Comparative approaches to the genetics of human neuropsychiatric disorders



Hyun Ji Noh

Department of Physiology, Anatomy and Genetics
Pembroke College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity Term 2012

Comparative approaches to the genetics of human neuropsychiatric disorders

A thesis submitted for the degree of
Doctor of Philosophy

Hyun Ji Noh, Pembroke College, University of Oxford

Trinity Term 2012

Abstract

In this thesis, I investigate the genetics of neuropsychiatric disorders by analysing large data sets derived from high-throughput experiments, using novel comparative genomics approaches. In the first project, I explore characteristics of rare, *de novo* copy number variants identified among autism patients by employing various bioinformatics resources including Mouse Genome Informatics phenotypes, Gene Ontology terms, and protein-protein interactions. I describe how I objectively identified a number of mouse model phenotypes that are significantly associated with autism, and that provide insight into the aetiologies for both copy number deletions and duplications. In the second project, I investigate the genetics of obsessive-compulsive disorder by resequencing genomic regions of human case-control cohorts and the best spontaneous disease model organisms, namely dogs with canine compulsive disorder, and breed-matched controls. Targeted sequencing experiments yielded a large number of high-quality genetic variants in both humans and dogs. I prioritised variants and genes using case-control comparisons and functional annotations such as types of mutation, evolutionary conservation status and regulatory marks. In turn, I generated several hypotheses that are experimentally tractable. Replication of these findings in a larger cohort is necessary, although it lies beyond the scope of this thesis. Results from both projects indicate that the analytical frameworks employed in this thesis could be profitably applied to other neuropsychiatric disorders.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Chris P. Ponting, for his continuous support, invaluable ideas, time and patience during the course of my DPhil study. I am also grateful to Dr. Caleb Webber for his supervision on the autism project and Prof. Kerstin Lindblad-Toh for her supervision on the obsessive-compulsive disorder project. Their scientific contributions made these research projects possible.

I also thank my colleague scientists in the Ponting group and the Webber group at the University of Oxford for their help and advice. I also would like to thank my collaborators Dr. Ruqi Tang and Dr. Elinor Karlsson at the Broad Institute of Harvard and MIT for stimulating discussions.

Lastly, I owe my deepest gratitude to my family for their moral support and encouragement outside of the office.

Word count

Chapters	*Main*	*Appendix*
1	6,497	NA
2	2,031	NA
3	6,622	1,527
4	7,205	19,319
5	10,297	4,689
6	12,977	2,291
7	865	NA
Total	**46,494**	27,826

Contents

List of Figures **xiii**

List of Tables **xxi**

1 Introduction **1**

1.1 Comparative genomics 2

1.1.1 Sources of variation - the 'raw material' of evolution 3

1.1.2 Strategies and lessons learned from comparative genomics 4

1.1.3 Implications for human disease studies 6

1.2 Genetic architecture of neuropsychiatric disorders 8

1.2.1 Insights from epidemiological data 10

1.2.1.1 Prevalence and age of onset 10

1.2.1.2 Reproductive disadvantage 11

1.2.1.3 Heritability 12

1.2.1.4 Recent origin of susceptibility variants 12

1.2.2 Insights from evolutionary perspectives 13

1.2.2.1 Balancing selection 14

1.2.2.2 Mismatch hypothesis 16

1.2.2.3 Polygenic mutation-selection balance hypothesis 17

1.2.3 Implications for genetic architecture of mental disorders 18

1.2.4 Implications for genetic studies of neuropsychiatric disorders . . 20

1.3 Project rationales and thesis structure 21

1.3.1 ASD project 21

1.3.2 OCD project 22

2 Materials and methods **25**

2.1 Datasets . . . 26

2.1.1 Rare *de novo* copy number variants (CNVs) found in ASD patients 26

2.1.2 Targeted sequence data sets from case-control cohorts of OCD and CCD . . . 27

2.2 Functional genomics resources . . . 28

2.2.1 Mouse Genome Informatics (MGI) . . . 28

2.2.2 Functional attributes and protein interaction . . . 28

2.2.3 Evolutionary conservation status . . . 29

2.2.4 Regulatory marks . . . 29

2.3 Statistical analyses . . . 29

2.3.1 Hypergeometric test . . . 30

2.3.2 Permutation test . . . 30

2.3.3 False discovery rate (FDR) . . . 31

2.4 Frequently used software packages . . . 32

3 Autism-relevant mouse models and phenotypes objectively identified from rare copy number variants **33**

3.1 Abstract . . . 34

3.2 Introduction . . . 35

3.2.1 Autism Spectrum Disorders (ASD) . . . 35

3.2.1.1 Classification and diagnosis . . . 35

3.2.1.2 Comorbidity . . . 36

3.2.1.3 Genetics . . . 38

3.2.2 Mouse as a model for human autism . . . 38

3.2.2.1 Social interaction deficits in mice . . . 39

3.2.2.2 Communication deficits in mice . . . 40

3.2.2.3 Repetitive/restricted behaviours in mice . . . 41

3.3 Materials and Methods . . . 42

3.3.1 *De novo* CNVs identified in individuals with ASD . . . 42

3.3.2 Assigning genes to CNVs . . . 43

3.3.3 Mouse Genome Informatics (MGI) phenotypes . . . 43

3.3.4 Statistical Analysis . . . 44

3.4 Results . . . 46
3.4.1 Identifying mouse model phenotypes associated with ASD *dn* CNVs 46
3.4.1.1 Hypothesis-free approach . . . 46
3.4.1.2 Hypothesis-led approach: examining *Nervous System* phenotypes . . . 51
3.4.2 Comparison with *ASD-implicated* genes . . . 52
3.4.3 Model phenotype associations offer causal hypotheses and aetiological insights . . . 53
3.5 Discussion . . . 54
3.5.1 Similarities between associated mouse phenotypes and human ASD 54
3.5.2 Mouse model phenotypic associations provide insights into the aetiology of ASD . . . 56
3.5.3 Linking mouse genotype to ASD phenotype . . . 57
3.5.4 Post hoc assessment of methodology . . . 58
3.5.4.1 Handling incomplete annotation of mouse phenotype . 58
3.5.4.2 Handling effect of gene and CNV size . . . 58
3.5.4.3 The null of mouse phenotypic association test . . . 59
3.5.4.4 Limitations . . . 59
3.5.5 Conclusions . . . 60
3.6 Appendices . . . 61
3.6.1 Confirming the null . . . 61
3.6.2 ASD *dn* CNV sets . . . 63
3.6.3 Mouse phenotypic enrichments and associated genes . . . 69

4 Functional genomics identifies convergent aetiologies for both deletions and duplications found in ASD 95
4.1 Abstract . . . 96
4.2 Introduction . . . 97
4.2.1 Functional Enrichment Analysis (FEA) . . . 97
4.2.1.1 Current FEA methods . . . 98
4.2.1.2 FEA for studying ASD genetics . . . 101
4.3 Materials and Methods . . . 102
4.3.1 *De novo* CNVs identified in ASD patients . . . 102

4.3.2 Gene Ontology (GO) . . . 102
4.3.3 Protein-Protein Interaction (PPI) network . . . 102
4.3.4 Statistical Analysis . . . 103
4.4 Results . . . 104
4.4.1 Mouse functional genomics offers candidate genes for ASD . . . 104
4.4.2 ASD candidate genes identified via functional mouse genomics form a significantly interconnected PPI network . . . 104
4.4.3 PPI analysis offers additional candidate genes for ASD . . . 107
4.4.4 Separate functional biases for deletions and duplications suggest convergent aetiologies for both directions of copy number change 110
4.5 Discussion . . . 113
4.5.1 *AST-associated* and *AAP-associated* genes offer an extensive interaction hub for ASD . . . 113
4.5.2 Functions of identified candidate genes for ASD . . . 114
4.5.3 *AST* genes yield behavioural phenotypes. . . . 115
4.5.4 Identified interactome is highly connected to two recently reported autism-networks . . . 115
4.5.5 Convergent aetiologies between deletions and duplications in ASD 115
4.5.5.1 Altered presynaptic exocytosis . . . 116
4.5.5.2 Reduced activity of MAPK pathway . . . 118
4.5.5.3 Altered calcium homeostatis . . . 120
4.5.5.4 Up-regulation of *Wnt* signalling . . . 122
4.5.5.5 Crosstalk between signalling cascades . . . 124
4.5.6 Conclusions . . . 126
4.6 Appendices . . . 127
4.6.1 Genes identified via mouse phenotypic enrichments . . . 127
4.6.2 GO functional enrichments . . . 128
4.6.3 REVIGO-summarised GO terms . . . 186
4.6.4 PPI network connectivity to recently published exome studies . . 193

5 Targeted resequencing of human genomic regions for obsessive-compulsive disorder 195

5.1 Abstract . . . 196
5.2 Introduction . . . 197
5.2.1 Obsessive-Compulsive Disorder (OCD) . . . 197
5.2.1.1 Classification and prevalence . . . 197
5.2.1.2 Neurobiology . . . 198
5.2.1.3 Genetics . . . 200
5.3 Materials and Methods . . . 202
5.3.1 DNA preparation and pooling . . . 202
5.3.2 Target selection . . . 202
5.3.3 Sequencing and mapping . . . 203
5.3.4 Variant discovery and allele frequency estimation . . . 203
5.3.5 Pathway analysis . . . 204
5.3.6 Association tests . . . 205
5.4 Results . . . 206
5.4.1 Pooled targeted sequencing yielded large amounts of high-quality sequence alignments . . . 206
5.4.2 SYZYGY identified a number of functional genetic variants and polymorphisms . . . 211
5.4.3 Comparison with genotyping array confirmed the accuracy of allele frequency (AF) estimation . . . 212
5.4.4 Many of the detected variants were rare . . . 213
5.4.5 Case-control comparison of AFs yielded strongly associated variants, genes, and functional themes for OCD . . . 216
5.4.5.1 Strongly associated variants may have functional effects 218
5.4.5.2 Examining burden of multiple variants per gene among cases identified significantly associated genes . . . 220
5.4.5.3 Associated genes were enriched for functions related to synaptic formation and development of the cortex and the basal ganglia . . . 221
5.4.6 Power calculation . . . 225
5.5 Discussion . . . 227

5.5.1 Associated coding variants may be less likely to be causative to OCD . . . 227
5.5.2 A variant in the 3'-UTR of *THBD* and its flanking sequence might serve as a new transcription factor binding site . . . 227
5.5.2.1 *Chr20:23028063 T→C* is uncommon in the general population . . . 228
5.5.2.2 Reference allele T at Chr20:23028063 is evolutionarily conserved . . . 228
5.5.2.3 Its neighbouring region overlaps regulatory marks . . . 228
5.5.2.4 *Chr20:23028063 T→C* may affect transcription factor binding . . . 231
5.5.3 A variant in the 3'-UTR of *CLDN1* might affect functional non-coding RNA . . . 233
5.5.4 Evolutionary conservation status is a good indication to detect functional variants associated with OCD . . . 236
5.5.5 Groupwise association test allowed a prioritisation of candidate genes for OCD . . . 237
5.5.6 Pathway analysis identified functional themes for OCD and prioritised further candidate genes . . . 238
5.5.7 Inclusion of non-coding regulatory region increased statistical power 238
5.5.8 Conclusions . . . 239
5.6 Appendices . . . 240
5.6.1 Handling duplicated reads . . . 240
5.6.2 Population stratification . . . 243
5.6.3 Comparison of AF distribution between cases and controls . . . 246
5.6.4 Q-Q plot of p-values for single variant association test . . . 247
5.6.5 Assessing groupwise association tests . . . 248
5.6.6 Top associated variants . . . 252
5.6.7 Samples and target regions . . . 254
5.6.8 Coverage and experimental quality (per pool statistics) . . . 276
5.6.9 Variant discovery per pool . . . 280
5.6.10 Associated genes from burden test . . . 281
5.6.11 Pathway analyses . . . 300

5.6.12 Power calculation . . . 305
5.6.13 Tracks for regulatory marks . . . 310
5.6.14 Transcription factor binding DNA motif . . . 312
5.6.15 Tracks for evolutionary conservation . . . 312

6 Resequencing of genomic regions for compulsive disorder in dogs 315
6.1 Abstract . . . 316
6.2 Introduction . . . 317
6.2.1 The domestic dog: its unique population history . . . 317
6.2.2 Dog as a model organism for human disease . . . 317
6.2.3 Trait mapping strategies for the dog genome . . . 319
6.2.3.1 Association mapping . . . 320
6.2.3.2 Selection mapping . . . 320
6.2.4 Canine Compulsive Disorder (CCD) . . . 322
6.2.4.1 Face validity of CCD . . . 322
6.2.4.2 Predictive and construct validity of CCD . . . 323
6.2.4.3 Previous genetic mappings for CCD . . . 324
6.3 Materials and Methods . . . 325
6.3.1 Sample collection . . . 325
6.3.2 Target resequencing and raw data processing . . . 325
6.3.3 SNPs and micro-indels discovery and genotyping . . . 326
6.3.4 Structural variant discovery . . . 326
6.3.5 Phenotype correlation analyses . . . 328
6.3.5.1 SEQscoring for individual variants . . . 328
6.3.5.2 Haplotype Likelihood Ratio (HLR) test for regions . . . 329
6.4 Results . . . 331
6.4.1 Targeted resequencing of 5Mb canine genomic regions yielded extensive high-quality alignments . . . 331
6.4.2 A large number of SNPs and micro-indels including functional variants were detected via GATK . . . 333
6.4.3 Replication of variant calls and the Ti/Tv ratio . . . 334
6.4.4 Detected variants were prioritised by case-control comparison and using a known risk haplotype . . . 336

