neuronal functions observed at the 10% FDR threshold, when using (**a**) the whole genome or (**b**) non-NAP genes as backgrounds during GO enrichment analysis. (**Right panel**) Over-represented terms persist when performing the analysis using human 1:1 orthologues. The x-axes depict the p-values adjusted for the number of GO categories tested using the Benjamini-Hochberg method (BH p). NAP genes, neuroanatomical phenotype genes, defined as those associated with neuroanatomical deviations in knockout compared to wildtype mice (significant at 5% FDR).

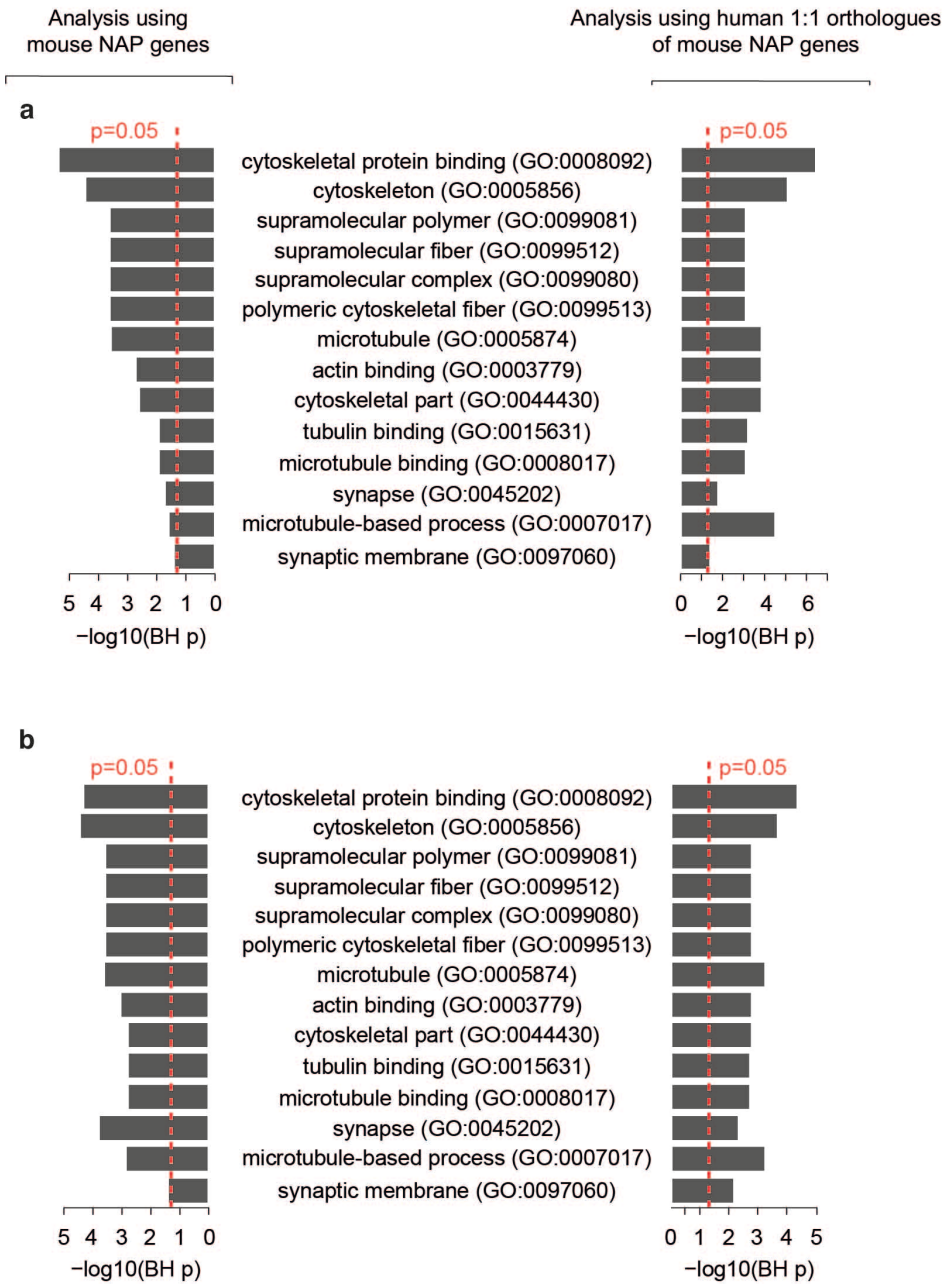


Fig S6.3 | GO enrichment analysis of NAP genes determined at 1% FDR. (**Left panel**) Mouse NAP genes determined at 1% FDR recapitulate the synaptic and neuronal functions observed at the 10% FDR threshold, when using (**a**) the whole genome or (**b**) non-NAP genes as backgrounds during GO enrichment analysis. (**Right panel**) Over-represented terms persist when performing the analysis using human 1:1 orthologues. The x-axes depict the p-values adjusted for the number of GO categories tested using the Benjamini-Hochberg method (BH p). NAP genes, neuroanatomical phenotype genes, defined as those associated with neuroanatomical deviations in knockout compared to wildtype mice (significant at 1% FDR).

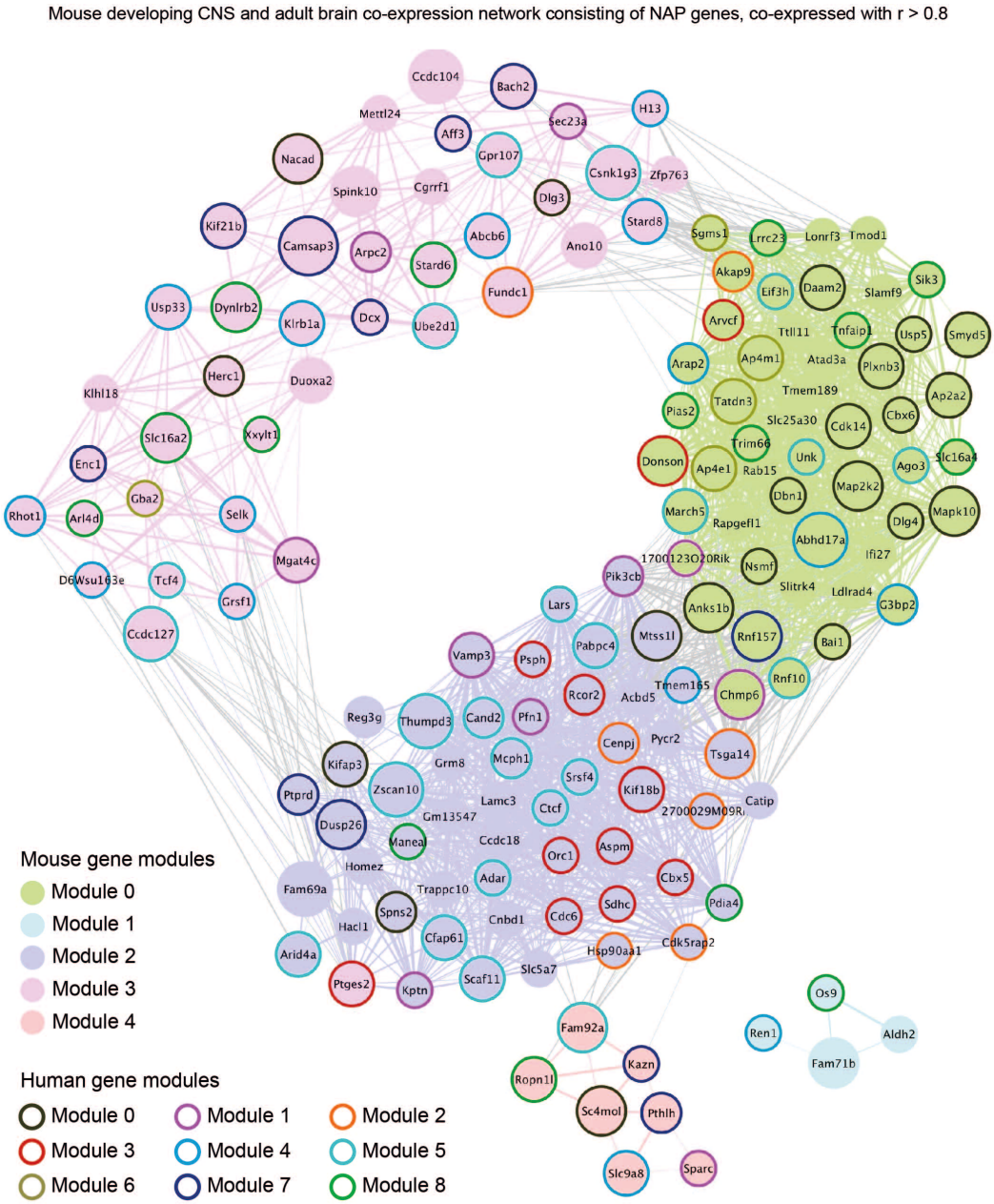


Fig S6.4 | Developing CNS and adult brain gene co-expression network of NAP genes. The mouse developing CNS and adult brain gene co-expression sub-network (derived from mouse developing CNS and adult brain expression data) consists of 2,115 links between 155 NAP genes. Shared node and border colours depict common memberships to mouse and human modules, respectively; the thickness of the edge is proportional to the co-expression coefficient of the respective genes. Human modules are discussed in **section 6.4.4**. NAP genes, neuroanatomical phenotype genes, defined as those associated with neuroanatomical deviations in knockout compared to wildtype mice (significant at 10% FDR).