6.4.4.1 More functional variants were exclusively observed among cases . . . 336
6.4.4.2 The number of case-only variants within breed is consistent with the breed genetic structure . . . 341
6.4.4.3 Many variants showed phenotypic correlations . . . 343
6.4.4.4 Resequencing allowed the pinpointing of candidate variants within *CDH2*, the only gene with a genome-wide significant association . . . 346
6.4.5 GenomeSTRiP reveals a number of CNV deletions . . . 350
6.4.6 Canine genes with excessive case-only variants are orthologous or functionally related to human candidate genes for OCD . . . 353
6.5 Discussion . . . 357
6.5.1 Genetic structure of breeds may have affected number of identified variants . . . 357
6.5.2 Chr7:61693835 and its flanking region may have regulatory role for CCD . . . 359
6.5.3 A 1.2kb deletion of an exon in *PGCP* may play a role in CCD . . . 364
6.5.4 Candidate genes may play a role in synaptic formation, and may function in concert . . . 367
6.5.5 Perturbed synaptic formation may be a common mechanism for compulsive disorder in both humans and dogs . . . 369
6.5.6 Conclusions . . . 371
6.6 Appendices . . . 372
6.6.1 Study design . . . 372
6.6.2 Sequencing statistics . . . 378
6.6.3 Variant statistics . . . 387
6.6.4 Prioritising variants . . . 390
6.6.5 Batch effect in calling CNVs . . . 393
6.6.6 Case- and control-only variants within genes . . . 394
6.6.7 Identifying transcription factor binding DNA motifs . . . 405

7 Conclusions and future perspectives **407**

7.1 ASD project 408

7.2 OCD project 409

References **411**

List of Figures

1.1 Comparison of genomes at different phylogenetic distances can address different questions 5
1.2 Using constraint information to identify candidate mutations for disease 7
1.3 Heterozygote advantage in sickle cell trait 15
1.4 Architecture of variable genetic aetiology of mental disorders inferred from epidemiological data and evolutionary theories 19

3.1 Three core domains of severe dysfunction in ASD in the context of normal population variation 36
3.2 Flow diagram of ASD *dn* CNV analysis 45
3.3 Enrichments of phenotypic categories among AGP ASD *dn* CNVs and AGP control CNVs 47
3.4 Relationships of mouse model phenotypic terms enriched among ASD *dn* CNV genes. 49
3.5 Enriched cochlear and mechanoreception-associated mouse model phenotypes among duplication AGP ASD *dn* CNV genes. 50
3.6 Enrichments of the *abnormal CNS synaptic transmission* phenotypes . . 51
3.7 Distributions of p-values for the surfeits of genes associated with 3 phenotypic categories obtained from 500 randomised sets of case-matched CNVRs 62

4.1 Protein-protein interaction network among *AST* and *AAP* genes 106
4.2 An extensive *AAP-associated* protein-protein interaction network 109
4.3 Copy number changes in candidate genes including *STX1A* and its protein interacting partners implicate presynaptic exocytosis in ASD 117

4.4 Copy number changes of candidate genes within the MAPK pathway implicate a reduced activation of MAPK-dependent transcription in ASD 119
4.5 Copy number changes of candidate genes related to IP3 production implicates altered calcium homeostasis in ASD 121
4.6 Copy number changes of candidate genes within *Wnt* signalling pathway implicates its constitutive activation in ASD 123
4.7 Copy number changes in the ASD protein-protein interaction network implicate MAPK pathway, *Wnt*/β-catenin signalling pathway and altered calcium homeostasis in ASD. 125
4.8 REVIGO-summarised Biological Process GO terms enriched within *AST-associated* network genes 187
4.9 REVIGO-summarised Cellular Component GO terms enriched within *AST-associated* network genes 188
4.10 REVIGO-summarised Molecular Function GO terms enriched within *AST-associated* network genes 189
4.11 REVIGO-summarised Biological Process GO terms enriched within *AAP-associated* network genes 190
4.12 REVIGO-summarised Cellular Component GO terms enriched within *AAP-associated* network genes 191
4.13 REVIGO-summarised Molecular Function GO terms enriched within *AAP-associated* network genes 192

5.1 Diagram of the cortico-striatal-thalamo-cortical (CSTC) circuit 199
5.2 Analysis steps and relevant file formats for next generation sequencing (NGS) data 204
5.3 Experimental design for targeted pooled resequencing 207
5.4 Heatmap for read depth of target region per pool 210
5.5 Comparison of AFs between pooled sequencing data and SNP array data 212
5.6 Distribution of coding and non-coding variants across different AFs . . . 214
5.7 Distribution of non-coding variants within evolutionarily conserved elements across different AFs 215
5.8 Distribution of the association levels across different AFs 217
5.9 The most strongly associated variants are non-coding 218

5.10 GO FAT terms that are significantly associated (FDR <5%) with 123 investigative genes and their fold-enrichments 223
5.11 Significantly enriched GO FAT terms and the ratio of the fold-change in 123 investigative genes to the fold-change in 611 resequenced genes . . . 224
5.12 Quantile-quantile (Q-Q) plots for groupwise association test in sub-sampled data 226
5.13 UCSC tracks showing a variant *Chr20:23028063* $T \rightarrow C$ and its overlapping gene (hg19) 228
5.14 Chr20:23028063 and its flanking region with regulatory marks 229
5.15 Brain histone H3K4Me3 ChIP-seq marks in Chr20:23028063 and its flanking region 230
5.16 Transcription factor binding DNA motifs detected for Chr20:23028063 and its neighbouring region 231
5.17 UCSC tracks showing a non-coding variant *Chr3:190025964* $C \rightarrow T$ and its overlapping gene in hg19 233
5.18 A non-coding RNA transcript overlaps Chr3:190025964 position 234
5.19 Correlation between mean target coverage and duplicate rates 241
5.20 Overview of GATK data processing pipeline 242
5.21 A multi-dimensional scaling (MDS) plot of samples in the OCD cohort . 245
5.22 Q-Q plot of AF distribution in cases and controls 246
5.23 Q-Q plot for the distribution of p-values from single variant association test 247
5.24 Q-Q plot for the distribution of p-values from three different groupwise association tests 250
5.25 Q-Q plot for the distribution of $-log_{10}P$ from burden test on genic regions and their 50kb up- & down- stream regions 251
5.26 Chr20:23028063 position is evolutionarily conserved 252
5.27 Reference allele at Chr3:190025964 is evolutionarily conserved 253
5.28 Number of total reads per pool 276
5.29 Number of unique reads per pool 277
5.30 Percentage of unique reads to total reads per pool 277
5.31 Mean target coverage per pool 278
5.32 Percent target bases with >20x read depth 278

5.33 Fraction of detected singletons (per pool) to total number of discovered variants 279
5.34 Fraction of known variants from the 1000 genome project to the total number of discovered variants per pool 280
5.35 Number of variants found per gene and its size for all 611 genes 281
5.36 Associated genes from burden test (all variants) 282
5.37 Associated genes from burden test (nonsynonymous variants) 283
5.38 Associated genes from burden test (coding variants) 284
5.39 Associated genes from burden test (variants in evolutionarily conserved region) 285
5.40 Associated genes from burden test (nonsynonymous variants in evolutionarily conserved region) 286
5.41 Associated genes from burden test (MAF< 10%) 287
5.42 Associated genes from burden test (MAF< 1%) 288
5.43 Associated genes from burden test (MAF<10% in conserved region) . . 289
5.44 Associated genes from burden test (MAF<1% in conserved region) . . . 290
5.45 Associated genes from burden test (all variants, adjusted for gene size) . 291
5.46 Associated genes from burden test (nonsynonymous variants, adjusted for gene size) 292
5.47 Associated genes from burden test (coding variants, adjusted for gene size) 293
5.48 Associated genes from burden test (variants in evolutionarily conserved region, adjusted for gene size) 294
5.49 Associated genes from burden test (nonsynonymous variants in evolutionarily conserved region, adjusted for gene size) 295
5.50 Associated genes from burden test (MAF<10%, adjusted for gene size) . 296
5.51 Associated genes from burden test (MAF<1%, adjusted for gene size) . 297
5.52 Associated genes from burden test (MAF<10% in conserved region, adjusted for gene size) 298
5.53 Associated genes from burden test (MAF< 1% in conserved region, adjusted for gene size) 299
5.54 REVIGO-summary of the significantly associated (FDR <5%) Biological Process GO FAT terms among investigative genes 303

5.55 REVIGO-summary of the significantly associated ($FDR < 5\%$) Cellular Component GO FAT terms among investigative genes 304
5.56 Q-Q plot for original dataset and doubled dataset 305
5.57 Q-Q plots for original dataset and permuted datasets where the sample size matches to the original data 306
5.58 Q-Q plots for sub-sampled datasets consisting of 96 individuals for each case and control from the original dataset 307
5.59 Q-Q plots for sub-sampled datasets consisting of 64 individuals for each case and control from the original dataset 308
5.60 Q-Q plots for sub-sampled datasets consisting of 48 individuals for each case and control from the original dataset 309
5.61 The CCAAT enhancer binding protein gamma (CEBPG) motif 312
5.62 Phylogeny of the 46 species in the Vertebrate Multiz Alignment and Conservation track 313

6.1 Population history and haplotype structure of *Canis familiaris* 318
6.2 Two-stage trait mapping strategy in dogs 321
6.3 Analytical framework for Genome STRucture in Populations (GenomeSTRiP) 327
6.4 Overview of SEQscoring modules 329
6.5 Sequenced samples and depth of coverage 332
6.6 Histogram of detected variants and conserved elements in target regions 335
6.7 Phenotypic correlation scoring scheme for dominant/complex traits and histogram of phenotype scores for the detected variants 344
6.8 Genotypes and HLR scores within *CDH2* among Doberman Pinschers . 348
6.9 Genotypes and HLR scores within a risk haplotype in *CDH2* 349
6.10 Size distribution of deletions identified among sequenced dogs 350
6.11 Copy number estimation of a deletion at Chr29:44178339-44179516 . . . 351
6.12 Pileup data in the region of 1.2kb deletion at Chr29 352
6.13 Exon deleting CNV in *PGCP* 353
6.14 Gene plot based on the number of genic case- and control-only variants across all four breeds 355
6.15 Gene plot based on the number of genic case- and control-only variants within conserved regions 356

6.16 Variants *Chr7:61693835* $T \rightarrow C$ and *Chr7:61693855* $A \rightarrow T$, and their neighbouring regions lifted on to the human genome hg19 361
6.17 Transcription factor binding sites mapped to Chr7:61693835 region by TRANSFAC® 362
6.18 Two CCD candidate regions in canine chromosome 7 363
6.19 Association signal with CCD in *PGCP* region 365
6.20 Expression profile of PGCP in the brain 366
6.21 A schematic cartoon of the cadherin-catenin complex in synapse assembly 368
6.22 GWAS for CCD in Doberman Pinschers 373
6.23 A schematic cartoon illustrating a genome with deletion 376
6.24 The difference occurred between inferred insert size and expected insert size due to a deletion in a subject genome 376
6.25 Workflow of modules in GenomeSTRiP 377
6.26 Variant types per pool 387
6.27 Variant novelty per pool 388
6.28 Functional variants per pool 388
6.29 Ti/Tv ratio per pool 389
6.30 Copy number deletions per pool 389
6.31 Variants plotted for conservation status and pattern score 390
6.32 Variant statistics within *CDH2* 391
6.33 Coding and UTR changes in *CDH2* region 392
6.34 Batch effect in calling CNVs 1 393
6.35 Batch effect in calling CNVs 2 393
6.36 Case- and control-only variants in gene across all breeds excluding Doberman Pinschers 394
6.37 Case- and control-only variants in conserved regions of gene across all breeds excluding Doberman Pinschers 395
6.38 Case- and control-only variants in gene for Doberman Pinschers 396
6.39 Case- and control-only variants in conserved regions of gene for Doberman Pinschers 397
6.40 Case- and control-only variants in gene for German Shepherds 398
6.41 Case- and control-only variants in conserved regions of gene for German Shepherds 399

6.42 Case- and control-only variants in gene for Jack Russell Terriers 400
6.43 Case- and control-only variants in conserved regions of gene for Jack Russell Terriers 401
6.44 Case- and control-only variants in gene for Shetland Sheepdogs 402
6.45 Case- and control-only variants in conserved regions of gene for Shetland Sheepdogs 403

LIST OF FIGURES

List of Tables

1.1 Genetic epidemiology of six representative neuropsychiatric disorders . . 9

3.1 Domains of dysfunctions in autism spectrum disorders (ASD) 37
3.2 Genomic extent and ENSEMBL gene content for ASD-associated *de novo* CNVs and AGP control CNVs. 42
3.3 Overlapping genomic regions of ASD *dn* CNV sets 43
3.4 73 *De novo* CNVs acquired from the AGP consortium (hg18) 63
3.5 28 *De novo* CNVs acquired from Marshall *et al.* (hg18) 64
3.6 94 *De novo* CNVs acquired from Levy *et al.* (hg18) 65
3.7 67 *De novo* CNVs acquired from Sanders *et al.* (hg18) 67
3.8 Phenotypic category enrichments for AGP ASD *dn* CNVs and control CNVs 69
3.9 Significantly enriched mouse model behavioural phenotypes among the orthologues of AGP *de novo* CNVs genes 70
3.10 Significantly enriched mouse model hearing/vestibular/ear phenotypes among the orthologues of AGP *de novo* CNV genes 71
3.11 Significantly enriched mouse model nervous phenotypes among the orthologues of AGP *de novo* CNV genes 74
3.12 Candidate genes from AGP *de novo* CNVs 76
3.13 AGP *de novo* CNVs and number of candidate genes in each CNV 77
3.14 Significantly enriched mouse model nervous phenotypes among the orthologues of Marshall *et al. de novo* CNV genes. 78
3.15 Candidate genes from Marshall *et al. de novo* CNVs 79
3.16 Marshall *et al. de novo* CNVs and number of candidate genes in each CNV 80

3.17 Nominally significantly enriched mouse model behavioural phenotypes among the orthologues of Levy *et al. de novo* CNV genes 81
3.18 Significantly enriched mouse model nervous phenotypes among the orthologues of Levy *et al. de novo* CNV genes 82
3.19 Candidate genes from Levy *et al. de novo* CNVs 84
3.20 Levy *et al. de novo* CNVs and number of candidate genes in each CNV 85
3.21 Nominally significantly enriched mouse model behavioural phenotypes among the orthologues of Sanders *et al. de novo* CNV genes 86
3.22 Nominally significantly enriched mouse model nervous phenotypes among the orthologues of Sanders *et al. de novo* CNV genes 87
3.23 Candidate genes from Sanders *et al. de novo* CNVs 88
3.24 Sanders *et al. de novo* CNVs and number of candidate genes in each CNV 89
3.25 Twenty two human genes deemed previously to have been implicated in ASD 90
3.26 Significantly enriched mouse model phenotype categories among *ASD-implicated* genes 91
3.27 Significantly enriched mouse model nervous phenotypes among *ASD-implicated* genes 91
3.28 Significantly enriched mouse model behavioural phenotypes among *ASD-implicated* genes 93
3.29 Significantly enriched mouse model behavioural phenotypes among *ASD-implicated* genes 94

4.1 PPIs among protein products of *AST* and *AAP* genes 105
4.2 PPIs among proteins of *AST* and *AAP* genes and their directly interacting partners encoded by ASD *dn* CNV genes 107
4.3 Enrichment of GO regulation of biological processes among *AST-associated* and *AAP-associated* genes 110
4.4 Candidate gene sets derived from ASD *dn* CNVs via mouse functional genomics described in Chapter 3 127
4.5 GO terms significantly enriched among *AST-associated* genes 128
4.6 GO terms significantly enriched among *AST-associated* genes from duplication CNVs 147

4.7 GO terms significantly enriched among *AST-associated* genes from deletion CNVs . . . 151
4.8 GO terms significantly enriched among *AAP-associated* genes . . . 157
4.9 GO terms significantly enriched among *AAP-associated* genes from duplication CNVs . . . 176
4.10 GO terms significantly enriched among *AAP-associated* genes from deletion CNVs . . . 180
4.11 PPI network connectivity between this study and two recently published exome-sequencing studies for ASD . . . 193

5.1 Sequencing statistics showing read length and number of high quality (HQ) aligned reads and bases for each pool . . . 208
5.2 Target coverage information per pool . . . 209
5.3 Functional annotations of detected variants . . . 211
5.4 Top associated coding variants . . . 219
5.5 Top ten non-coding variants in conserved region . . . 220
5.6 Proportion of strongly associated ($p < 0.01$) genes to all associated genes for each test variant type . . . 236
5.7 Number of associated genes ($p < 0.05$) from three different gene-based association tests . . . 248
5.8 Overlap of associated genes among three association tests . . . 249
5.9 List of targeted genes . . . 254
5.10 Case cohort . . . 267
5.11 Control cohort . . . 271
5.12 43 GO FAT terms that are significantly enriched (FDR< 5%) among 123 investigative genes from burden test . . . 301
5.13 40 GO FAT terms that are significantly enriched (FDR< 5%) both in 123 investigative genes and 611 target genes . . . 302
5.14 Cell line information assayed in the ENCODE project . . . 311

6.1 Human neuropsychiatric disorders and its canine counterparts . . . 322
6.2 Colour-codes for detected variants based on their evolutionary conservation and genotype . . . 330

6.3 Counts for detected variants in terms of mutation types and evolutionary conservation . . . 334
6.4 Case unique variants . . . 338
6.5 Case unique SNPs . . . 339
6.6 Case unique micro-indels . . . 339
6.7 The number of case- or control-only variants found in each dog . . . 340
6.8 Heterozygosity of each breed, number of case-control pairs, and number of breed-specific case unique variants . . . 341
6.9 Number of case-control pairs and case-only variants across 2 breeds . . . 342
6.10 Case only variants across three breeds . . . 343
6.11 Top 16 variants with high pattern scores ($\geq$60) . . . 345
6.12 Variants in a haplotype within *CDH2* among Doberman Pinschers . . . 347
6.13 Average membership coefficient of breeds to genetic clusters inferred from 85 domestic dog breeds . . . 358
6.14 Candidate genes identified for compulsive disorder in both human and dog 370
6.15 Sample information . . . 372
6.16 P-values for top GWA-intervals . . . 374
6.17 Target regions for resequencing . . . 375
6.18 Functional variant counts per sample . . . 378
6.19 Sequencing statistics - numbers of aligned reads . . . 383
6.20 Sequencing statistics - read depth of coverage . . . 384
6.21 Concordance between sequencing and SNP array data . . . 385
6.22 Ti/Tv ratio per individual dog . . . 386
6.23 Number of case- and control-only variants in conserved regions and their case to control ratio . . . 404

1

Introduction

A major challenge in biomedical research is to understand the genetic basis of human disease. While model organisms have proved a key experimental tool for studying human biology and disease, the availability of the whole genome sequences of humans and model organisms have immediately allowed links to be forged between clinical and basic research conducted on model organisms to humans, via a comparison of their genomes; there are concepts that shall be used in Chapters 3 and 4. Moreover, comparative genomics is a powerful tool to discover functional sequences in the human genome that can be used to interpret disease-associated variants (Chapters 5 and 6).

In this chapter, I will first introduce this relatively recent concept of comparative genomics, discussing the lessons learnt from concurrent research approaches that are providing important insights into the function of the human genome. Then, I will introduce an emerging view of the genetic architecture of human neuropsychiatric disorders [within which novel comparative genomics approaches were applied in this thesis] on the basis of epidemiological data and evolution theories. Finally, I will describe the rationales of the two main approaches presented in this thesis and present an overview of the following chapters.

1.1 Comparative genomics

A key goal in genomics research is to identify all functional sequences and their roles. An analysis of a single genome provides substantial insights into genome structure, but little insight into genome function (1)(2). Recent technical and analytical advances in DNA sequencing have made it possible to generate and compare the genomes of many species (3), thereby providing unprecedented tools for the functional annotation of genomes. The basic principle of comparative genomics used for such an achievement was simple: shared features between different organisms would be encoded within the DNA sequences that are conserved through vertical descent between the organisms. Based on this principle, comparative genomics takes advantage of information provided by signatures of selection present in the genomes. For instance, purifying (negative) selection, in which deleterious mutations are preferentially removed from a population, results in sequences changing more slowly throughout evolution, while Darwinian (positive) selection, in which beneficial mutations are preferentially retained

in a population, results in sequences changing more rapidly. The remaining nonfunctional sequences change at neutral rates that vary across the genome, within species, and along branches of the phylogenetic tree (reviewed in (4)). Comparing sequences of different genomes can identify such evolutionary signatures, thereby permitting the inference of functional sequences. These predictions are confirmed or complemented by large-scale experiments such as the Encyclopedia of DNA Elements (ENCODE) project (5)(6).

In this section, I will begin by introducing the sources of variation that are required as substrates for evolutionary selection. Subsequently, I will discuss how to choose an appropriate model organism to compare with human in order to address different types of questions, and what is known about the human genome. Lastly, I will discuss applications of comparative genomics to studying human diseases.

1.1.1 Sources of variation - the 'raw material' of evolution

For any evolutionary changes to take place, some level of genetic variation must exist in the population. The sources of genetic variation can be either internal or external to the population. Genetic variations from external sources may be introduced into the population by: *(i)* new alleles or genes being transferred from other populations of the same species via migration or dispersion of seeds, spores, or pollen (*gene flow*), *(ii)* related species being hybridised via repeated backcrossing of an interspecific hybrid with one of its parent species; in this way some portion of the genome of one species becomes incorporated into the genome of another, while the two parental species remain distinct (*introgression, also known as introgressive hybridisation*), or *(iii)* genetic materials being exchanged among distantly related species via cross-infection of retroviruses, which can incorporate host DNA into its own genome then transfer the host DNA into a new host of a different species (*horizontal gene transfer*).

Nevertheless, a major source of genetic variation arises internally to a population. Such genetic variations may result from spontaneous or induced (by mutagen) failures of the biological machinery that is responsible for DNA replication, recombination, or DNA repair, and/or from a transposable element (TE), i.e. a DNA stretch that has an ability to change its location within a genome. The mutations that occur in the germ line may then be transmitted to its offspring, providing much of the 'raw material' for evolutionary selection to then act upon. The types of resulting mutations

include small-scale variations such as point mutations, insertions, and deletions, and larger structural variations such as gene duplications and deletions, and chromosomal inversions and translocations. Once these novel genetic variations are introduced into the population, the fate of these novelties may be determined by selective pressures that act to remove (purifying selection) or to retain (positive selection) them, or by stochastic effects (specifically, genetic drift).

1.1.2 Strategies and lessons learned from comparative genomics

Different types of functional sequences may be found by comparing genomes at different phylogenetic distances (Figure 1.1). Genomic comparisons at very long phylogenetic distances may provide insights about the protein-coding genes, particularly the core set of proteins that is essential to make up an organism. Over the course of long evolutionary time, genomic regions for vital proteins are likely to remain unchanged, while the order of genes and regulatory sequences may not be conserved. Thus, any genomic regions preserved over such a long evolutionary distance are likely to be vital to the organisms.

However, such high degrees of divergence also indicate that functional sequences that are class-specific (e.g. mammal-specific regulatory sequences) cannot be identified. A way to account for such sequences is to compare related genomes at moderate phylogenetic distances, e.g. human and mouse or dog (∼90 million years of divergence from their common ancestors (7)). At such phylogenetic distances, both functional and nonfunctional sequences may be preserved. However, functional sequences will be signified by purifying selection, showing less change than nonfunctional or neutral sequences (8). These identified functional sequences then allow functional classes of DNA, such as coding exons, noncoding RNAs, and gene regulatory motifs to be identified. Comparisons of very similar genomes (e.g. human and chimpanzee, ∼6.5 million years of divergence) are appropriate for detecting functional sequences that account for the differences between the organisms. Such sequences are under positive selection, changing more rapidly compared to other sequences. However, detecting sequences that are truly under positive selection is problematic, as most nucleotide change is neutral and some may appear adaptive but is actually due to biased gene conversion (9)(10)(11).

The first comparative analysis of the human genome was only reported about a decade ago: a comparison with the mouse genome (13). This work provided important

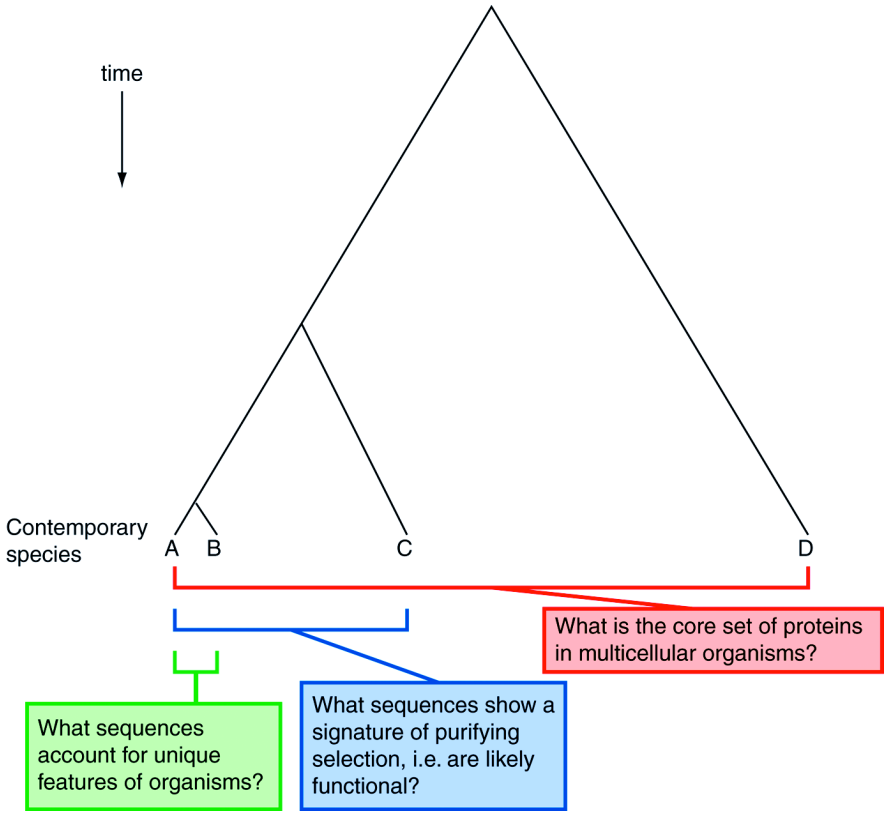


Figure 1.1: Comparison of genomes at different phylogenetic distances can address different questions - Comparing the genomes of closely related species (A & B) may be appropriate to find sequences that account for unique features of organisms. Comparison of genomes at moderate phylogenetic distances (A & C) may allow functional sequences to be identified that show signatures of purifying selection. Genomic comparisons at very long phylogenetic distances (A & D) may reveal the core set of proteins essential for multicellular organisms. Adapted from (12).

insights into the structure, function, and evolution of the human genome, as well as a powerful tool to directly link the results of extensive biomedical researches conducted on mice to humans (13). Since then, many closely and distantly related organisms have been sequenced and analysed. Most recently, sequencing and comparison of 29 mammalian genomes were reported, which substantially enhanced the resolution of evolutionary constraints in the human genome (3). It has been shown that the genomes of most mammals including human, mouse, and dog have roughly the same size (i.e. ∼3 billion base pairs) and a similar numbers of genes (i.e. ∼20,000-25,000 protein-coding genes). Furthermore, there is essentially a 1:1 correspondence between most protein-coding genes in any two mammalian species. At genome-wide level, ∼6.5%-10% of the human genome has been estimated to have undergone purifying selection (14), and hence is likely to be functional. Of these, ∼30% are associated with protein-coding transcripts, ∼27% overlap chromatin states, ∼1.5% are predicted to have RNA structures, and ∼3% may be conserved regulatory motif instances, leaving ∼40% of constrained elements still unexplained to date (3).