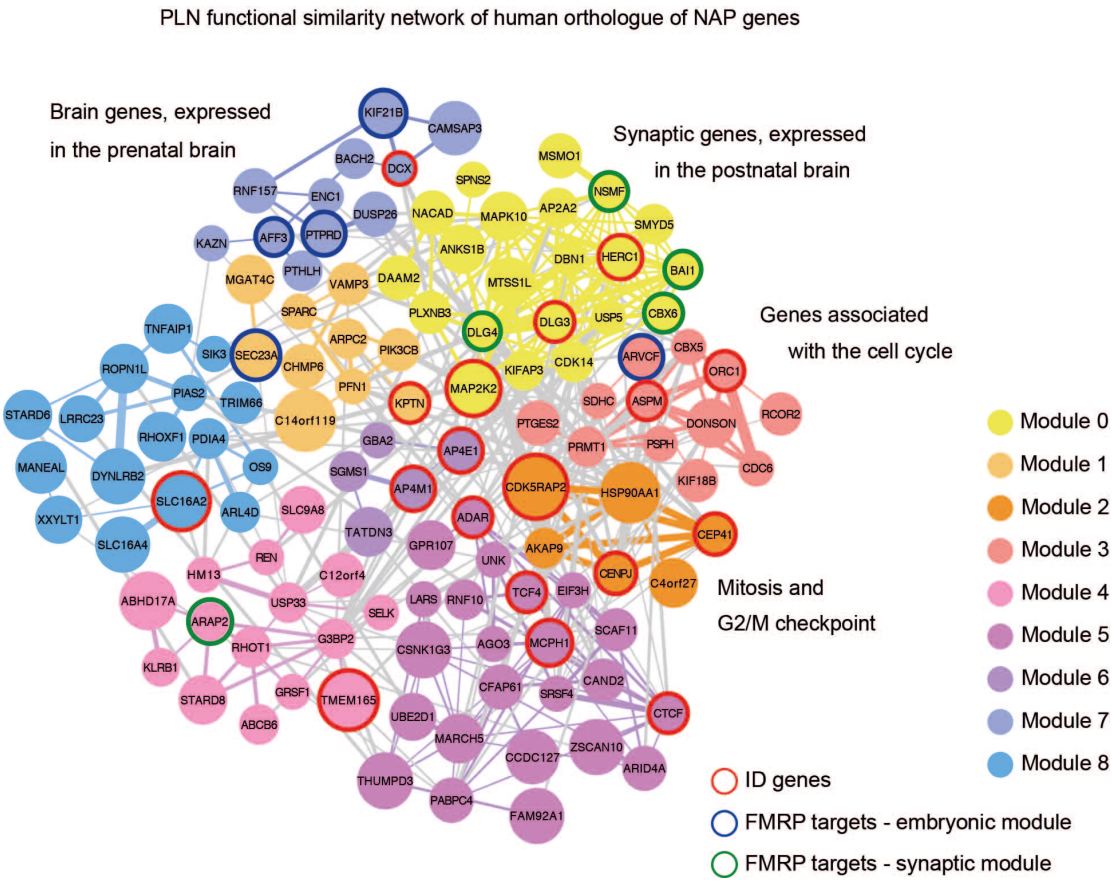


Fig S6.5 | Functional similarity network of human NAP orthologues. The PLN (Phenotypic Linkage Network) sub-network, consisting 381 links between 121 genes of human orthologues of mouse NAP genes, was partitioned into nine modules of functionally similar genes, illustrated by the colour of the nodes and edges. The thickness of the edges is proportional to the functional similarity score within the PLN. Grey edges depict inter-module links. Red node borders depict known ID genes, whereas blue and green borders refer to the set of embryonically-expressed and mature FMRP target genes, respectively. Module descriptions were added, where clearly discernable. NAP genes, neuroanatomical phenotype genes, defined as those associated with neuroanatomical deviations in knockout compared to wildtype mice (significant at 10% FDR).

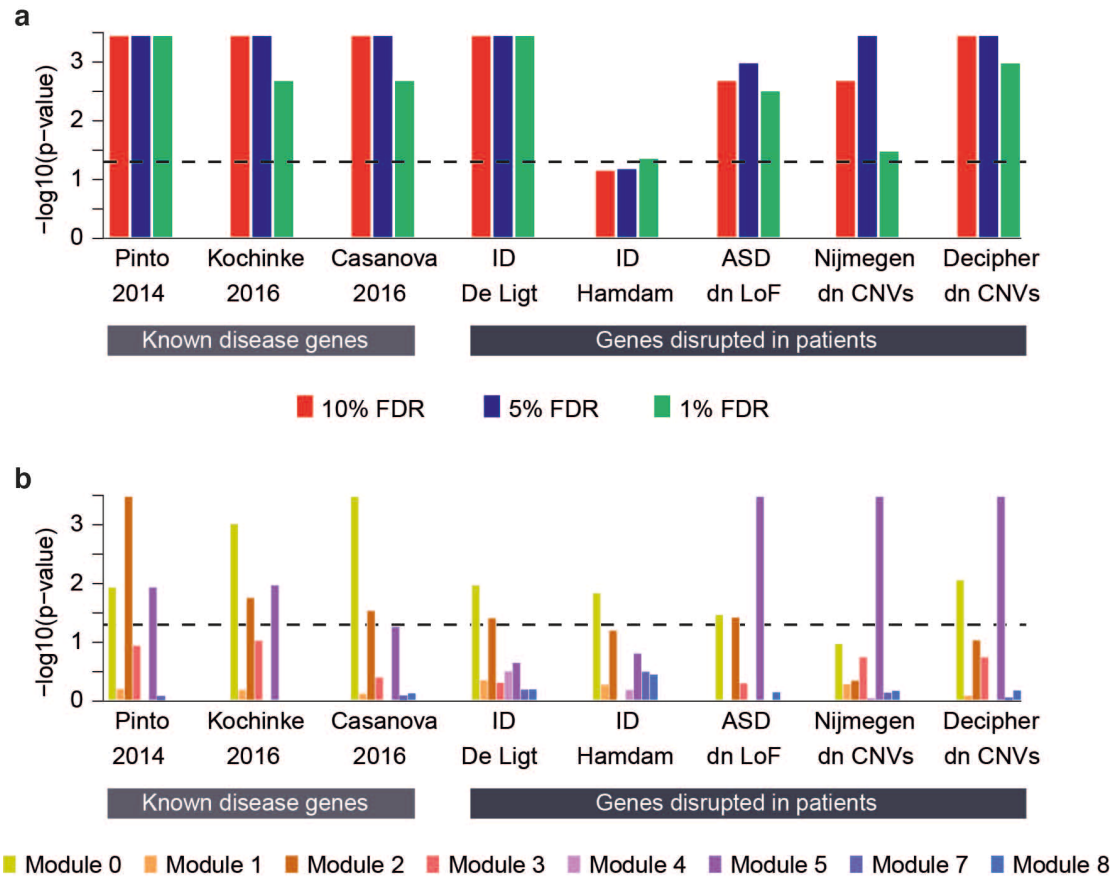


Fig S6.6 | Functional similarity of human orthologues of NAP genes or human module genes and disease genes. We evaluated the functional connections between (**a**) human orthologues of NAP genes or (**b**) human module genes with known ID genes or genes affected by *de novo* base substitutions (SNVs) and CNVs in ID or ASD patients (**Table 2.1**). For this, we compared their interconnectedness to that of randomly sampled (**a**) non-NAP orthologues (matched for CDS and degree) and (**b**) NAP orthologues. We derived empirical p-values, reflecting the proportion of randomisations (n = 1000) during which random genes were more functionally similar within the PLN (Phenotypic Linkage Network) to the sets tested (**section 2.5.3**). NAP genes, neuroanatomical phenotype genes, defined as those associated with neuroanatomical deviations in knockout compared to wildtype mice (significant at 10% FDR).

SUPPLEMENTARY TABLES

Table S3.1 | Transcriptional gene regulation is affected in GTS patients.

GO term	P-value	Odds ratio	# patients with hit in pathway
DNA binding (GO:0003677)	0.024	1.542	38
RNA biosynthetic process (GO:0032774)	0.066	1.393	37
aromatic compound biosynthetic process (GO:0019438)	0.021	1.509	45
heterocycle biosynthetic process (GO:0018130)	0.045	1.424	42
nucleic acid binding (GO:0003676)	0.106	1.279	48
nucleic acid-templated transcription (GO:0097659)	0.048	1.46	35
nucleobase-containing compound biosynthetic process (GO:0034654)	0.041	1.437	42
organic cyclic compound biosynthetic process (GO:1901362)	0.011	1.579	47
regulation of RNA biosynthetic process (GO:2001141)	0.135	1.249	46
regulation of RNA metabolic process (GO:0051252)	0.143	1.24	46
regulation of nucleic acid-templated transcription (GO:1903506)	0.135	1.249	46
regulation of nucleobase-containing compound metabolic process (GO:0019219)	0.13	1.245	50
regulation of transcription, DNA-templated (GO:0006355)	0.135	1.249	46
transcription, DNA-templated (GO:0006351)	0.048	1.46	35

Considering only those GO terms (i.e. pathways) significantly over-represented among case genes, we evaluated whether an excess of Tourette (GTS) patients harbours CNVs mapping to genes annotated with these terms compared to healthy controls. Over-representations were assessed using right-tailed Fisher's tests.

Table S3.2 | GO enrichment analysis stratified by CNV size and direction.

CNV direction	CNV size	GO term	P-value	BH-adjusted p-value	# genes
Gain	20 to 100 kb	DNA binding (GO:0003677)	0.011	0.307	15
		nucleic acid binding (GO:0003676)	0.018	0.307	20
		regulation of transcription, DNA-templated (GO:0006355)	0.025	0.296	18
		regulation of nucleic acid-templated transcription (GO:1903506)	0.027	0.296	18
		plasma membrane part (GO:0044459)	0.027	0.976	17
	100 to 500 kb	RNA biosynthetic process (GO:0032774)	< 1 x 10-8	< 1 x 10-8	31
		nucleobase-containing compound biosynthetic process (GO:0034654)	< 1 x 10-8	< 1 x 10-8	33
		heterocycle biosynthetic process (GO:0018130)	< 1 x 10-8	< 1 x 10-8	33
		aromatic compound biosynthetic process (GO:0019438)	< 1 x 10-8	< 1 x 10-8	33
		organic cyclic compound biosynthetic process (GO:1901362)	< 1 x 10-8	< 1 x 10-8	33
		DNA binding (GO:0003677)	< 1 x 10-8	< 1 x 10-8	29
		transcription, DNA-templated (GO:0006351)	< 1 x 10-8	< 1 x 10-8	27
		nucleic acid-templated transcription (GO:0097659)	< 1 x 10-8	< 1 x 10-8	27
		cellular nitrogen compound biosynthetic process (GO:0044271)	< 1 x 10-8	< 1 x 10-8	33
		regulation of transcription, DNA-templated (GO:0006355)	< 1 x 10-8	< 1 x 10-8	34
		regulation of nucleic acid-templated transcription (GO:1903506)	< 1 x 10-8	< 1 x 10-8	34
		nucleic acid binding (GO:0003676)	< 1 x 10-8	< 1 x 10-8	35
	> 500 kb	receptor activity (GO:0004872)	< 1 x 10-8	0.005	11
		molecular transducer activity (GO:0060089)	< 1 x 10-8	0.005	11
		signal transducer activity (GO:0004871)	0.001	0.012	11
		plasma membrane (GO:0005886)	0.014	0.519	16
		establishment of localization in cell (GO:0051649)	0.261	0.991	5
Loss	20 to 100 kb	signal transducer activity (GO:0004871)	0.016	0.264	8
		receptor activity (GO:0004872)	0.023	0.264	7
		molecular transducer activity (GO:0060089)	0.023	0.264	7
		single-multicellular organism process (GO:0044707)	0.045	1	9
		single-organism metabolic process (GO:0044710)	0.059	1	10
	100 to 500 kb	plasma membrane (GO:0005886)	0.017	0.425	4
		integral component of plasma membrane (GO:0005887)	0.059	0.471	2
		intrinsic component of plasma membrane (GO:0031226)	0.065	0.471	2
		regulation of biological quality (GO:0065008)	0.188	1	2
		plasma membrane part (GO:0044459)	0.186	0.838	2
	> 500 kb	protein complex (GO:0043234)	0.126	1	4
		extracellular region part (GO:0044421)	0.133	1	4
		extracellular space (GO:0005615)	0.189	1	2
		protein localization (GO:0008104)	0.212	1	2
		macromolecule localization (GO:0033036)	0.215	1	2

Considering CNV deletions (i.e. losses) and duplications (i.e. gains) of sizes 20 to 100 kb, 100 to 500 kb and > 500 kb, we assessed GO term over-representation among case genes mapping to the corresponding case CNVs compared to genes mapping to the corresponding CT CNVs. P-values were derived using right-tailed hypergeometric tests and corrections for multiple testing were applied the Benjamini-Hochberg (BH) method to control for the number of terms tested. The table shows the top five or twelve GO terms with smallest BH p-values for each CNV set.

Table S3.3 | Case gene clustering in GTEx, Brainspan and PPI networks.

Test set	Background		GTEx r>0.5	GTEx r>0.6	GTEx r>0.7	Brainspan r>0.5	Brainspan r>0.6	Brainspan r>0.7	PPIs (String)
Case genes	Genome	Number of genes in network	225	218	190	192	191	181	207
		Empirical p-value	0.003	0.002	0	0.1	0.046	0.009	0.887
Case genes	CT genes	Number of genes in network	225	218	190	192	191	181	207
		Empirical p-value	0	0	0	0.04	0.002	0.0	0.022
GTS CNVR genes	CT CNVR genes	Number of genes in network	196	205	163	180	179	169	212
		Empirical p-value	0.033	0.031	0.19	0	0.005	0	0.953

The similarity of expression patterns of case genes within the GTEx and Brainspan co-expression networks (with different minimum co-expression coefficient thresholds) was compared to that of random genome genes or control (CT) genes, matched for coding sequence length and network connectivity. The interconnectedness of case genes was also compared to that of the same number of CT genes matched for number of genes per GTS CNVR (CNV region). Shown here are the associated p-values and the number of genes co-expressed with one another in the evaluated networks. Using the same methodology, we compared the number of physical protein-protein interactions (PPI) between case and CT genes using the String PPI network.

Table S3.4 | Likelihood of haploinsufficiency of case genes and module gene.

Case set		Control set	P-value	Median HIS case	Median HIS control
Case genes		Genome	0.106	0.517	0.506
		CT genes	0.548	0.517	0.516
GTS CNVR co-expressed genes		Genome	0	0.542	0.506
		CT genes	0	0.542	0.516
		CT CNVR co-expressed genes	0.085	0.542	0.534
		GTS CNVR non co-expressed genes	0	0.542	0.444
Modules	Module 0	Genome	0	0.5935	0.50593
	Module 1		0.574	0.5158	0.50593
	Module 2		0.033	0.5621	0.50593
	Module 4		0.008	0.5491	0.50593
	Module 5		0.002	0.5518	0.50593
	Module 7		0.018	0.5741	0.50593
	Module 8		0.634	0.4553	0.50593
	Module 11		0.108	0.5785	0.50593
	Module 0	CT genes	8.06E-05	0.5935	0.51588
	Module 1		0.57061	0.5158	0.51588
	Module 2		0.03299	0.5621	0.51588
	Module 4		0.00317	0.5491	0.51588
	Module 5		0.00099	0.5518	0.51588
	Module 7		0.01139	0.5741	0.51588
	Module 8		0.79292	0.4553	0.51588
	Module 11		0.07153	0.5785	0.51588
	Module 0	CT CNVR co-expressed genes	0.00194	0.5935	0.53374
	Module 1		0.92583	0.5158	0.53374
	Module 2		0.09709	0.5621	0.53374
	Module 4		0.11032	0.5491	0.53374
	Module 5		0.07075	0.5518	0.53374
	Module 7		0.05468	0.5741	0.53374
	Module 8		0.93207	0.4553	0.53374
	Module 11		0.16272	0.5785	0.53374
	Module 0	GTS CNVR co-expressed genes	0.00372	0.5935	0.5371
	Module 1		0.98819	0.5158	0.55145
	Module 2		0.12695	0.5621	0.54117
	Module 4		0.2856	0.5491	0.5371
	Module 5		0.20402	0.5518	0.53941
	Module 7		0.0975	0.5741	0.54047
	Module 8		0.95957	0.4553	0.54223
	Module 11		0.23998	0.5785	0.54117

Genome-wide haploinsufficiency (HIS) scores (Steinberg *et al.*, 2015) (**Table 2.1**) of case genes, GTS CNVR co-expressed genes and module genes compared to random genome genes (matched for coding DNA sequence length), CT genes, CT CNVR co-expressed genes or remaining GTS CNVR co-expressed genes. P-values were derived using a permutation-based approach (**section 2.7.2**) when comparisons were performed using genome genes and otherwise using right-tailed Mann-Whitney U tests. CNVR, CNV region; CT, control.

Table S3.5 | Gene set enrichments of GTS module genes.

Gene set tested		Module	P-value	Odds Ratio	# module / rest GTS CNVR co-expressed genes
Brain and synaptic functions	Brain-expressed	Module 0	0.8872	0.6247	14 / 108
		Module 1	0.9986	0.2515	8 / 114
		Module 2	0.7105	0.9348	6 / 116
		Module 4	0.0098	5.1088	33 / 89
		Module 5	0.0462	2.4824	40 / 82
		Module 7	0.3421	2.4811	8 / 114
		Module 8	0.9075	0.4697	3 / 119
		Module 11	0.2652	inf	5 / 117
	Brain-specific	Module 4	0.043	3.6101	5 / 2
		Module 11	0.0244	13.9244	2 / 5
	FMRP targets	Module 2	0.088	5.4839	2 / 7
		Module 4	0.3681	1.5688	3 / 6
		Module 5	0.3114	1.6325	4 / 5
	Postaynaptic density	Module 0	0.7184	0.8397	1 / 8
		Module 1	0.6657	0.987	1 / 8
		Module 2	0.3946	2.3794	1 / 8
		Module 4	0.8687	0.4949	1 / 8
		Module 5	0.7398	0.7813	1 / 8
		Module 11	0.2714	4.1299	1 / 8
	Synaptosome	Module 4	0.3275	4.6237	1 / 0
Regulation of gene expression	Constrained	Module 1	0.6258	1.1169	1 / 7
		Module 2	0.3625	2.6901	1 / 7
		Module 4	0.3129	1.7751	3 / 5
		Module 5	0.4488	1.355	3 / 5
	Transcription factors	Module 0	0.6619	0.906	4 / 29
		Module 1	0.9071	0.4827	2 / 31
		Module 4	0.0161	2.5513	14 / 19
		Module 5	0.7314	0.8328	8 / 25
		Module 7	0.2942	1.9157	3 / 30
		Module 8	0.6896	0.9623	1 / 32
High confidence disease genes	MIM Morbid	Module 0	0.0933	2.3685	6 / 17
		Module 1	0.5083	1.1901	3 / 20
		Module 2	0.3394	1.9756	2 / 21
		Module 4	0.5947	0.9928	5 / 18
		Module 5	0.9566	0.4269	3 / 20
	ASD (SFARI)	Module 1	0.417	2.2798	1 / 3
		Module 4	0.6323	1.1463	1 / 3
		Module 5	0.3931	1.8079	2 / 2
	ID (Pinto *et al.*, 2014)	Module 5	0.4495	2.3109	1 / 1
		Module 7	0.5265	1.7895	1 / 1
	ID (Kochinke *et al.*, 2016)	Module 0	0.5281	1.5456	1 / 4
		Module 4	0.3668	1.8715	2 / 3
		Module 5	0.4783	1.439	2 / 3
	ID (Casanova *et al.*, 2016)	Module 4	0.5498	1.5354	1 / 2
		Module 5	0.3025	2.4213	2 / 1