1.1.3 Implications for human disease studies

One apparent application of comparative genomics to human disease study is the mapping all genetic, experimental findings from different organisms onto the human genome. An ultimate approach to ascribe a function to a gene is the manipulation of genes in intact animals. The International Knockout Mouse Consortium (IKMC) (15) is one such large-scale, ongoing effort, aiming to knockout all 21,000 mouse genes of which ∼40% have been already knocked out (16). The completion of such a resource will significantly benefit studies of gene activity in human disease models. In parallel, the Mouse Genome Database (MGD) integrates current genetic, genomic and phenotypic information about laboratory mice (17). MGD permits scientists to survey functions of candidate genes for human disease using 1:1 orthologous relationships between human and mouse genes. I applied this concept to identify candidate genes and mouse models for autism in Chapter 3. This was followed by an identification and a functional characterisation of further candidate genes using various functional genomics resources in Chapter 4.

While protein-coding portion of the genome has been intensely explored for their functional roles, only ∼1.2% of the human genome encodes protein sequence. How-

ever, almost 10% of the human genome has been under purifying selection, and hence probably functional (14). Identifying such non-coding portion of functional sequences is particularly relevant to biomedical research, given that the variants identified in genome-wide association studies (GWASs) often lie in non-coding sequence (18). In fact, diseases-associated genetic variants are enriched for constrained regions compared to common genetic polymorphisms (3). This implies that constrained elements identified from comparative analysis perhaps provides a powerful lens with which to focus on causative variants among several disease-associated variants that are merely in linkage disequilibrium (Figure 1.2). Accordingly, I employed constraint information of genomic regions, together with other experimental resources indicating functional regulatory sequence (e.g. chromatin status and DNase1 hypersensitivity sites), in order to prioritise variants identified from case-control resequencing data in Chapters 5 and 6.

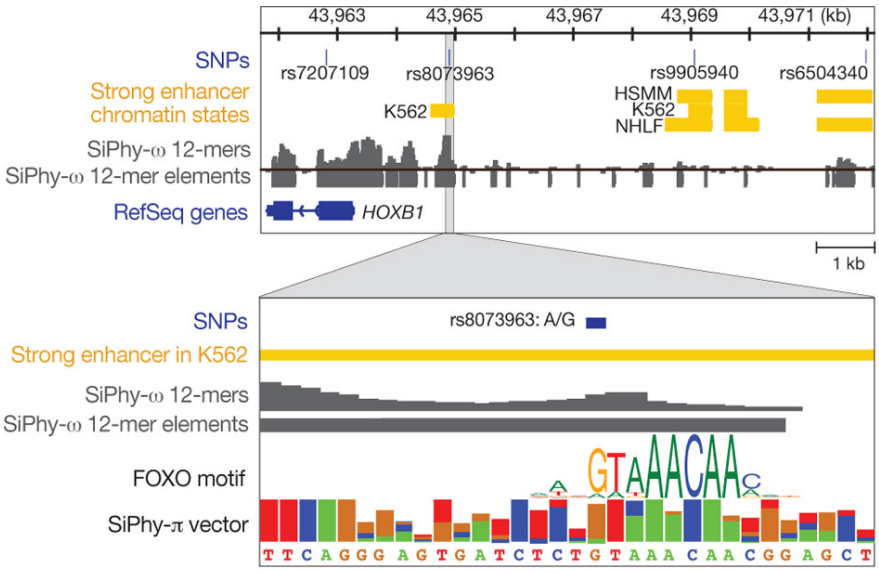


Figure 1.2: Using constraint information to identify candidate mutations for disease - Conservation status measured from the comparison of multiple genomes may be informative in identifying causative mutations. Among multiple variants that are associated with a disease, ones that disrupt conserved functional elements are likely to have deleterious effects. In this example, the reported SNP (rs6504340) for tooth development phenotypes is not in a conserved region, while a linked intergenic SNP (rs8073963) sits in a strongly conserved region. This intergenic SNP also disrupts a consensus Forkhead-Family (FOXO) motif instance in a strong enhancer, thus becoming a promising candidate mutation to follow-up.

1.2 Genetic architecture of neuropsychiatric disorders

Strong heritability of neuropsychiatric disorders implies a large contribution of genetic factors to disease aetiology. For several decades, the common disease/common variant (CDCV) hypothesis has been the most popular theory for explaining the genetics of common diseases including neuropsychiatric disorders. This predicts that common genetic variants of mild effect size explain a majority of the risk for common diseases (19)(20)(21). Genome-wide association study (GWAS), whose success heavily relies on the CDCV hypothesis, has brought substantial advances in understanding the genetic basis of many common phenotypes, mainly physical disorders (22). On the other hand, such an approach in neuropsychiatric disorders has been less successful, given that most GWASs produced unreplicated associations or associations with very small effect size that explains only a small proportion of the variance (23)(24). Several factors have been attributed to this, so-called, "missing heritability": *(i)* an initial overestimation of the contribution of the genetic component, *(ii)* involvement of hundreds to thousands of common risk variants with extremely small effect sizes, and *(iii)* a considerable role for rare alleles in disease aetiology.

Indeed, heritability estimates among many neuropsychiatric disorders may have been inflated, as a twin-based study, the most commonly used study design for heritability estimation, cannot distinguish genetic effects from shared familial risks such as poverty, parenting and migration. Comparisons with adoption studies resolve this problem, but such data are difficult to collect. Nonetheless, heritability estimates from large family and adoption studies in a couple of mental diseases are available, in which their heritability has shown to be lower than those from twin studies, but were still substantial (25). In terms of allelic effect size, an overestimation of allelic effect size has led to severely underpowered cohort sizes in initial GWASs (18), which would then be susceptible to cryptic differences between case and control cohorts. This might be a partial reason why identification and/or replication of genetic associations for neuropsychiatric disorders have been difficult. More recently, an important role of rare variants in neuropsychiatric disorders has emerged. Rare point mutations, deletions, duplications, and larger chromosomal abnormalities appear to be a considerable proportion of autism and schizophrenia causes (26). In addition, patients with autism

have shown to carry excessive rare copy number variants (CNVs) compared to control populations (27)(28)(29).

The notion that rare mutations may underlie common diseases such as neuropsychiatric disorders may appear counterintuitive at first. Furthermore, their strong heritability is even more perplexing, given that patients usually have severe reproductive deficits (Table 1.1, see section 1.2.1.2). When genetic variants are associated with a severly 'fitness-reducing' trait, they are subject to strong purifying selection, hence are preferentially purged from the population (see sections 1.1 and 1.2.2). Nevertheless, neuropsychiatric disorders remain common.

In this section, I will explore two major problems in genetic studies of neuropsychiatric disease: *(i)* why it is difficult to produce replicated genetic associations and *(ii)* the evolutionary mechanisms that could have left harmful, yet strongly heritable traits, in the population. Six representative disorders that cover the spectrum of reproductive disadvantage and severity [as a measure of 'fitness' in evolution] are considered to address these problems, namely, schizophrenia, bipolar affective disorder, unipolar depression, anxiety disorders, anorexia nervosa and autism, of which epidemiological data are relatively comprehensive (reviewed in (24)). I will then introduce an evolution-informed model for the genetic architecture of neuropsychiatric disorders, and discuss where my approaches to the genetics of neuropsychiatric disorders sit within this framework.

Disorders	Prevalence (%)	Age onset	Mortality	Fertility	Heritability	Paternal age efect
Autism	0.3	1	2	0.05	0.9	1.4
Anorexia	0.6	15	6.2	0.33	0.56	-
Schizophrenia	0.7	22	2.6	0.4	0.81	1.4
Bipolar	1.25	25	2	0.65	0.85	1.2
Unipolar	10.22	32	1.8	0.9	0.37	1
Anxiety	28.8	11	1.2	0.9	0.32	-

Table 1.1: Genetic epidemiology of six representative neuropsychiatric disorders - The disorders column shows the six representative neuropsychiatric disorders, namely, autism, anorexia nervosa, schizophrenia, bipolar affective disorder, unipolar depression, and anxiety disorders; Lifetime prevalence is shown as a percentage; median age of onset is in years; mortality ratios of >1 indicate increased mortality compared to the general population; fertility ratios of <1 indicate decreased fertility compared to the general population; all heritabilities shown here are based on twin studies for consistency; paternal age effect is shown as a risk ratio for the 10-year increase in a father's age above 30 years, where available. Adapted from (24).

1.2.1 Insights from epidemiological data

1.2.1.1 Prevalence and age of onset

As life expectancy and reproduction age in humans have significntly increased during a relatively short period of time, genetic variants underlying diseases with late age of onset, such as cardiovascular diseases and type II diabetes, probably have not affected reproductive fitness of the human population in the past, and thus have been retained common in current human population. However, given that almost half of all humans exhibit a mental illness at least once during their life, and half of these cases develop before the age of 14 (30), such a scenario cannot explain the high prevalence of neuropsychiatric disorders in modern humans.

Despite their general pattern of high prevalence and early age of onset, some variations exist among different neuropsychiatric disorders. Autism is rarer than other psychiatric illness (strict autism affects 0.03% of the human population and autism spectrum disorder affects ∼0.6%) and develops in a patient's infancy (31). Anorexia nervosa is also uncommon (0.6% prevalence) and has an average onset age of 15 (32). Schizophrenia and bipolar affective disorder have medium levels of prevalence (0.7% and 1.25%, respectively), and mostly develop at ages of 22-25, affecting the typical period of reproduction (33)(34). Unipolar depression is common, affecting >10% of the population and has a varied age of onset with a median at 32 years (35). Anxiety disorders are very common, prevalent in almost 30% of the population, and the median age of onset is 15 (36).

Importantly, prevalence of some of these disorders varies across different geographic regions and cultures. For instance, the prevalence of unipolar depression differs >10 times between different countries and cultures (35)(37)(38). Incidence of anorexia nervosa shows the most striking geographical and cultural differences, highly susceptible to female gender, caucasian ethnicity, and western culture (32)(39). The prevalence of schizophrenia or bipolar disorders is relatively constant, with mild variations that may depend on urban environment and ethnic minority status (33)(40).

In summary, neuropsychiatric disorders are common and their onset is often early, affecting people in or before their reproductive age. Occurrence of some mental disorders, however, varies across different locations.

1.2.1.2 Reproductive disadvantage

During evolution, selective pressure on a trait depends on its capability to enhance fitness for the organism's survival and reproduction. Survival fitness for mental illness may be measured in terms of mortality among patients. While increased mortality from both natural and unnatural causes has been reported for all mental disorders (41)(42)(43), there are considerable variations in degree of their increments. Excess mortality is strikingly high for anorexia nervosa at over six-fold higher than the mortality rate of the general population (44). Schizophrenia (33)(45), autism (46)(47), bipolar disorder (48), and unipolar depression (49)(50) also show substantial excess mortality that is around twice the normally expected mortality. Anxiety disorders have moderate excess mortality at 1.2 times higher than the expected mortality (41)(51)(52). As a large fraction of excess mortality in these disorders occurs at a young age (e.g. suicides), increased mortality also affects the reproductive period (41).

The fertility rate (i.e. the average number of children that would be born to a woman over her lifetime) may be a good metric for individual reproductive fitness. Most neuropsychiatric disorders are associated with reduced fertility. Autism patients almost never marry or have children (53). Female sufferers from anorexia nervosa have been reported to have only one third the fertility expected by the general population (54). Men and women with schizophrenia (55) and bipolar disorders (56) also show significant reductions in reproduction, having only 40% and 65% of the normally expected fertility, respectively. Unipolar depression (57) and anxiety disorders (58) have more subtle effects on reproduction, with only a 10% decrease compared to that in the general population. Importantly, these reduced fertilities do not appear to be compensated by the greater reproduction of unaffected relatives (55)(59). Another indication for fitness may be the performance of offspring from affected parents. Children of affected individuals have less children (56), higher risk of disability including mental and physical health problems (60), and decreased chances of surviving until adulthood (61).

In summary, mental disorders are associated with reduced fitness for survival and reproduction that further extends to their offspring and that are not compensated for by greater fertility among their relatives who may be a carrier of risk variants with no symptoms. Notably, there is an inverse correlation between reproductive handicap and prevalence among mental disorders. This is probably because a variable degree of

purifying selection has acted upon different traits depending on the fitness of individuals with the traits (Table 1.1).

1.2.1.3 Heritability

Twin studies have suggested that a major contributor to autism (62)(63), bipolar disorder (64)(65) and schizophrenia (66) are genetic, estimating their heritabilities among monozygotic twins at 90%, 85%, and 81%, respectively. Heritability estimates for anorexia nervosa are lower yet substantial at 56% (67), and are moderate for unipolar depression (68) and anxiety disorders (69) at 37% and 32%, respectively. Conservative estimates for heritability obtained from a large combined adoption and extended family study are available for schizophrenia (64%) and bipolar disorder (59%), in which substantial contributions of genetic component to disease aetiology were confirmed (25). Interestingly, the most deleterious disorders that have the strongest reproductive handicap appear to be those that are most heritable (Table 1.1). This observation provokes a question of how such harmful traits under strong purifying selection could be strongly inherited and yet persist in the population.