The overlap between various gene sets (**Table 2.1**) and modules was evaluated against remaining GTS CNVR co-expressed genes using right-tailed Fisher's tests. Results are only displayed for modules, which comprise at least one gene present in the evaluated gene set. CNVR, CNV region.

Table S3.6 | GO enrichment analysis of GTS modules.

Module	GO term	P-value	BH-adjusted p-value	# genes	# modules genes in GO / total
0	DNA metabolic process (GO:0006259)	4.20E-05	0.048	6	16 / 21
	DNA recombination (GO:0006310)	0.00082	0.4678	3	16 / 21
	regulation of cell activation (GO:0050865)	0.00153	0.5818	4	16 / 21
	chromosomal part (GO:0044427)	0.00183	0.2115	5	16 / 21
	nuclear chromosome part (GO:0044454)	0.00234	0.2115	4	16 / 21
1	regulation of actin cytoskeleton organization (GO:0032956)	0.00145	0.508	3	13 / 18
	regulation of actin filament-based process (GO:0032970)	0.00194	0.508	3	13 / 18
	positive regulation of cell migration (GO:0030335)	0.00373	0.508	3	13 / 18
	positive regulation of cell motility (GO:2000147)	0.00407	0.508	3	13 / 18
	regulation of cytoskeleton organization (GO:0051493)	0.00412	0.508	3	13 / 18
2	actin binding (GO:0003779)	0.00134	0.3	3	8 / 8
	trans-Golgi network (GO:0005802)	0.00241	0.2939	2	8 / 8
	Golgi subcompartment (GO:0098791)	0.00414	0.2939	2	8 / 8
	organelle subcompartment (GO:0031984)	0.00487	0.2939	2	8 / 8
	cytoskeletal protein binding (GO:0008092)	0.01108	1	3	8 / 8
4	protein localization to membrane (GO:0072657)	0.00018	0.1877	4	16 / 35
	single-organism cellular localization (GO:1902580)	0.00033	0.1877	5	16 / 35
	cilium (GO:0005929)	0.00151	0.273	3	16 / 35
	cell projection organization (GO:0030030)	0.0028	0.58	4	16 / 35
	single-organism membrane organization (GO:0044802)	0.00283	0.58	4	16 / 35
5	positive regulation of I-kappaB kinase/NF-kappaB signaling (GO:0043123)	0.00298	1	3	26 / 45
	positive regulation of defense response (GO:0031349)	0.00361	1	4	26 / 45
	ncRNA processing (GO:0034470)	0.00374	1	4	26 / 45
	positive regulation of immune system process (GO:0002684)	0.00398	1	6	26 / 45
	ubiquitin-like protein transferase activity (GO:0019787)	0.00421	0.9433	4	26 / 45
7	regulation of transmembrane receptor protein serine/threonine kinase signaling pathway (GO:0090092)	0.00249	1	2	7 / 9
	regulation of cellular response to growth factor stimulus (GO:0090287)	0.00291	1	2	7 / 9
	negative regulation of protein modification (GO:0031400)	0.01942	1	2	7 / 9
	negative regulation of cell differentiation (GO:0045596)	0.02215	1	2	7 / 9
	promoter sequence-specific DNA binding (GO:0001046)	0.02248	1	1	7 / 9
8	monocarboxylic acid transport (GO:0015718)	0.02163	1	1	4 / 5
	carboxylic acid transmembrane transporter activity (GO:0046943)	0.02545	1	1	4 / 5
	organic acid transmembrane transporter activity (GO:0005342)	0.02609	1	1	4 / 5
	tumor necrosis factor-mediated signaling pathway (GO:0033209)	0.02788	1	1	4 / 5
	protein ubiquitination involved in ubiquitin-dependent protein catabolic process (GO:0042787)	0.02894	1	1	4 / 5
11	glycosaminoglycan biosynthetic process (GO:0006024)	0.01303	1	1	5 / 5
	aminoglycan biosynthetic process (GO:0006023)	0.01315	1	1	5 / 5
	terpenoid metabolic process (GO:0006721)	0.01495	1	1	5 / 5
	isoprenoid metabolic process (GO:0006720)	0.01735	1	1	5 / 5
	glycosaminoglycan metabolic process (GO:0030203)	0.01915	1	1	5 / 5

For each gene module, we evaluated the over-representation of GO terms using the whole genome as a background. Enrichments were performed using right-tailed hypergeometric tests. Corrections for multiple testing were applied using the Benjamini-Hochberg (BH) method to control for the number of terms tested. The table shows the top 5 GO terms with smallest p-values for each module.

Table S5.1 | Summary of GO terms over-represented among GTEx modules.

Mod	GO term	P	#
0	neuron part (GO:0097458)	< 1E-16	57
	synapse part (GO:0044456)	< 1E-16	37
	neuron projection (GO:0043005)	7.44E-15	37
	synaptic membrane (GO:0097060)	2.44E-14	23
	transmembrane transporter complex (GO:1902495)	5.97E-13	23
	transporter complex (GO:1990351)	8.71E-13	23
	cell projection (GO:0042995)	1.11E-12	47
	ion channel complex (GO:0034702)	4.82E-12	21
	modulation of synaptic transmission (GO:0050804)	5.84E-13	22
	postsynaptic membrane (GO:0045211)	1.70E-10	17
1	DNA metabolic process (GO:0006259)	1.89E-14	36
	nuclear part (GO:0044428)	1.80E-13	85
	response to DNA damage stimulus (GO:0006974)	1.72E-13	34
	cellular response to stress (GO:0033554)	2.33E-13	47
	DNA repair (GO:0006281)	4.68E-13	27
	DNA recombination (GO:0006310)	3.24E-09	15
	cell cycle process (GO:0022402)	1.30E-08	34
	mitotic cell cycle process (GO:1903047)	3.88E-08	27
	cell cycle checkpoint (GO:0000075)	2.04E-05	9
	cell division (GO:0051301)	3.55E-05	14
2	epithelial cell differentiation (GO:0030855)	0.00016	6
	anchoring junction (GO:0070161)	0.01441	6
	apical plasma membrane (GO:0016324)	0.00775	4
	cell junction (GO:0030054)	0.01165	9
	extracellular exosome (GO:0070062)	0.0151	14
	extracellular organelle (GO:0043230)	0.01575	14
	extracellular vesicle (GO:1903561)	0.01571	14
	membrane (GO:0016020)	0.0062	31
	plasma membrane (GO:0005886)	0.00503	21
	plasma membrane region (GO:0098590)	0.01166	7
3	assembly involved in morphogenesis (GO:0010927)	9.03E-10	16
	cilium organization (GO:0044782)	2.77E-09	13
	cilium assembly (GO:0042384)	9.50E-09	12
	cell projection organization (GO:0030030)	1.34E-08	25
	microtubule (GO:0005874)	1.41E-07	17
	microtubule organizing center (GO:0005815)	4.62E-07	20
	cell projection assembly (GO:0030031)	1.51E-07	14
	microtubule-based process (GO:0007017)	2.44E-07	19
	cytoskeletal part (GO:0044430)	1.71E-06	34
	cell projection morphogenesis (GO:0048858)	3.28E-05	11
4	ion channel complex (GO:0034702)	3.04E-08	11
	contractile fiber part (GO:0044449)	1.33E-07	9
	transmembrane transporter complex (GO:1902495)	8.92E-08	11
	transporter complex (GO:1990351)	1.07E-07	11
	Z disc (GO:0030018)	4.65E-07	7
	protein complex (GO:0043234)	7.89E-07	38
	calcium ion transporter activity (GO:0015085)	1.62E-06	7
	cation channel activity (GO:0005261)	9.23E-07	10
	channel activity (GO:0015267)	1.32E-06	12
	ion channel activity (GO:0005216)	4.32E-07	12