1.2.1.4 Recent origin of susceptibility variants

One explanation for an observation of very harmful traits present in the population is that the responsible mutations arose only very recently, thus these mutations have had insufficient time to be purged from the population. The mutations of most recent origin in the population at a time point would be those that arose during gametogenesis. The sex cells that harbour mutations may then be fertilised, and hence the mutations are transmitted to offspring. The occurrence of such extremely recent mutations can be easily inferred from the age of a father at the time he conceives a child (i.e. paternal age). Female germ cells undergo 24 mitotic divisions and stop dividing after puberty, while male germ cells undergo ∼23 divisions per year (70). As spontaneous mutations occur during each cell division at a relatively constant rate (see section 1.1), each division accumulates new mutations. Thus, sperm produced from an aged father that have undergone a greater number of cell divisions are more likely to harbour newly arising mutations (70). Indeed, such a paternal age effect on newly arising mutations that increased the risk of disease has been shown (71), and hence paternal age is a good indicator that reflects the burden of newly arising mutations. For recently emerged,

inherited mutations, the effect of a grandparent's age to the risk of mental disease would be a good measure, but such data are currently lacking.

A paternal age effect has been identified for neuropsychiatric disorders (Table 1.1). Autism has been shown to be subjective to very strong paternal age effects: offspring of men aged >30 are 1.4 times more likely to have autism than those of men aged ≤29 (72)(73), and the risk ratio was larger if one considers older fathers (2.2-fold risk for paternal age of >50) (74). Paternal age effect is also strong for schizophrenia: it has been shown that children of men aged >30 have a 1.4 times higher risk of having schizophrenia (75)(76). The risk of bipolar disorder has been shown to increase by 1.2-fold due to their paternal age effect (77)(75). Other mental disorders including unipolar depression have shown either insignificant changes in disease risk by paternal age effect (75), or data are not available. These risk associations with increased paternal age remain significant after controlling for maternal age and socio-economic status (72)(76). Together, the reported paternal age effects suggest a contribution of newly arising *de novo* mutations to autism, schizophrenia, and bipolar disorder.

Indeed, evidence indicating an important role for rare variants in neuropsychiatric disorders is accumulating. Especially in autism, roles for rare chromosomal abnormalities, rare *de novo* copy number variations, and rare gene mutations have been well documented. *De novo* risk variants identified from CNV or exome sequencing studies account for 10-15% of autism probands. This number is likely to rise as whole genome sequencing reveals more rare variants that could not be detected previously (78).

Nevertheless, one may argue that if a severely fitness-reducing trait was to be largely driven by *de novo* events that occur so infrequently (i.e. $\sim 2.5 \times 10^{-8}$ per nucleotide per generation) (79), the resulting trait should be extremely rare in the population. However, given that at least over a half of all coding DNA is probably involved in human brain function and development (80), the brain has an enormous mutational target size (81). Accordingly, spontaneous mutations may have a higher chance of affecting brain function and development, subsequently behaviours to some extent (82), while individual susceptibility mutations remain very rare in the population.

1.2.2 Insights from evolutionary perspectives

Several evolutionary theories have been proposed to explain the paradox of common, yet harmful and heritable mental disorders (reviewed in (81)). *(i) Balancing selection* may

occur when a variant was associated with both fitness-reducing and fitness-increasing traits, thus the variant is retained in the population. *(ii) Mismatch hypothesis* may occur when a trait was fitness-neutral in an ancestral environment but is now harmful in the modern environment. *(iii) Polygenic mutation-selection balance hypothesis* suggests a more flexible scenario, where the fate of novel mutations in the population varies depending on the mutation rate and the severity of the associated trait. Below, I explore possibilities for these hypotheses as being evolutionary mechanisms for neuropsychiatric disorders, which, in turn, may provide useful insights into the complex genetics of neuropsychiatric disorders.

1.2.2.1 Balancing selection

Several versions of balancing selection have been proposed for mental disorders. These include a susceptibility to schizophrenia balanced by a fitness-increasing trait involving language (83), depression by an ability to evoke care and sympathy (84), bipolar affective disorder by an ability to survive long winters (85), and anorexia nervosa by tolerance to starvation (86).

However, these hypotheses fail to tally with the epidemiology of mental disorders. A genetic mutation under balancing selection remains common in the population mainly via heterozygote advantage (Figure 1.3), through which unaffected carriers enjoy the balancing trait, thereby increasing their fitness. However, it has been shown that the decreased fertility of schizophrenia patients was not compensated for by increased fertilities of their relatives (55)(59). Moreover, not many examples for balancing selection have been found thus far except for a few examples such as sickle-cell disease (Figure 1.3) (81). Such infrequent observation is also consistent with the concept that multiple occasions of balancing selection would have an unfavourable net outcome for a species, thus alternative traits would eventually be selected (81).

In addition, the balancing traits proposal conflicts with the variable prevalence of some psychiatric disorders across different regions. For example, balancing traits for depression and anorexia nervosa such as evoking care and tolerance to famine are likely to be universally beneficial in most environments, whereas the two disorders have shown the greatest variability in prevalence across different regions and cultures (see section 1.2.1.1). Also, benefits from an ability to survive long winters, i.e. a proposed balancing trait for bipolar disorder, is likely to be limited to regions with high latitude,

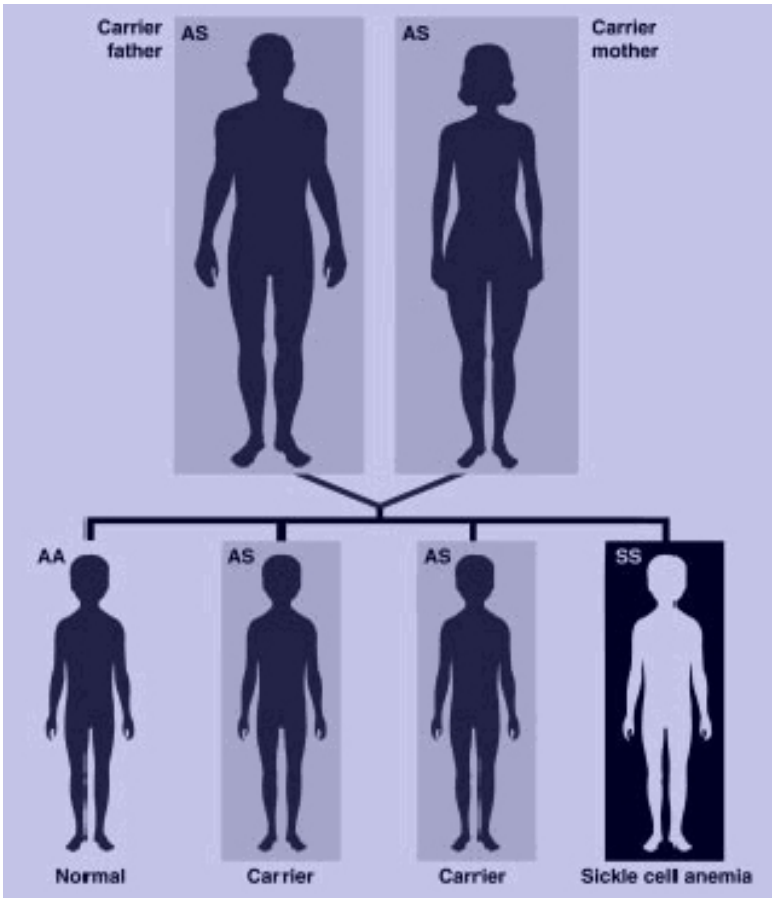


Figure 1.3: Heterozygote advantage in sickle cell trait - When each parent carries one sickle cell susceptibility allele (S) and one wildtype allele (A), each child has: a 1 in 4 chance of having two susceptibility alleles (SS) resulting in sickle cell anaemia, a 1 in 4 chance of inheriting two wildtype alleles (AA) thus unaffected, and a 1 in 2 chance of carrying both alleles (AS) thus unaffected but carrying sickle cell trait. In regions of malaria endemic, heterozygotes have a survival advantage over unaffected people who have two wildtype alleles (AA). Together, both homozygotes have reduced-fitness (AA, death by malaria; SS, sickle cell anaemia), while heterozygotes (AS) have increased fitness due to its resistance to malaria. Thus, S allele is both advantageous and disadvantageous to individual fitness, thereby balance the selective pressure. Adapted from (87).

but the prevalence of bipolar disorder has been shown to be relatively consistent across different regions (see section 1.2.1.1). Together, current evidence suggests that balancing selection may be a rather implausible scenario to explain most mental disorders (81)(88). However, reproductive disadvantage or neutrality of patients' relatives has not been confirmed for all mental disorders and other unexplored balancing traits might still remain a possibility.

1.2.2.2 Mismatch hypothesis

A mismatch between the ancestral and modern environments might lead to an unexpected prevalence of a trait that is maladaptive in the modern environment, but which used to be adaptive or neutral in the ancient environment. This hypothesis is supported by an observation that an ancestral APOE*4 allele, which may have been associated with an adaptive 'thrifty phenotype' in pre-agriculture period, has largely been replaced by alternative alleles (89). This replacement may be a consequence of APOE*4 allele's association with Alzheimer dementia and cardiovascular disease that are maladaptive in modern environment. This maladaptivity is largely because disease susceptibility is linked to a recent increase in human lifespan (see section 1.2.1.1). Although these disease traits are now maladaptive to our current environment, the reproductive disadvantage of affecteds are limited as the diseases tend to occur at a late age. Consequently, responsible mutations are purged only slowly and the diseases remain common in the population. However, such a mismatched adaptivity due to increased lifespan is unlikely to occur for neuropsychiatric disorders as the age of onset for these disorders is early, and hence affects the reproductive age (see section 1.2.1.1). Furthermore, given that affected individuals tend to have low social status and poor marriage rates in traditional rural societies (90), mental illness was probably as maladaptive as it is now in pre-modern society.

Nonetheless, if a susceptibility allele is maladaptive only in environments in which certain conditions fluctuate (e.g. parasite endemics), and hence its effect on fitness is averaged to be almost neutral over the ancestral period, the allele may remain common in the population. For example, it has been shown that heritable (avian) personality variations that are also associated with different levels of reproductive fitness tend to be preserved, when the selective pressure coincides with repeated changes in environmental condition (i.e. masting of beeches) (91). Such conditional mismatch may

provide a theoretical basis for a contribution of gene-environment interactions that involve common genetic variants to mental health (92)(93). In addition, mental disorders that have moderate heritability may be explained by gene-environment interactions to a certain extent, as estimates from twin studies include interactions between genetic factors and shared environment within family. However, this mechanism alone is unlikely to account for the prolonged existence of fitness-reducing variants directly and their consistent association with mental illness. Taken together, gene-environment interactions may provide a partial explanation for some mental disorders, especially for those showing regional and temporal variability in incidence, such as unipolar depression and eating disorders (see section 1.2.1.1). This is consistent with quantitative and molecular studies that have shown that gene-environment interactions may have a major contribution to depression (94)(93).

1.2.2.3 Polygenic mutation-selection balance hypothesis

Under polygenic mutation-selection balance, newly arising mutations with fitness-reducing effect are purged from the population at a rate proportional to their negative effect on reproductive fitness (79). While the rate of *de novo* mutations is low, the mutational target size of human mental health is relatively large (see section 1.2.1.4). It has been estimated that an average human carries 500, largely variable, mutations that may have negative effects on brain function (81). According to this mechanism, most mutations responsible for mental illness with severe reproductive disadvantage will be rare (<1%) and of evolutionarily recent origin (<100 generations). The combined burden of these mutations will then form a continuous distribution of susceptibility to mental disorder (24). Indeed, autism and schizophrenia, whose patients have shown the most severe reproductive disadvantage, have also been associated with strong paternal age effect (72)(76), and rare chromosomal aberrations (95)(96). Also, rare variants of recent origin, particularly many *de novo* CNVs, have been associated with these disorders (97)(29), inferring a major contribution of newly arising mutations to disease susceptibility (Table 1.1). This hypothesis is also compatible with the high heritability estimates of these disorders, as even *de novo* mutations are normally shared between monozygotic twins. However, other mental disorders with mild reproductive disadvantage have not been associated with paternal age effect (Table 1.1), and hence are less

likely to be affected by *de novo* mutations. Also, the regional and temporal variations in their prevalence cannot be explained by newly arising mutations, which is a consequence of stochastic events.

1.2.3 Implications for genetic architecture of mental disorders

Considering the epidemiological evidence for the six representative mental disorders, mental disorders appear to have heterogeneous genetic aetiologies contributed by different evolutionary mechanisms, mainly by gene-environment interactions and mutations of recent origin.

The degree of contribution from each mechanism to the genetic aetiology of different disorders can be evaluated based on multiple lines of epidemiological evidence (Figure 1.4). Mental disorders such as autism, which has a strong reproductive disadvantage (shown as a product of mortality ratio and the reciprocal of fertility ratio in Figure 1.4) and paternal age effect, but has little regional variation in incidence, are likely to be largely contributed by rare variants of recent origin. On the other hand, disorders such as unipolar depression, which has a mild reproductive disadvantage and little paternal age effect, but has large regional variation in incidence, are likely to be largely influenced by common genetic variants that are involved in gene-environment interactions. The genetic aetiology of disorders that show strong paternal age effect but mild regional variation in incidence, such as schizophrenia, is likely to be affected greatly by both mechanisms, to a greater extent by rare variants of recent origin and to a lesser extent by common variants with gene-environment interactions. Disorders with moderate effect on both paternal age effect and regional variation, such as bipolar affective disorder, are likely to have contribution from both mechanisms equally.

As paternal age effects of anorexia nervosa and anxiety are currently not available, it is difficult to directly infer the degree of contribution of rare variants of recent origin to their genetic aetiology. However, given the essentially normal reproductive fitness in anxiety, the contribution of newly arising mutations is probably less than their contribution to anorexia nervosa which has a severe reproductive disadvantage. In order to assess the contribution of common variants related to gene-environment interactions to anxiety, evidence for regional variations in incidence may be required. Ideally, systematic molecular and/or genetic data for gene-environment interactions for all mental disorders would confirm the level of contribution by such mechanisms.

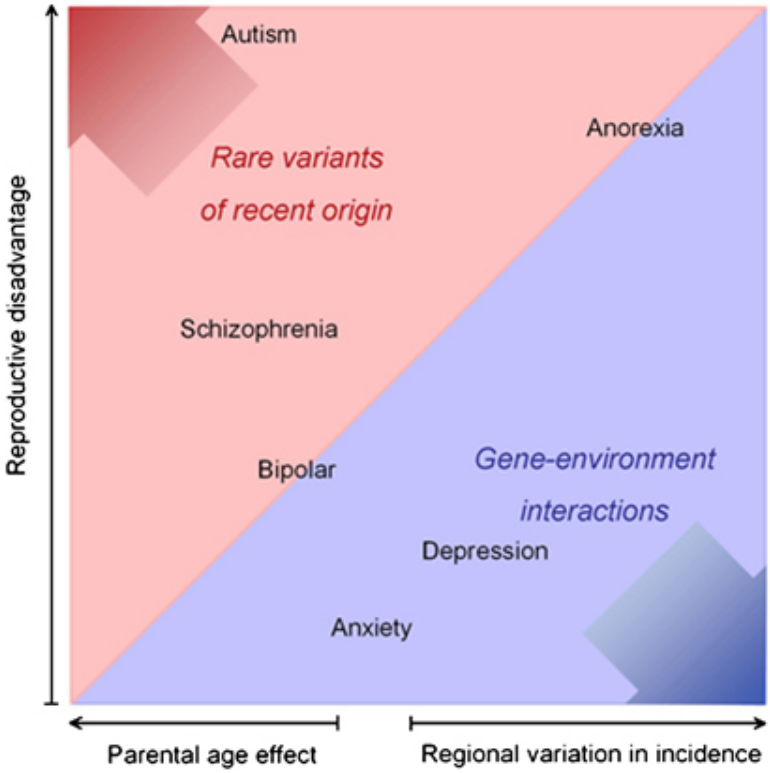


Figure 1.4: Architecture of variable genetic aetiology of mental disorders inferred from epidemiological data and evolutionary theories - The y-axis represents reproductive disadvantage (i.e. log[mortality ratio×(1-fertility ratio)]). The x-axis reflects two types of epidemiological data as markers for corresponding plausible evolutionary theories and types of contributing variants: paternal age effects (to the left) as a marker for rare variants of recent origin and regional variation in incidence (to the right) as a marker for common variants retained in the population via gene-environment interaction. The location of each disorder in the framework is represented by a spectrum rather than being discrete, which implies that there are overlaps between mental disorders and heterogeneous contribution of genetic aetiology. Adapted from (24).

1.2.4 Implications for genetic studies of neuropsychiatric disorders

The framework implies that unreplicated, heterogeneous associations from studies on mental disorders might be due to gene-environment interactions. Once an objective measurement for environmental conditions relevant to mental disorders is developed (98), and hence a genome-wide study for gene-environment interactions becomes more feasible (99)(100), common variants for mental disorders may be effectively detected. While identification of common susceptibility variants remains an ongoing challenge, increasing numbers of rare variants have been associated with neuropsychiatric disorders, taking advantage of a recent advance in genome-wide arrays and whole genome sequencing technologies. Nevertheless, these identified variants are individually rare, thus cannot serve as accurate predictors for disease causation. Rather, individual rare susceptibility variants would affect a set of relevant biological processes that have converging functional effects that lead to medical symptoms and phenotypes. Indeed, single gene mutations and rare structural variants identified among autism patients have consistently been shown to be involved with functions related to the assembly and functioning of synapses (reviewed in (23)). Therefore, given the proposed genetic architecture and increasing numbers of rare variants being associated with mental disorders, a current key research task is to systematically identify disease-relevant biological processes, and to interpret them in a pathological context. These are the key aims of Chapter 3 and Chapter 4 in this thesis (see section 1.3).

Most importantly, when investigating the genetics of neuropsychiatric disorders, it should be noted that most mental disorders lie on a continuum. Comorbidity between mental disorders is high and this is a likely consequence of overlaps in their genetic aetiologies (101). Genetic overlaps have already been reported for pairs of mental disorders: schizophrenia-bipolar disorder (25)(102), bipolar disorder-unipolar depression (65), and unipolar depression-anxiety (103). These overlaps between mental disorders are also consistent with their proximity within the framework (Figure 1.4), supporting the concept of their shared genetic aetiology driven by common evolutionary mechanisms.

1.3 Project rationales and thesis structure

In this thesis, I investigated the genetic aetiologies of two neuropsychiatric disorders: autism spectrum disorders (ASD) and obsessive-compulsive disorder (OCD).

1.3.1 ASD project

Given the expected and observed strong contribution of rare variants to the genetic aetiology of autism (see section 1.2.1.4), I employed rare *de novo* CNVs identified from simplex ASD families (i.e. families with one affected offspring and his/her unaffected parents), and analysed them using several large-scale publicly available databases including the Mouse Genome Informatics (MGI) phenotypes, Gene Ontology (GO) terms, and protein-protein interactions. To this end, I sought to identify and interpret autism-relevant biological processes and/or pathways, in which disruptions of any points may lead to autistic symptoms.

- In **Chapter 3**, I analysed 4 sets of *de novo* CNVs identified from individuals with ASD (ASD *dn* CNVs), using simple 1:1 mouse:human orthologue relationships (see section 1.1) and extensive gene-phenotype information available in mice. As these individual CNVs are rare in both cases and controls, a conventional case-control comparison was not possible. Instead, I tested whether ASD *dn* CNV genes are significantly over-represented with specific mouse model phenotypes compared to that expected from the whole genome. This approach allowed an identification of objective mouse model phenotypes, as the methodology does not use any *priori* information about human ASD presentation. In addition, ASD *dn* CNV genes that contributed to these mouse phenotypic associations became candidate genes for ASD, of which functional investigation may shed further light on the pathology of ASD.

- In **Chapter 4**, I analysed these ASD candidate genes that were determined by a statistically-robust methodology (from Chapter 3) in order to identify a biological network related to ASD. To do this, first the connectivity among these ASD candidate genes was tested using publicly available protein-protein interaction information. Then, I performed a systematic functional analysis on the identified ASD network, not only in entirety, but also for duplicated and deleted ASD

network genes separately. This allowed me to examine whether there is any functional difference between duplicated and deleted genes found in ASD patients.

1.3.2 OCD project

While OCD is classified as a type of anxiety disorder, its symptoms are distinctive and epidemiological data suggest a mixed genetic aetiology from both rare and common variants. The prevalence of OCD has been shown to be 1-3% (104)(105), while its socio-cultural variance is controversial (106)(107). The median age of onset is 18, affecting the reproductive age (108)(109). Current evidence suggests that OCD has a low effect on individual mortality (110). The fertility of OCD patients appears to be lower than that of individuals with bipolar disorder and similar to that of schizophrenia (111). Estimated heritability of OCD is as high as 80-87% (112), and an advanced paternal age effect has been shown among the fathers of OCD patients (113). The overall reproductive disadvantage of OCD, based on its effect on mortality and fertility, is probably similar to that for bipolar affective disorder. Other epidemiological data such as prevalence, heritability and moderate paternal age effect are also similar to the epidemiology of bipolar disorders, hence placing OCD near bipolar disorder in the genetic-evolutionary framework discussed before (Figure 1.4). In Chapters 5 and 6, high-throughput sequencing was employed to study OCD, which allows an identification of a large number of both rare and common variants in cases and controls. Considering the important role of non-coding regulatory regions in disease (see section 1.1), targeted regions included both coding and non-coding sequences.

- In **Chapter 5**, I took advantage of a brain circuitry model for OCD that has been implicated based on functional evidence. I performed resequencing of a human OCD case-control cohort, targeting genes including those expressed in the cortico-striatal-thalamo-cortical (CSTC) brain circuit and functionally related genes, candidate genes from linkage and association studies, and animal studies. By comparing the allele frequencies of detected variants between cases and controls, I sought to identify genetic variants that are associated with OCD. Considering the limited cohort size (200 cases and 200 controls) and the mixed genetic aetiology of OCD, I also examined an aggregated effect of multiple variants per gene in order to identify candidate genes that are significantly associated

with OCD. Finally, I performed a functional analysis of these candidate genes to identify OCD relevant biological processes.

- In **Chapter 6**, I used dogs as the best natural animal model for human OCD. Not only does canine compulsive disorder (CCD) best resemble human OCD in its disease progression, drug response, and age of onset among other model organisms, but limited genetic heterogeneity within modern breeds provides a unique advantage in the identification of susceptibility variants and genes for complex disorders. Target regions were determined by strong associations with CCD identified from a previous GWAS and haplotype tests between dog breeds that are enriched for CCD, based on the assumption that the regions responsible for CCD behaviour may be commonly shared among these predisposed dog breeds. By resequencing 12 case dogs and 12 control dogs, I sought to identify genetic variants for CCD given their phenotypic concordance and potential functional consequence inferred from evolutionary conservation (see section 1.1.3). Finally, I compared the top candidate genes from the human (Chapter 5) and the dog studies (Chapter 6) to explore common mechanisms for compulsive disorder across both species.

In **Chapter 7**, I discuss the significance of these research projects in the context of the current field of neuropsychiatric disease genetics. Also, I will discuss potential future research investigation that may be suggested as a consequence of my analyses described in Chapter 3, 4, 5, and 6.

2

Materials and methods

As the two projects undertaken in this thesis employed different types of large-scale datasets that were obtained from various platforms and technologies, I needed to apply diverse analytical methods. Thus, in this chapter, I will describe the general approaches to each project and the differences between the techiques and/or the study designs that were used to generate the raw data. Also, I will summarise functional genomics resources and statistical methods that were frequently used across these projects. Materials and Methods that are specific to one project are described at the beginning of the relevant chapter.

2.1 Datasets

2.1.1 Rare *de novo* copy number variants (CNVs) found in ASD patients

In Chapters 3 and 4, I employed 4 sets of *de novo* CNVs identified from individuals with ASD (termed ASD *dn* CNVs), namely, the Autism Genome Project (AGP) (28), Marshall *et al.* (27), Levy *et al.* (114), and Sanders *et al.* (115). *De novo* CNVs were detected by investigating ASD families, each comprised of at least one affected offspring and both unaffected parents. Such an approach using trios permits the distinction of whether the CNVs were inherited or have newly arisen. All of the ASD patients considered in these ASD *dn* CNV sets met the standard diagnostic criteria described in the Autism Diagnostic Interview - Revised (ADI-R) (116) and the Autism Diagnostic Observation Schedule (ADOS) (117).

Marshall *et al.* ASD *dn* CNVs are contained in the oldest set that was reported in 2008. They used an Affymetrix GeneChip Human Mapping 500K Array that genotypes ∼500,000 single nucleotide polymorphisms (SNPs). The identified genome-wide SNP pattern from this array was then used to infer CNVs. CNVs were called based on allele-specific signal intensities that were integrated on SNP spacing and SNP allele frequencies information using a hidden Markov model (HMM) algorithm (118). The AGP ASD *dn* CNV set (reported in 2010) and the Sanders *et al.* set (reported in 2011) were both detected with a higher resolution SNP array, the Illumina 1M microarray. A single Illumina 1M array can genotype ∼1 million SNPs, providing a twice-fold higher resolution compared to the microarray used in Marshall *et al.* study. For some of the Sanders *et al.* ASD *dn* CNVs, the Illumina 1M Duo microarray was used, which

provides an even higher resolution. CNVs detected from the AGP control cohort were also employed in my study as a comparison to ASD *dn* CNVs. Levy *et al.* ASD *dn* CNVs (reported in 2011) were identified using comparative genomic hybridisation (CGH) (119)(120), with the NimbleGen HD2 microarray platform which contains 2.1 million probes.

Calling CNVs using array CGH is different from that using SNP arrays. Array CGH permits the direct measurement of gain or loss of DNA, while SNP arrays allow the calling of copy number changes based on SNP patterns. In array CGH, patient and control DNA that are each labelled with different fluorescence compete to hybridise with oligonucleotide probes on a microarray, and the intensities of fluorescent signals from control and patient DNA are compared to identify local DNA loss or gain. Balanced chromosomal abnormalities such as inversion and translocations thus cannot be detected using array CGH.

Finally, it should be noted that the same ASD families were analysed in both Levy *et al.* and Sanders *et al.* studies. Nevertheless, the CNVs detected from each study may differ as the studies employed different platforms and analytical methods. Genomic overlaps between these 4 CNV sets are described in Chapter 3.

2.1.2 Targeted sequence data sets from case-control cohorts of OCD and CCD

In Chapters 5 and 6, ∼13Mb human and ∼5Mb canine genomic regions were selectively sequenced among affected individuals and matched controls. Genomic regions were targeted using the NimbleGen hybrid capture technique, followed by Illumina's high-throughput sequencing. Specific hybrid capture arrays and sequencing platforms slightly differ for each species and these details are provided in Chapters 5 and 6. The major difference in the sequence datasets between human and dog studies is whether individual samples are identifiable or not. In the OCD study, human samples were grouped into pools of multiple individuals with the same phenotypes (i.e. either case or control). As a result, each pool, rather than an individual, was identifiable in the human sequence dataset. On the other hand, dog sequences were labelled individually, and thus each dog was identifiable. Due to this difference, I needed to apply different analytical approaches to each data set. Briefly, in humans, my analysis was based on

the comparison of estimated allele frequencies between cases and controls. In dogs, I accounted for individual dogs' genotypes and breeds when comparing cases and controls. Details of these approaches are described in Chapters 5 and 6.