Mod	GO term	P	#
5	chromatin modification (GO:0016569)	< 1E-16	36
	chromatin organization (GO:0006325)	< 1E-16	40
	RNA metabolic process (GO:0016070)	< 1E-16	89
	histone modification (GO:0016570)	2.61E-13	24
	transcription, DNA-templated (GO:0006351)	1.61E-10	59
	regulation of gene expression (GO:0010468)	5.78E-10	85
	cellular protein modification process (GO:0006464)	9.60E-09	62
	chromatin binding (GO:003682)	1.99E-07	20
	RNA processing (GO:0006396)	4.94E-07	27
	histone acetylation (GO:0016573)	1.59E-06	9
6	cell adhesion (GO:0007155)	1.78E-14	37
	regulation of cell motility (GO:2000145)	1.18E-11	25
	regulation of cell migration (GO:0030334)	1.91E-11	24
	cell-substrate adhesion (GO:0031589)	2.97E-11	13
	adherens junction (GO:0005912)	1.45E-10	23
	anchoring junction (GO:0070161)	2.55E-10	23
	extracellular matrix component (GO:0044420)	8.59E-10	11
	positive regulation of locomotion (GO:0040017)	6.49E-09	17
	regulation of developmental process (GO:0050793)	2.39E-08	39
	regulation of cell morphogenesis (GO:0022604)	1.78E-07	16
10	carboxylic acid metabolic process (GO:0019752)	2.45E-06	7
	oxoacid metabolic process (GO:0043436)	2.56E-06	7
	organic acid metabolic process (GO:0006082)	4.83E-06	7
	small molecule metabolic process (GO:0044281)	3.12E-05	8
	lipid modification (GO:0030258)	4.60E-05	4
	monocarboxylic acid metabolic process (GO:0032787)	5.51E-05	5
	microbody (GO:0042579)	0.00018	3
	peroxisome (GO:0005777)	9.66E-05	3
	fatty acid metabolic process (GO:0006631)	0.00015	4
	small molecule catabolic process (GO:0044282)	0.00018	4
11	side of membrane (GO:0098552)	0.00012	7
	GTPase activator activity (GO:0005096)	0.00033	7
	GTPase regulator activity (GO:0030695)	0.0007	7
	enzyme activator activity (GO:0008047)	0.00178	8
	ligase activity (GO:0016874)	0.00281	7
	zinc ion binding (GO:0008270)	0.00303	13
	histone modification (GO:0016570)	0.00018	8
	cation binding (GO:0043169)	0.0048	30
	cell death (GO:0008219)	0.00153	10
	chromatin modification (GO:0016569)	0.00148	8

Displayed here are the number of genes (#) annotated with ten of the most strongly associated GO terms with each module (Mod) and the associated nominal p-values (P). We note that significances persist following Benjamini-Hochberg (BH) correction for multiple testing. GO term descriptions were shortened to fit the table.

Table S5.2 | Summary of GO terms over-represented among Brainspan modules.

Mod	GO term	P	#
0	gene silencing (GO:0016458)	7.50E-10	11
	chromatin organization (GO:0006325)	3.34E-09	18
	DNA metabolic process (GO:0006259)	6.82E-09	19
	cell cycle process (GO:0022402)	2.39E-08	22
	cellular response to stress (GO:0033554)	5.97E-08	24
	negative regulation of gene expression (GO:0010629)	6.12E-07	23
	DNA recombination	8.59E-07	23
	regulation of histone modification (GO:0031056)	4.45E-06	7
	mitotic cell cycle (GO:1903047)	7.75E-06	15
	chromatin remodelling (GO:0006338)	9.25E-06	7
1	RNA processing (GO:0006396)	1.91E-08	30
	mRNA metabolic process (GO:0016070)	5.35E+07	23
	response to DNA damage stimulus (GO:0006974)	1.48E-06	24
	regulation of gene expression (GO:0010468)	2.19E-06	77
	regulation of transcription (GO:0006355)	9.68E-06	65
	DNA repair (GO:0006281)	1.49E-05	17
	cellular response to stress (GO:0033554)	4.06E-05	32
	chromatin organisation (GO:0006325)	9.00E-05	19
	ubiquitin-protein transferase activity (GO:0004842)	1.58E-04	14
	histone binding (GO:0042393)	0.00032	8

Mod	GO term	P	#
2	neuron part (GO:0097458)	4.33E-12	53
	synapse part (GO:0044456)	2.07E-09	30
	regulation of GTPase activity (GO:0043087)	7.51E-09	33
	cell projection (GO:0042995)	8.49E-08	46
	neuron projection (GO:0043005)	1.51E-07	31
	synaptic membrane (GO:0097060)	4.85E-07	17
	modulation of synaptic transmission (GO:0050804)	1.05E-06	17
	regulation of membrane potential (GO:0042391)	1.28E-06	19
	cell junction (GO:0030054)	2.49E-06	43
	axon part (GO:0033267)	1.23E-05	13
3	chromatin modification (GO:0016569)	1.38E-13	30
	chromatin organization (GO:0006325)	3.39E-12	33
	histone modification (GO:0016570)	6.06E-12	24
	nervous system development (GO:0007399)	2.20E-10	20
	histone acetylation (GO:0016573)	5.76E-08	11
	transcription, DNA-templated (GO:0006351)	2.00E-06	55
	regulation of axonogenesis (GO:0050770)	5.78E-06	11
	neuron projection development (GO:0010975)	3.34E-05	16
	growth cone (GO:0030426)	9.40E-05	8
	cytoskeleton organisation (GO:0007010)	5.84E-04	20

Displayed here are the number of genes (#) annotated with ten of the most strongly associated GO terms with each module (Mod) and the associated nominal p-values (P). We note that significances persist following Benjamini-Hochberg (BH) correction for multiple testing. GO term descriptions were shortened to fit the table. Module 5 is excluded from the table as no terms are over-represented.

Table S5.3 | Summary of GO terms over-represented among PLN modules.

Mod	GO term	P	#
0	anchoring junction (GO:0070161)	< 1E-16	42
	cell adhesion (GO:0007155)	< 1E-16	65
	intracellular signal transduction (GO:0035556)	< 1E-16	59
	cell migration (GO:0016477)	2.22E-16	41
	neuron projection development (GO:0010975)	1.61E-13	27
	regulation of axonogenesis (GO:0050770)	8.49E-12	17
	regulation of neurogenesis (GO:0050767)	3.47E-11	31
	cell projection organisation (GO:0030030)	9.68E-10	30
2	transmembrane transporter activity (GO:0022804)	< 1E-16	25
	ATPase act & substance movement (GO:0043492)	< 1E-16	17
	transmembrane transport (GO:0055085)	< 1E-16	33
	ion transport (GO:0006811)	1.11E-16	27
	establishment of localisation (GO:0051234)	2.33E-14	40
	ATPase activity (GO:0016887)	1.64E-13	17
	lipid transport (GO:0006869)	3.28E-11	12
	anion transmembrane transport (GO:0098656)	5.50E-10	10
3	mRNA metabolic process (GO:0016071)	< 1E-16	31
	posttranscriptional regulation (GO:0010608)	3.70E-12	22
	regulation of translation (GO:0006417)	4.72E-12	19
	mRNA processing (GO:0006397)	1.07E-08	17
	RNA splicing (GO:0008380)	4.10E-08	15
	chromsome organisation (GO:0051276)	3.90E-07	14
	DNA repair (GO:0006281)	8.38E-07	15
	regulation of gene expression (GO:0010468)	9.05E-07	55
4	cytoskeletal part (GO:0044430)	< 1E-16	35
	microtubule (GO:0005874)	< 1E-16	20
	mitotic cell cycle process (GO:1903047)	< 1E-16	23
	cell cycle G2/M phase transition (GO:0044839)	2.22E-16	13
	microtubule organising center (GO:0044770)	8.60E-12	16
	centrosome (GO:0005813)	3.00E-09	12
	cell division (GO:0051301)	1.39E-08	11
	cytoskeleton organisation (GO:0007010)	2.09E-07	13
5	carboxylic acid metabolic process (GO:0019752)	< 1E-16	34
	oxoacid metabolic process (GO:0043436)	< 1E-16	34
	lipid metabolic process (GO:0006629)	1.33E-13	30
	lipid transport (GO:0006869)	8.43E-07	10
	mitochondrion (GO:0005739)	3.31E-07	22
	mitochondrial protein complex (GO:0098798)	2.65E-05	7
	oxidoreductase activity (GO:0016491)	7.75E-04	12
	electron carrier activity (GO:0009055)	0.00357	4
6	calcium channel activity (GO:0005262)	< 1E-16	18
	calcium transmembrane transporter act (GO:0015085)	< 1E-16	20
	calcium ion transmembrane transport (GO:0070588)	< 1E-16	20
	voltage-gated ion channel activity (GO:0005244)	2.65E-13	14
	muscle system process (GO:0003012)	1.48E-11	14
	muscle contraction (GO:0006936)	1.74E-11	13
	regulation of membrane potential (GO:0042391)	1.21E-10	15
	ion homeostasis (GO:0050801)	3.49E-09	17