2.2 Functional genomics resources

2.2.1 Mouse Genome Informatics (MGI)

MGI is an international database resource for laboratory mice. Projects contributing to this database include: the Mouse Genome Database (MGD) project (17) relating to gene characterisation, nomenclature, mapping, gene homologies, sequences, phenotypes, allelic variants and mutants; Gene Expression Database (GXD) project (121) integrating gene expression information such as endogenous gene expression during development; Mouse Tumor Biology (MTB) database project (122) concerning frequency, incidence, genetics, and pathology of neoplastic disorders in different mouse strains; and MouseCyc project (123) regarding biological pathways focussing on *Mus musculus* metabolism such as biosynthesis, degradation, energy production, and detoxification. In particular, gene - phenotype information for ∼6,000 mouse orthologues was extensively used to identify statistically robust autism-relevant mouse model phenotypes in Chapter 3.

2.2.2 Functional attributes and protein interaction

The Gene Ontology (GO) (124) provides a description of gene and gene product attributes across species and multiple databases using standardised GO terms. GO annotations cover three domains: *(i) cellular component* that indicates the parts of a cell or its extracellular environment, *(ii) molecular function* that describes the primary activities of a gene product at the molecular level (e.g. binding or catalysis), and *(iii) biological process* that relates to operations or sets of molecular events that is relevant to the functioning of integrated living units such as cells, tissues, organs, and organisms (124). GO terms were referred throughout Chapters 4, 5, and 6 for functional investigation of genes of interest. InWeb database (125)(126) provides 169,810 high-confidence pair-wise interactions from 12,793 proteins including human proteins and their orthologues. These known protein-protein interactions were used to evaluate interconnectivity of candidate genes for ASD in Chapter 4.

2.2.3 Evolutionary conservation status

Constrained elements from the genomes of 29 mammals and the elements' associated conservation scores (3) were used in Chapters 5 and 6 in order to prioritise detected variants (see section 1.1.3). Sequence alignments of 46 vertebrates (127) were also used to investigate the conservation of specific bases across different species. Details of the conservation and alignment tracks are described in section 5.6.15.

2.2.4 Regulatory marks

The ENCODE (128) is an international collaborative consortium that aims to build a comprehensive annotation of functional elements in the human genome. Functional elements may act at the protein, RNA, and/or DNA levels and may regulate gene activities in different cells and circumstances. While an increasing number of new assays for specific cell types are being integrated into the ENCODE database (6), the data most frequently used herein include: RNA profiling data (showing transcription level), chromatin immunoprecipitation sequencing (ChIP-seq) data (showing regional histone modification or transcription factor binding), and DNase1 hypersensitivity assay (for regions of open chromatin status). For a more detailed description of these data, see section 5.6.13). The original publication and cell types are referenced individually.

The TRANScription FACtor database (TRANSFAC®) (129) contains published data on eukaryotic transcription factors, their experimentally-proven binding sites, and regulated genes. The regulatory motif instances are also assigned with positional weight metrics that were calculated based on the compilation of binding sites integrated in their knowledge-base. In Chapters 5 and 6, I used these regulatory motif sequences and associated positional metrics in order to predict functions of candidate non-coding variants. I searched TRANSFAC® against sequences of interest that were determined by an association with disease phenotypes and regulatory evidence. The detailed search algorithm is described in section 6.6.7.

2.3 Statistical analyses

Various statistical tests were employed in this thesis. While details of the statistical methods are described in each relevant chapter, here I will describe statistical tests that were frequently used throughout the thesis.

2.3.1 Hypergeometric test

A test for an over-representation of a term (e.g. MGI phenotype or GO term) in a set of genes was achieved using the hypergeometric distribution. This distribution is used for sampling without replacement. Let us assume that among n genes in my set, k genes are annotated with a specific term. I want to know how likely this outcome could occur by random chance, given that there are a total number of N genes and of these M genes are annotated with the specific term. This situation can be interpreted as an urn model with black and white stones, each stone as a gene and black being annotated with a tested term. A theoretical experiment for this test would be to draw n stones (without replacement) from an urn, in which there are N number of stones and of these M genes are black. The probability of having at least k black stones drawn from the urn can then be calculated using the cumulative distribution function of the hypergeometric distribution. Intuitively, a probability for having n black stones drawn, when sampled n times, would be very small. The hypergeometric test was a major analytical tool for the work described in Chapters 3, 4, and 5, to investigate phenotypical or functional enrichments among genes of interest.

2.3.2 Permutation test

The permutation test (also known as a randomisation test) is a type of statistical significance test. In this test, the distribution of the test statistic under the null is obtained by calculating all possible values of the test statistic resulting from permutation of labels on the observed data points.

The basic steps are as follows. First, decide on a metric to measure the effect in question. For example, in Chapter 4, I used the number of direct interactions identified among the protein products of a given gene set as a metric. In Chapter 5, I used the sum of differences in allele frequencies between the case and control population as a metric.

Next, calculate the test statistic from the data (i.e. observed test statistics), and repeat the same calculation on data that have been randomly shuffled N times, where N is sufficiently large. For instance, in Chapter 5, the allele frequency metric was calculated on case-control permuted data and this procedure was repeated 1000 times. In Chapter 4, given the difficulty of obtaining control *de novo* CNVs, data were permuted

against an appropriate genomic background rather than using a control population. For example, 500 randomised gene sets, matched by the number of genes to a test gene set, were sampled from gene orthologues that are annotated with a *Behavior/Neurological* category. In this way, one can calculate the test statistics expected from genes annotated with this category, from which the candidate gene set was selected.

Finally, count the occasions when the test statistics from randomisation performed better than the observed test statistics and calculate the proportion of such cases given N times randomisation. This then provides the probability observing the result (or better) by chance. When calculating the probability, 1 is added to the numerator in order to avoid zero probabilities.

2.3.3 False discovery rate (FDR)

Controlling the false discovery rate (FDR) is a powerful statistical method to correct for issues arising from multiple comparisons. As more hypotheses are tested, a small number of type I errors (i.e. incorrectly rejecting the null hypothesis when it should be accepted) (130) that occur in each test may accumulate and hence may result in a greatly increased rate of false positive observations (131). Applying a FDR threshold addresses this multiple testing problem by allowing only an expected proportion of "false discoveries" among statistically significant findings (131).

A simple form of FDR estimation can be achieved as follows. For a specified score threshold, count the number s_b of observed scores $\geq$t. The number s_n of scores $\geq$t resulting from the null score distribution is also counted. Assuming that the total number of observed scores and null scores are equal, the estimated FDR is s_n/s_b (132).

In this thesis, the number of false discoveries was controlled using the Benjamini-Hochberg procedure, which relies on p-values being uniformly distributed under the null hypothesis (131). The procedure is as follows (131)(133). First, the p-values are ordered such that $p_{(1)} \leq ... \leq p_{(m)}$. Second, each $p_{(i)}$ value is compared with $q\frac{i}{m}$, where q is the desired FDR level (e.g. 5%). Lastly, with $k = max(i : p_{(i)} \leq q\frac{i}{m})$, all hypotheses belonging to $p_{(1)}$,...,$p_{(k)}$ are rejected, thus considered as being statistically significant. In the reference (131), it was shown that when test statistics are independent this procedure controls for the expected ratio E(V/R) of the number of false positives (V) among all significant tests (R) at level $\leq q$.

2.4 Frequently used software packages

Software packages were frequently used throughout these projects as tools for handling large-scale datasets. A substantial part of the code for data processing and association tests was written with R packages. Other code was written with programming languages including Perl, Python, and UNIX shell. BEDTools and PLINK were also used. BEDTools (134) is a flexible suite of codes that is useful for comparing, manipulating, and annotating genomic features. PLINK (135) is a whole genome association analysis toolset that assists in basic data management and applying various statistical tests. PLINK was used for handling GWAS data (for comparison with the sequence data set), population stratification detection (see section 5.6.2), and multimarker haplotypic tests (see section 6.3.5.2).

3

Autism-relevant mouse models and phenotypes objectively identified from rare copy number variants

3.1 Abstract

Autism is a highly heritable neuropsychiatric disorder that is characterised by functional deficits in social interaction and communication, and restricted interests and repetitive behaviours. The genetics of autism is largely unknown, but the low rate of reproduction among affected individuals suggests that autism-predisposing genetic variants are less likely to be common. Focussing on rare *de novo* copy number variants (CNVs), I tested whether the genes affected by *de novo* CNVs identified among autism patients (ASD *dn* CNVs) are significantly enriched among model phenotypes resulting from disruption of mouse orthologues, using 1:1 human:mouse orthologue relationships. Employing 4 sets of ASD *dn* CNVs, I objectively identified 76 different model phenotypic associations, including an *abnormal CNS synaptic transmission* phenotype that was triplicated (False Discovery Rate $< 5\%$). Other phenotypic associations were found among many behavioural phenotypes that are related to abnormal social interaction, abnormal learning/conditioning, and stereotypic behaviour. These significantly associated behavioural phenotypes were readily comparable to the presentations of human autism and its frequently comorbid symptoms. Corroboratively, an independent set of genes that were previously implicated in autism also yielded significant associations with similar mouse model phenotypes. Finally, these results offer a causal hypothesis for $\sim$30% of the patients in the cohorts and identify 59 largely novel candidate genes for autism.

3.2 Introduction

3.2.1 Autism Spectrum Disorders (ASD)

3.2.1.1 Classification and diagnosis

Autism is a severe neurodevelopmental disorder which affects approximately 1 in 150-200 of the human population (136)(137)(138). It is characterised by functional impairments in three behavioural/cognitive domains: reciprocal social interaction, language and communication, and repetitive behaviours and restricted interests (Figure 3.1) (139). According to the current version (IV) of the Diagnostic and Statistical Manual of Mental Disorders, autism falls into the category of "pervasive developmental disorder (PDD)", which also includes Asperger syndrome, childhood disintegrative disorder, Rett syndrome, and pervasive developmental disorder not otherwise specified (PDD-NOS) (140). The term "autism spectrum disorders" (ASD) includes Asperger syndrome, in which affected individuals have normal language ability, and PDD-NOS, which is diagnosed when the full requirement for autistic disorder or Asperger syndrome is not satisfied.

Like many other psychiatric disorders, the clinical definition of ASD relies on observation of behavioural phenotypes, without objective diagnostic tools such as laboratory tests, neuroimaging, or biomarkers. Furthermore, ASD diagnostic criteria (i.e. functional impairments in the three behavioural/cognitive domains) are not categorical, but quantitative and continuous. In addition, as ASD diagnosis is not currently based on aetiology or pathophysiology, ASD is not unitary conditions, but rather, clinical syndromes (141). For instance, among severe autistic individuals who are affected in all three core domains, IQ varies widely: some individuals may have normal or even above average IQ, although 30-50% of autistic individuals also possess severe intellectual disability (142). These limitations in clinical classification make diagnosis of ASD somewhat arbitrary, hence challenging to study.

It has been suggested that autism should be viewed as a continuum of normal population variation in behaviour and cognition rather than a set of discrete features (141). This notion that ASD consists of several components of normal variation may be evaluated by observing "endophenotypes" within families of affected individuals. An endophenotype is a heritable feature associated with a disorder that is observed in non-affected family members more often than the general population. Familial studies have

shown that distinct ASD-associated endophenotypes (i.e. features involving language (143)(144), social responsiveness (145)(146), and restrictive behaviours (147)(148)) are more frequently observed among parents and siblings of ASD probands than the general population. From this perspective, there would be few genes causing ASD per se. Rather, a combination of genetic variation in behavioural and cognitive functions may predispose to ASD (Figure 3.1).

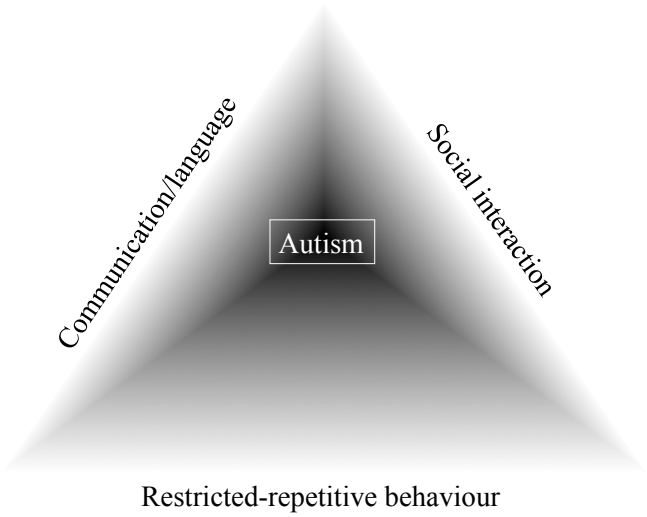


Figure 3.1: Three core domains of severe dysfunction in ASD in the context of normal population variation - The current diagnosis of autism requires deficits in social interaction, communication, and restricted interests and repetitive behaviours. Normal variation (area shown in white / light grey) of heritable features in the three core components may contribute to ASD susceptibility. High levels of dysfunctions in these three components (area shown in black) lead to ASD.

3.2.1.2 Comorbidity

Besides the three core behavioural/cognitive domains required for diagnosis (Figure 3.1), several other medical features are frequently observed among individuals diagnosed with autism (Table 3.1). For example, sensory abnormalities are observed in more than 90% of individuals with autism. Motor signs including hypotonia and apraxia are also common, affecting 60-80% of individuals with ASD. Other symptoms such as sleep disturbance, gastrointestinal disturbance, and epilepsy are also frequently observed among affected individuals.