Mod	GO term	P	#
7	protein modification by conjugation (GO:0032446)	< 1E-16	37
	protein ubiquitination (GO:0016567)	< 1E-16	35
	protein modification process (GO:0036211)	9.55E-13	58
	microtubule organising centre (GO:0005815)	3.54E-09	21
	centrosome (GO:0005813)	1.27E-07	16
	mitotic cell cycle process (GO:1903047)	1.44E-04	17
	cell projection organisation (GO:0030030)	0.00016	16
	Wnt signaling pathway (GO:0016055)	6.22E-04	10
8	neuron part (GO:0097458)	2.67E-13	36
	ion channel activity (GO:0005216)	1.00E-09	18
	synapse part (GO:0044456)	3.63E-09	20
	axon part (GO:0033267)	2.49E-08	12
	regulation of transport (GO:0051049)	1.68E-09	37
	modulation of synaptic transmission (GO:0050804)	9.54E-09	14
	bhevaiour (GO:0007610)	1.12E-07	17
	learning and memory (GO:0007611)	3.09E-07	11
9	Golgi apparatus part (GO:0044431)	2.51E-09	11
	Golgi membrane (GO:0000139)	2.73E-09	10
	cytoplasmic part (GO:0044444)	1.18E-06	20
	endoplasmic reticulum part (GO:0044432)	5.31E-06	9
	Golgi vesicle transport (GO:0048193)	1.72E-06	6
	endoplasmic reticulum lumen (GO:0005788)	7.59E-05	4
	ER to Golgi vesicle-mediated transport (GO:0006888)	4.79E-05	4
	protein transport (GO:0015031)	0.00015	8
11	GTPase mediated signal transduction (GO:0051056)	4.76E-12	13
	positive regulation of GTPase activity (GO:0043547)	7.58E-11	16
	positive regulation of hydrolase activity (GO:0051345)	2.08E-10	18
	regulation of Ras signal transduction (GO:0046578)	5.61E-10	10
	regulation of Rho signal transduction (GO:0035023)	2.98E-09	8
	GTPase regulator activity (GO:0030695)	1.06E-06	9
	synapse organisation (GO:0050808)	4.00E-05	5
	regulation of cell communication (GO:0010646)	5.23E-05	22
13	chromatin binding (GO:0003682)	< 1E-16	32
	chromatin modification (GO:0016569)	< 1E-16	53
	histone modification (GO:0016570)	< 1E-16	40
	regulation of gene expression (GO:0010468)	< 1E-16	112
	transcription, DNA-templated (GO:0006351)	< 1E-16	83
	transcription fcator binding (GO:0008134)	7.77E-16	30
	DNA conformation change (GO:0071103)	2.92E-07	10
	cell fate commitment (GO:0045165)	6.24E-07	10

Displayed here are the number of genes (#) annotated with eight of the most strongly associated GO terms with each module (Mod) and the associated nominal p-values (P). We note that significances persist following Benjamini-Hochberg (BH) correction for multiple testing. GO term descriptions were shortened to fit the table. PLN, Phenotypic Linkage Network.

Table S5.4 | Summary of MGI phenotypes over-represented among GTEx modules.

Mod	MGI phenotype	P	#
0	behavior/neurological phenotype (MP:0005386)	< 1E-16	70
	nervous system physiology (MP:0003633)	< 1E-16	60
	behavior (MP:0004924)	< 1E-16	70
	nervous system phenotype (MP:0003631)	2.09E-14	73
	synaptic transmission (MP:0003635)	9.99E-14	32
	CNS synaptic transmission (MP:0002206)	1.56E-11	27
	motor coordination/ balance (MP:0001516)	1.92E-10	30
	motor capab/coordination/movement (MP:0002066)	6.38E-10	46
	behavioral response to xenobiotic (MP:0009745)	1.14E-09	19
	sensory capab/reflexes/nociception (MP:0002067)	4.03E-09	28
1	cell physiology (MP:0005621)	2.90E-06	26
	cellular phenotype (MP:0005384)	2.18E-05	26
	cell death (MP:0000313)	0.00014	14
	embryonic lethality (MP:0008762)	0.00017	23
	induced cell death (MP:0008942)	0.00044	7
	sensitivity to induced cell death (MP:0008943)	0.00046	6
	apoptosis (MP:0001648)	0.00047	10
	prenatal lethality (MP:0002080)	0.00071	26
	increased classified tumor incidence (MP:0010273)	0.00072	9
	increased organ/body tumor incidence (MP:0010274)	0.00083	9
2	epidermal layer morphology (MP:0001216)	4.93E-05	6
	hair follicle morphology (MP:0000377)	0.00653	3
	increased single-positive T cell number (MP:0008082)	0.00672	3
	increased T cell number (MP:0005015)	0.00721	4
	aorta morphology (MP:0000272)	0.00938	3
	integument phenotype (MP:0010771)	0.01027	9
	heart rate (MP:0001629)	0.0103	3
	skin morphology (MP:0002060)	0.0133	6
	fluid regulation (MP:0001784)	0.01686	5
	heartbeat (MP:0004085)	0.02067	3
3	kidney morphology (MP:0002135)	7.35E-07	16
	nephron morphology (MP:0003881)	7.91E-06	11
	preweaning lethality (MP:0010770)	8.31E-06	37
	survival (MP:0010769)	1.02E-05	40
	mortality/aging (MP:0010768)	9.53E-05	40
	cranium morphology (MP:0000438)	0.00012	11
	hearing/vestibular/ear phenotype(MP:0005377)	0.0014	11
	craniofacial bone morphology (MP:0002116)	0.00018	11
	hearing physiology (MP:0001963)	0.00032	8
	growth/size phenotype (MP:0005378)	0.004	30
4	muscle physiology (MP:0002106)	3.33E-11	20
	muscle phenotype (MP:0005369)	3.97E-11	24
	skeletal muscle morphology (MP:0000759)	4.12E-08	12
	muscle contractility (MP:0005620)	6.88E-08	11
	muscle morphology (MP:0002108)	2.63E-07	16
	impaired muscle contractility (MP:0000738)	7.33E-07	9
	skeletal muscle fiber morphology (MP:0003084)	9.68E-07	8
	muscle fiber morphology (MP:0004087)	1.47E-06	10
	kyphosis (MP:0000160)	0.00101	5
	limb grasping (MP:0001513)	0.00522	4

Mod	MGI phenotype	P	#
5	prenatal lethality (MP:0002080)	1.11E-08	40
	survival (MP:0010769)	1.07E-07	56
	preweaning lethality (MP:0010770)	2.37E-07	51
	embryonic lethality (MP:0008762)	9.09E-07	31
	mortality/aging (MP:0010768)	2.52E-06	56
	embryonic lethality in organogenesis (MP:0006207)	9.21E-05	19
	complete fetal lethality (MP:0006208)	0.00013	12
	fetal lethality (MP:0011099)	0.00014	9
	small liver (MP:0000601)	0.00015	8
	embryonic lethality in organogenesis (MP:0011098)	0.00017	16
6	cardiovascular development (MP:0002925)	1.36E-07	21
	vascular development (MP:0000259)	9.86E-06	15
	embryonic lethality in organogenesis (MP:0006207)	2.10E-05	19
	cardiovascular system phenotype (MP:0005385)	2.61E-05	31
	blood vessel morphology (MP:0001614)	2.87E-05	20
	cardiovascular system morphology (MP:0002127)	8.17E-05	25
	extraembryonic tissue morphology (MP:0002086)	0.00012	16
	embryonic tissue morphology (MP:0002085)	0.00015	18
	angiogenesis (MP:0000260)	0.00017	11
	muscle phenotype (MP:0005369)	0.00018	21
10	food intake (MP:0005449)	0.00366	3
	eating behavior (MP:0001431)	0.00906	3
	other phenotype (MP:0005395)	0.01256	2
	eating/drinking behavior (MP:0002069)	0.01347	3
	liver physiology (MP:0000609)	0.02947	2
	hepatobiliary system physiology (MP:0002139)	0.03587	2
	hepatobiliary system morphology (MP:0002138)	0.0372	3
	homeostasis/metabolism phenotype (MP:0005376)	0.04398	6
	liver/biliary system phenotype (MP:0005370)	0.05283	3
	complete postnatal lethality (MP:0011085)	0.06625	2
11	enlarged spleen (MP:0000691)	0.00278	6
	partial neonatal lethality (MP:0011088)	0.00521	5
	spleen size (MP:0002224)	0.00626	7
	hearing/vestibular/ear phenotype (MP:0005377)	0.00913	7
	spleen morphology (MP:0000689)	0.00923	8
	perinatal lethality (MP:0002081)	0.01015	10
	immune system organ morphology (MP:0002722)	0.01672	9
	tail morphology (MP:0002111)	0.01999	4
	diencephalon morphology (MP:0000830)	0.02033	4
	immune cell physiology (MP:0001819)	0.02544	10

Displayed here are the number of genes (#) annotated with ten of the most strongly associated mouse phenotypes with each module (Mod) and the associated nominal p-values (P). We note that significances persist following Benjamini-Hochberg (BH) correction for multiple testing. MGI term descriptions were shortened to fit the table.

Table S5.5 | Summary of MGI phenotypes over-represented among Brainspan modules.

Mod	MGI phenotype	P	#
0	mortality/aging (MP:0010768)	5.33E-05	33
	preweaning lethality (MP:0010770)	0.00022	28
	survival (MP:0010769)	0.00039	30
	neural tube morphology/development (MP:0002151)	0.00059	9
	somite development (MP:0001688)	0.00063	6
	cochlear outer hair cell morphology (MP:0004399)	0.00067	5
	developmental patterning (MP:0002084)	0.00069	9
	thymus morphology (MP:0000703)	0.00076	8
	sensory ganglion morphology (MP:0000960)	0.00076	6
	embryonic tissue morphology (MP:0002085)	0.00089	12
1	developmental patterning (MP:0002084)	4.62E-06	15
	embryogenesis phenotype (MP:0005380)	2.31E-05	27
	embryogenesis/ development (MP:0001672)	2.31E-05	27
	gastrulation (MP:0001695)	4.04E-05	10
	prenatal lethality (MP:0002080)	9.67E-05	29
	embryonic growth/weight/body size (MP:0002088)	0.00011	17
	placenta morphology (MP:0001711)	0.00019	11
	embryonic growth arrest (MP:0001730)	0.00029	9
	embryonic lethality (MP:0008762)	0.00038	23
	placenta development (MP:0001712)	0.00052	7

Mod	MGI phenotype	P	#
2	behavior/neurological phenotype (MP:0005386)	9.94E-13	69
	behavior (MP:0004924)	9.94E-13	69
	synaptic transmission (MP:0003635)	1.08E-08	27
	seizures (MP:0002064)	3.40E-08	20
	motor coordination/ balance (MP:0001516)	7.08E-08	28
	motor capab/coordination/movement (MP:0002066)	1.64E-07	45
	behavioral response to xenobiotic (MP:0009745)	2.93E-07	17
	nervous system physiology (MP:0003633)	6.32E-07	46
	CNS synaptic transmission (MP:0002206)	6.42E-07	22
	nervous system phenotype (MP:0003631)	7.10E-07	65
3	preweaning lethality (MP:0010770)	6.81E-09	65
	survival (MP:0010769)	1.56E-07	68
	nervous system development (MP:0003861)	2.26E-07	30
	nervous system phenotype (MP:0003631)	2.62E-06	54
	prenatal lethality (MP:0002080)	5.23E-06	42
	mortality/aging (MP:0010768)	5.25E-06	68
	embryonic lethality (MP:0008762)	1.64E-05	34
	nervous system morphology (MP:0003632)	2.02E-05	44
	neural tube closure (MP:0003720)	9.30E-05	12
	fetal lethality (MP:0006208)	0.0001	14

Displayed here are the number of genes (#) annotated with ten of the most strongly associated mouse phenotypes with each module (Mod) and the associated nominal p-values (P). We note that significances persist following Benjamini-Hochberg (BH) correction for multiple testing. MGI term descriptions were shortened to fit the table. Module 5 is excluded from the table as no phenotypes are over-represented.

Table S5.6 | Summary of MGI phenotypes over-represented among PLN modules.

Mod	MGI phenotype	P	#
0	fetal and weaning lethality (MP:0010832)	1.31E-06	58
	forebrain morphology (MP:0000783)	7.48E-06	30
	brain morphology (MP:0002152)	8.51E-06	42
	telencephalon morphology (MP:0000787)	1.05E-05	25
	internal hemorrhage (MP:0001634)	1.42E-05	16
	nervous system morphology (MP:0003632)	1.63E-05	59
	nervous system development (MP:0003861)	1.75E-05	34
	nervous system phenotype (MP:0003631)	3.95E-05	69
	enlarged heart (MP:0000274)	4.01E-05	16
	cerebellar cortex morphology (MP:0004097)	4.17E-05	14
2	partial perinatal lethality (MP:0011090)	2.40E-05	7
	impaired hearing (MP:0006325)	0.00025	5
	urinary system physiology (MP:0005502)	0.00025	8
	kidney physiology (MP:0002136)	0.00026	7
	renal tubule morphology (MP:0002703)	0.00032	5
	renal/urinary system phenotype (MP:0005367)	0.00042	10
	nephron morphology (MP:0003881)	0.00074	6
	brainstem auditory potential (MP:0004765)	0.00156	4
	ear physiology (MP:0003878)	0.00172	6
	hearing physiology (MP:0001963)	0.00186	5
3	prenatal lethality (MP:0002080)	0.00011	20
	embryonic lethality (MP:0008762)	0.00045	16
	survival (MP:0010769)	0.00054	28
	preweaning lethality (MP:0010770)	0.00115	25
	embryogenesis phenotype (MP:0005380)	0.00125	16
	embryogenesis/ development (MP:0001672)	0.00125	16
	mortality/aging (MP:0010768)	0.00241	28
	interventricular septum morphology (MP:0000281)	0.00287	5
	heart septum morphology (MP:0006113)	0.00468	5
	developmental patterning (MP:0002084)	0.00765	7
4	cerebral cortex morphology (MP:0000788)	0.00191	4
	cerebrum morphology (MP:0008540)	0.00269	5
	conditioning behavior (MP:0002062)	0.00276	4
	nervous system morphology (MP:0003632)	0.00551	11
	learning/memory/conditioning (MP:0002063)	0.00606	5
	impaired coordination (MP:0001405)	0.00807	4
	nervous system physiology (MP:0003633)	0.00842	9
	neurodegeneration (MP:0002229)	0.00887	4
	telencephalon morphology (MP:0000787)	0.01089	5
	nervous system phenotype (MP:0003631)	0.01262	12
5	lipid homeostasis (MP:0002118)	1.49E-06	13
	homeostasis (MP:0001764)	8.10E-06	27
	homeostasis/metabolism phenotype (MP:0005376)	1.73E-05	29
	lipid level (MP:0001547)	1.73E-05	11
	circulating lipid level (MP:0003949)	4.93E-05	9
	triglyceride level (MP:0000187)	0.00029	7
	circulating glucose level (MP:0000188)	0.00052	8
	glucose homeostasis (MP:0002078)	0.00063	10
	decreased cholesterol level (MP:0003983)	0.00063	5
	decreased circulating glucose level (MP:0005560)	0.00071	6

Mod	MGI phenotype	P	#
6	muscle physiology (MP:0002106)	3.18E-08	20
	muscle contractility (MP:0005620)	3.28E-08	13
	impaired muscle contractility (MP:0000738)	1.47E-06	10
	muscle phenotype (MP:0005369)	1.21E-05	21
	behavioral response to xenobiotic (MP:0009745)	1.78E-05	10
	synaptic transmission (MP:0003635)	2.11E-05	14
	muscle fiber morphology (MP:0004087)	3.15E-05	10
	heartbeat (MP:0004085)	0.00052	7
	behavior/neurological phenotype (MP:0005386)	0.00056	28
	behavior (MP:0004924)	0.00056	28
7	survival (MP:0010769)	0.00025	35
	renal tubule morphology (MP:0002703)	0.00074	6
	mortality/aging (MP:0010768)	0.00147	35
	nephron morphology (MP:0003881)	0.00371	7
	preweaning lethality (MP:0010770)	0.00435	29
	brain development (MP:0000913)	0.00764	9
	embryonic lethality (MP:0008762)	0.00875	16
	complete lethality before implantation (MP:0011094)	0.01035	4
	urinary system morphology (MP:0000516)	0.0139	9
	impaired coordination (MP:0001405)	0.01597	6
8	nervous system physiology (MP:0003633)	1.05E-09	34
	synaptic transmission (MP:0003635)	1.44E-08	19
	nervous system phenotype (MP:0003631)	5.29E-08	42
	behavior/neurological phenotype (MP:0005386)	1.68E-07	37
	behavior (MP:0004924)	1.68E-07	37
	CNS synaptic transmission (MP:0002206)	2.91E-07	16
	sensory capab/reflexes/nociception (MP:0002067)	8.36E-07	18
	seizures (MP:0002064)	5.86E-06	12
	behavioral response to xenobiotic (MP:0009745)	3.77E-05	10
	neuron morphology (MP:0002882)	7.25E-05	21
11	skeletal muscle fiber morphology (MP:0003084)	0.00032	4
	muscle fiber morphology (MP:0004087)	0.00325	4
	dendritic cell physiology (MP:0002376)	0.00342	3
	skeletal muscle morphology (MP:0000759)	0.00464	4
	sensitivity to induced morbidity/mortality (MP:0009763)	0.00492	4
	decreased survivor rate (MP:0008770)	0.00713	3
	partial lethality (MP:0010831)	0.00832	3
	physiological sensitivity to xenobiotic (MP:0008873)	0.00962	3
	physical strength (MP:0004262)	0.00996	3
	spinal cord morphology (MP:0000955)	0.01668	3
13	preweaning lethality (MP:0010770)	4.79E-08	56
	survival (MP:0010769)	1.30E-07	60
	mortality/aging (MP:0010768)	3.37E-07	62
	embryonic growth arrest (MP:0001730)	4.59E-07	14
	prenatal lethality (MP:0002080)	2.72E-06	38
	erythropoiesis (MP:0000245)	5.16E-06	17
	extraembryonic tissue morphology (MP:0002086)	9.35E-06	20
	double-negative T cell number (MP:0005089)	9.58E-06	9
	complete embryonic lethality (MP:0011092)	1.21E-05	12
	embryogenesis phenotype (MP:0005380)	1.94E-05	32

Displayed here are the number of genes (#) annotated with ten of the strongly associated mouse phenotypes with each module (Mod) and the associated nominal p-values (P). We note that significances persist following Benjamini-Hochberg (BH) correction for multiple testing. MGI term descriptions were shortened to fit the table. Module 9 was excluded from the table as no terms are over-represented. PLN, Phenotypic Linkage Network.

Table S6.1 | Clustering of NAP genes and their human orthologues.

Organism	Network Type	Name	Threshold	Empirical p-value (number of genes) 10% FDR	5% FDR	1% FDR
Mouse	Functional network	MouseNet v2	all links	0.021 (140)	0.107 (106)	0.01 (66)
Human	Functional network	PLN	all links	0.001 (140)	0.003 (106)	0.035 (64)
			top 25% links	0.001 (138)	0.001 (105)	0.001 (63)
			top 10% links	0.001 (127)	0.001 (98)	0.001 (58)
			top 1% links	0.065 (73)	0.078 (58)	0.125 (37)
	PPI network	String	all links	0.006	0.007	0.003
	Co-expression network	GTEx	r>0.6	0.169 (136)	0.242 (103)	0.441 (65)
			r>0.7	0.134 (121)	0.23 (95)	0.454 (59)
			r>0.8	0.03 (77)	0.035 (60)	0.115 (40)
			r>0.9	0.002 (23)	0.009 (19)	0.031 (14)
		Brainspan	r>0.6	0.023 (133)	0.15 (100)	0.174 (61)
			r>0.7	0.057 (130)	0.127 (98)	0.087 (60)
			r>0.8	0.048 (108)	0.058 (81)	0.183 (52)
			r>0.9	0.005 (63)	0.032 (49)	0.297 (29)

Interconnectedness of mouse NAP genes, determined at 10%, 5% and 1% FDR, and their human orthologues within various gene networks as compared to randomly sampled genome genes, matched for coding DNA sequence length and network connectivity. We considered gene co-expression networks (GTEx and Brainspan), functional gene networks (MouseNet v2 and PLN; Phenotypic Linkage Network) and a PPI (protein-protein interaction) network (String). Various minimum co-expression coefficient thresholds ($r > 0.6$, $r > 0.7$, $r > 0.8$ and $r > 0.9$) were tested. Similarly, we also considered all links from the PLN as well as the top 25%, 10% and 1% strongest links. MouseNet v2 uses mouse genes, while the remaining networks use human unique 1:1 orthologues (**Table 2.3**). NAP genes, neuroanatomical phenotype genes, defined as those associated with neuroanatomical deviations in knockout compared to wildtype mice (significant at 10%, 5% or 1% FDR).

Table S6.2 | GO enrichment analysis of mouse modules.

Module	GO term	P-value	Adjusted p-value	Number genes
Module 0	postsynaptic density (GO:0014069)	0.00018	0.01761	5
	postsynaptic specialization (GO:0099572)	0.00018	0.01761	5
	synapse part (GO:0044456)	0.00086	0.04089	7
	synaptic membrane (GO:0097060)	0.00074	0.04089	5
	postsynaptic membrane (GO:0045211)	0.00224	0.0854	4
	positive regulation of nervous system development (GO:0051962)	9.36E-05	0.10335	8
	cell cortex part (GO:0044448)	0.0045	0.14328	3
	positive regulation of cell morphogenesis involved in differentiation (GO:0010770)	0.00067	0.36981	4
	regulation of cell morphogenesis involved in differentiation (GO:0010769)	0.00108	0.39673	5
	Golgi subcompartment (GO:0098791)	0.0162	0.4419	3
Module 2	cytoskeleton (GO:0005856)	0.00311	0.29655	11
	nucleolus (GO:0005730)	0.00177	0.29655	8
	ciliary basal body (GO:0036064)	0.00517	0.32894	3
	microtubule (GO:0005874)	0.0076	0.36306	5
	microtubule organizing center part (GO:0044450)	0.0096	0.36679	3
	cytoskeletal protein binding (GO:0008092)	0.00205	0.49042	8
	ciliary part (GO:0044441)	0.01821	0.57979	4
	cellular macromolecular complex assembly (GO:0034622)	0.00166	0.60992	7
	cellular protein complex assembly (GO:0043623)	0.00145	0.60992	5
	microtubule cytoskeleton organization (GO:0000226)	0.00142	0.60992	5
Module 3	Golgi apparatus (GO:0005794)	0.00499	0.17044	8
	integral component of organelle membrane (GO:0031301)	0.00456	0.17044	3
	intrinsic component of organelle membrane (GO:0031300)	0.00502	0.17044	3
	mitochondrial outer membrane (GO:0005741)	0.00465	0.17044	3
	organelle outer membrane (GO:0031968)	0.00603	0.17044	3
	outer membrane (GO:0019867)	0.00625	0.17044	3
	ubiquitin ligase complex (GO:0000151)	0.00448	0.17044	4
	cullin-RING ubiquitin ligase complex (GO:0031461)	0.00801	0.18926	3
	mitochondrial membrane part (GO:0044455)	0.00892	0.18926	3
	microtubule (GO:0005874)	0.0118	0.22545	4

GO terms over-represented at BH $p < 0.1$ or the top 10 terms with smallest p-values for each module. The whole genome was used as background and p-values were derived using right-tailed hypergeometric tests with Benjamini-Hocchberg (BH) corrections for multiple testing to control for the number of terms tested. Modules 1 and 4 were excluded from the table due to the lack of nominal GO enrichments.

Table S6.3 | GO enrichment analysis of human modules.

Module	GO Term	P-value	Adjusted p-value	Number genes
Module 0	postsynaptic density (GO:0014069)	1.28E-05	0.0015	4
	postsynaptic specialization (GO:0099572)	1.28E-05	0.0015	4
	postsynaptic membrane (GO:0045211)	5.57E-05	0.0042	4
	synaptic membrane (GO:0097060)	1.43E-04	0.0081	4
	cell cortex part (GO:0044448)	2.76E-04	0.0125	3
	organelle (GO:0043226)	6.18E-04	0.0233	18
	cytoskeleton (GO:0005856)	8.10E-04	0.0262	5
	membrane (GO:0016020)	0.0018	0.0398	15
	cell projection (GO:0042995)	0.0019	0.0398	6
	cell projection part (GO:0044463)	0.0019	0.0398	5
Module 1	cytoplasmic part (GO:0044444)	3.13E-04	0.0371	10
	synapse (GO:0045202)	3.28E-04	0.0371	3
	bounding membrane of organelle (GO:0098588)	0.0012	0.0867	5
	whole membrane (GO:0098805)	0.0022	0.0867	4
	cytoplasmic vesicle membrane (GO:0030659)	0.0024	0.0867	3
	vesicle membrane (GO:0012506)	0.0026	0.0867	3
	membrane protein complex (GO:0098796)	0.0027	0.0867	4
	neuron part (GO:0097458)	0.0035	0.0874	4
	cytosol (GO:0005829)	0.0038	0.0874	6
	cell projection (GO:0042995)	0.0044	0.0874	4
Module 2	G2/M transition of mitotic cell cycle (GO:0000086)	6.33E-10	4.55E-07	5
	cell cycle G2/M phase transition (GO:0044839)	6.79E-10	4.55E-07	5
	mitotic cell cycle phase transition (GO:0044772)	1.05E-08	4.06E-06	5
	cell cycle phase transition (GO:0044770)	1.21E-08	4.06E-06	5
	single-organism organelle organization (GO:1902589)	3.48E-07	9.32E-05	6
	mitotic cell cycle process (GO:1903047)	3.69E-06	8.25E-04	5
	centrosome (GO:0005813)	8.76E-06	0.002	4
	microtubule organizing center part (GO:0044450)	2.68E-05	0.0025	3
	microtubule organizing center (GO:0005815)	3.29E-05	0.0025	4
	cell cycle process (GO:0022402)	2.15E-05	0.0041	5
Module 3	nucleus (GO:0005634)	1.10E-04	0.0248	10
	cell cycle process (GO:0022402)	4.87E-05	0.0514	6
	regulation of cell division (GO:0051302)	9.82E-05	0.0514	3
	mitotic cell cycle process (GO:1903047)	1.15E-04	0.0514	5
	DNA replication (GO:0006260)	1.66E-04	0.0555	3
Module 4	Golgi apparatus (GO:0005794)	4.10E-04	0.0926	5

Module 5	heterocyclic compound binding (GO:1901363)	1.74E-04	0.0238	16
	organic cyclic compound binding (GO:0097159)	2.09E-04	0.0238	16
	nucleic acid binding (GO:0003676)	2.62E-04	0.0238	13
	poly(A) RNA binding (GO:0044822)	5.70E-04	0.0326	7
	RNA binding (GO:0003723)	5.97E-04	0.0326	8
	RNA metabolic process (GO:0016070)	3.98E-05	0.0533	13
	ribonucleoprotein complex assembly (GO:0022618)	1.06E-04	0.0562	4
	ribonucleoprotein complex subunit organization (GO:0071826)	1.26E-04	0.0562	4
	nucleic acid metabolic process (GO:0090304)	1.73E-04	0.0581	13
	cellular macromolecular complex assembly (GO:0034622)	2.58E-04	0.0693	6
Module 6	Golgi membrane (GO:0000139)	1.31E-05	0.003	4
	Golgi apparatus part (GO:0044431)	4.17E-05	0.0047	4
	bounding membrane of organelle (GO:0098588)	3.62E-04	0.026	4
	vacuolar lumen (GO:0005775)	4.61E-04	0.026	2
	Golgi subcompartment (GO:0098791)	7.87E-04	0.0349	2
	organelle subcompartment (GO:0031984)	9.27E-04	0.0349	2
	organelle membrane (GO:0031090)	0.00172	0.0554	4
Module 7	cytoplasm (GO:0005737)	6.94E-05	0.0157	9
	microtubule binding (GO:0008017)	2.49E-04	0.068	3
	tubulin binding (GO:0015631)	5.69E-04	0.0777	3
	cytoskeletal protein binding (GO:0008092)	0.0011	0.0984	4

Table S6.3 continued

GO terms over-represented at BH $p < 0.1$ or the top 10 terms with smallest p-values for each module. The whole genome was used as background and p-values were derived using right-tailed hypergeometric tests with Benjamini-Hochberg (BH) corrections for multiple testing to control for the number of terms tested.