Domain	**Autism** (149)(150)	**Asperger**	**PDD-NOS**	**ASD** (151)(152)
Social interaction	necessary	necessary	necessary	
Language	necessary	-	variable	
Repetitive/restrictive behaviours	necessary	necessary	variable	
Sensory abnormalities (153)	>90%	80%	variable	94%
Developmental regression	15-40%	?	?	15-40%
Motor signs (154)(155)	60-80%	60%	60%	60-80%
Gross motor delay	10%	?	?	5-10%
Sleep disturbance (156)(157)	55%	5-10%	40%	50%
Gastrointestinal disturbance	45%	4%	50%	4-50%
Epilepsy (158)(159)	10-60%	0-5%	5-40%	6-60%
Comorbid psychiatric diagnosis	70%	60%	>25%	25-70%

Table 3.1: Domains of dysfunctions in autism spectrum disorders (ASD) - Domains indicated as "necessary" are required for the diagnosis of a disorder shown in the column heading. A dashed line indicates domains that are not observed among affected individuals. The frequencies in the table are conservative estimates based on several references shown in the table. The ASD column shows an estimate for the broad category of ASD, including autism, Asperger syndrome, and PDD-NOS. Developmental regression refers to loss of function in either the language or social interaction domain, or in both. Motor signs include hypotonia, gait problems, tow walking (i.e. walking on one's toes without putting much weight on the heel or any other part of the foot), and apraxia. Gastrointestinal disturbance denotes diarrhoea, constipation, reflux, or bloating for at least six months. The broad range of Epilepsy estimates reflects the presence of other comorbid symptoms, such as intellectual disability or cerebral palsy, which significantly increase epilepsy risk by 30% and 60% respectively. Comorbid psychiatric diagnosis includes mood disorders, conduct disorders, aggression, and attention deficit/hyperactivity disorder (ADHD). Adapted from (141).

3.2.1.3 Genetics

ASD has a strong genetic component. Twin studies have suggested an estimated heritability of ∼90% for the broad ASD phenotype (160), and first-degree relatives are at 25 times higher risk than the general population (161). Despite the importance of genetic analysis shown by this high heritability, there have been no genetic linkage or association studies that confidently and consistently identified genetic loci for ASD. However, recent studies have shown an important role for structural variations in ASD. A pioneering experiment by Sebat *et al.* has shown a significant burden of *de novo* CNVs among autism patients. In their study, identified CNVs explain ∼10% of sporadic autism and ∼3% of familial cases (97). Similar results have been observed by Marshall *et al.*, where *de novo* CNVs may account for ∼7% of sporadic ASD and for ∼2% of familial cases (27). In addition to rare *de novo* CNVs, recurrent CNVs have also been identified in individuals with autism. Duplication of 15q11-13 (162)(27), duplication/deletion of 16p11 (163)(27), or deletion of 22q11-13 have been shown to account for ∼0.5-1% of autism cases (97)(164). Taken together, known structural variants may explain 10-20% of ASD causes (161).

3.2.2 Mouse as a model for human autism

Recently, behavioural neuroscientists have generated new phenotyping assays for laboratory mice by a collaboration with specialists for human neuropsychiatric disorders (165). Consequently, an extensive set of behavioural assays has been developed to test face validity (i.e. analogies to human disease phenotypes; see 6.2.4) of mouse models for human behavioural disorders (165). Those assays that interrogate social interaction, communication and repetitive/stereotyped behaviours are of particular use for studying autism as deficits in these three behavioural domains are required for autism diagnosis. To facilitate the understanding of mouse model phenotypes that are identified and discussed later in this chapter, I will describe here examples of behavioural assays that are highly relevant to types of behaviours in autism. A comprehensive list of mouse phenotyping assays for autism-relevant behaviours and detailed descriptions are available in (165).

3.2.2.1 Social interaction deficits in mice

Mus musculus is a highly social species. They engage in social reciprocal interactions (166), shared nesting (i.e. building nests and sleeping together in a group), sexual and nurturing activities (167), territorial scent marking (168)(169), and aggressive behaviours (170). Social deficits in mouse can be identified by observing mouse-specific social behaviours that can be parameterised as described below.

Reciprocal social interaction A sophisticated measure of social interactions in mice is possible by observing groups of juvenile or adult mice that are put together in a standard cage or arenas. Normal mouse behaviours including nose-to-nose sniffing, nose-to-anogenital sniffing, following, pushing each other, crawling over/under each other, chasing, climbing, and grappling are the standard parameters to assess reciprocal social interactions in mice (171)(172)(173)(166)(174). The experiments to observe those behavioural parameters need to be carefully designed, taking into account various factors such as session duration, time of day, prior social isolation, and group composition by age, sex and strain. The same mice can be repeatedly tested for the same assays, allowing the measurement of phenotypes in different developmental stages.

Social approach In assaying social approach, sociability is scored as the time a mouse spent with another new mouse compared to the time spent with a non-living object (175)(172)(173)(176)(177). This social approach task is typically conducted in a special chamber with widely spaced wire bars that permit olfactory, visual, auditory, and some physical contact while preventing aggressive and sexual interactions, to ensure the measurement of pure interest in approaching rather than other non-purely social activities. There are photocell-equipped devices that enable automatic tracking of social approach by detecting beam-breaks caused by the movement of a subject mouse that are converted to time duration spent with a novel mouse or object (177)(178). Such instruments can offer standardised higher-throughput assays, although detailed social behaviours cannot be evaluated.

Social partner preference Placing different social partners and measuring the time spent by a subject mouse with each target mouse is also used to score a mouse's

sociability. For example, if a subject mouse spends more time with a new mouse than with a familiar mouse, this provides a measure of social recognition (i.e. distinguishing familiar from novel mice) (179)(180)(181). A particular combination of different social partners such as pair-bonded mates or a kin offers metrics for social affiliation (e.g. parent-pup interaction, pair-bonding, mating or aggression) (182)(183). Also, partners may be introduced sequentially with time delays between presentations in order to evaluate social memory (i.e. time delay in recognising a familiar mouse) (184)(185).

3.2.2.2 Communication deficits in mice

How mice communicate with one another is not yet well understood, but olfactory signals appear to be a primary mechanism (167). Literature has shown that ultrasonic vocalisation and visual cues may also contribute to the communication between mice (186)(187)(188)(189)(190)(191). There are several behavioural tasks that are now routinely used to evaluate the olfactory and auditory cues generated by mice and their responses to such signals.

Urinary pheromone and social odours As in many other species, mice leave urinary pheromones as territorial scent marks. They also have strong interest in others' urinary scent, which is frequently exhibited as sniffing the anogenital region of a novel mouse and inspecting others' urinary scent marks in a cage (192)(193)(194)(195). The number of scent marks and counter-markings near the initial urinary scent marks may be an indicator of olfactory communication (196). Furthermore, normal mice spend significantly more time on sniffing social odours (e.g. urine of other mice) than non-social odours (e.g. almond extract or banana flavouring) (197)(198).

Ultrasonic vocalisations Mice tend to utter elaborate ultrasonic vocalisation in social situations. Sensitive ultrasonic microphones and headphones have allowed recording of detailed sonograms of mice's vocalisation and revealed that mice in various social situations produce distinctive ultrasonic vocalisations. Such cases include pups that are isolated from the nest (180)(179), resident females that are faced with an intruder (181)(187), and males detecting female urinary pheromones (189)(199).

3.2.2.3 Repetitive/restricted behaviours in mice

Repetitive/stereotyped behaviours Normal mouse behaviours include spontaneous motor stereotypies such as circling, jumping, backflips and self-grooming. These normal rodent behaviours may persist for unusually long periods of time, thus considered as repetitive behaviours (200)(201)(202)(203). For example, BTBR T+Tf/J mice may self-groom for up to 2 minutes while the control strains would self-groom only for 5-10 seconds (172)

Perseverative behaviours and restricted interests As perseverative behaviours are fairly common in mice, reversal learning tasks may be used to measure a mouse's flexibility to switch an established habit to a new habit. For instance, once a spatial habit of using certain paths in a maze is established by using a reinforcer (e.g. food reward), the reinforcer can be moved to a new location to test whether the mouse could alter its way. In this type of test, a mouse model for autism may well be able to acquire the first spatial habit, but unable to change the acquired habit to a new one. Another normal mouse behaviour that can be used to test restricted interests is their tendency to explore every feature of a new environment. These characteristics are displayed as exploratory locomotion of an open field, sniffing of novel objects, or nose poking into holes in the wall (204). If a mouse investigates newly introduced objects or areas only to a limited extent rather than all aspects, this may be comparable to restricted interests observed in individuals with autism.

3.3 Materials and Methods

3.3.1 *De novo* CNVs identified in individuals with ASD

The ASD patients considered in this study had been diagnosed under the standard methods, i.e. the Autism Diagnostic Interview - Revised (ADI-R) (116) and the Autism Diagnostic Observation Schedule (ADOS) (117) criteria. *De novo* variants were detected from ASD families, when a variant was found in a proband, but not among the unaffected parents.

	#Subjects	#CNVs	CNV median size	#Genes	Genome covered
AGP Control	1,146	3,096	91 Kb	2,789	326 Mb
AGP All	54	73	118 Kb	280	23 Mb
AGP Gain	26	34	134 Kb	156	14 Mb
AGP Loss	31	39	92 Kb	154	11 Mb
Marshall et al. All	24	28	3.3 Mb	758	119 Mb
Marshall et al. Gain	11	11	4.3 Mb	408	42 Mb
Marshall et al. Loss	15	17	3.2 Mb	402	80 Mb
Levy et al. All	82	94	318 Kb	644	69 Mb
Levy et al. Gain	35	39	527 Kb	295	30 Mb
Levy et al. Loss	52	55	90 Kb	350	40 Mb
Sanders et al. All	63	67	719 Kb	839	90 Mb
Sanders et al. Gain	29	32	850 Kb	417	40 Mb
Sanders et al. Loss	35	35	607 Kb	428	51 Mb

Table 3.2: Genomic extent and ENSEMBL gene content for ASD-associated *de novo* CNVs and AGP control CNVs. - Sets termed *All* are formed by combining both *Gain* (duplication) and *Loss* (deletion) CNVs. Note that all the individuals in Sanders *et al.* study were also analysed in Levy *et al.*

In this study, I employed four sets of *de novo* CNVs arisen from individuals with ASD (Table 3.2): The Autism Genome Project (AGP) set consists of 73 *de novo* CNVs identified in 54 (out of 996) individuals with ASD (Appendix Table 3.4). 39 of 73 *de novo* CNVs had been confirmed as being *de novo* by independent methods, while 34 were deemed to be *de novo* by CNV calling algorithms (28). The Marshall *et al.* (27) set consists of 28 *de novo* CNVs identified in 24 patients. The two patients who had been also included in the AGP set were excluded from the Marshall *et al.* set (Appendix Table 3.5). The Levy *et al.* (114) set consists of 94 CNVs identified in 82 patients (Appendix Table 3.6). The Sanders *et al.* (115) set consists of 67 *de novo* CNVs

identified in 63 patients (Appendix Table 3.7). Note that the individuals analysed in the Sanders *et al.* set wholly overlap with those analysed in the Levy *et al.* set (114)(115). Accordingly, the genomic overlap between these two sets is large compared to that between others (Table 3.3. Herein, *de novo* CNVs identified in patients with ASD will be termed "ASD *dn* CNVs".

	AGP	Marshall	Levy	Sanders
AGP		9.4 Mb (40.9%)	12.6 Mb (54.7%)	12.4 Mb (53.9%)
Marshall	9.4 Mb (7.9%)		49.5 Mb (41.6%)	27.3 Mb (22.9%)
Levy	12.6 Mb (18.2%)	49.5 Mb (71.7%)		60.6 Mb (87.8%)
Sanders	12.4 Mb (13.8%)	27.3 Mb (30.3%)	60.6 Mb (67.3%)	

Table 3.3: Overlapping genomic regions of ASD *dn* CNV sets - This table details the overlapping sequence between the four sets of *de novo* CNVs considered in the study. The percentage shown in brackets gives the proportion of the set listed in the row heading that is overlapped by the set listed in the column headings.

3.3.2 Assigning genes to CNVs

To assure whether expressed coding sequence is affected by ASD *dn* CNVs, a gene was conservatively assigned to a CNV when at least one exon of its every known transcript was overlapped by the CNV (Table 3.2). This gene assignment method did not yield any functional bias under the null (see Appendix text 3.6.1). Duplication and deletion, and their union, CNV sets were considered separately to allow investigation of distinct molecular aetiologies for duplications or deletions. Genes observed to be copy number variable in the same direction (duplication/deletion) in the AGP control CNVs were excluded from the ASD *dn* CNV genes to focus on genes affected by rare *de novo* variants. The coordinates of the genes were extracted from Ensembl Ensmart54 (205).

3.3.3 Mouse Genome Informatics (MGI) phenotypes

The phenotypic information resulting from disruption of mouse orthologues of ASD *dn* CNV genes was obtained from the Mouse Genome Informatics (MGI) resource (206)(17). Using the simple one-to-one gene orthology relationships described in the MGI resource, 5,671 human genes were associated with 5,283 MGI phenotypes that are annotated with Mammalian Phenotype (MP) ontology terms (207). MP ontology is organised at its highest level into 30 over-arching phenotypes, into which more

specific child-terms belong. Among the 8,670 specific MP ontology terms, only 4,055 reasonably populated phenotypes (i.e. a phenotype that includes more than 1% of all genes associated with its over-arching phenotypic category) were considered, to reduce uninformative results and to improve methodological power (208).

3.3.4 Statistical Analysis

Given the small number of ASD *dn* CNVs employed in this study and a lack of control sets of *de novo* CNVs, testing differences between cases and controls was not possible. Instead, I tested a null hypothesis that the frequency of genes overlapping ASD *dn* CNVs that are associated with a (mouse) phenotype is no different from that expected from the whole genome, using the hypergeometric distribution (see section 2.3). A False Discovery Rate (FDR) $< 5\%$ was applied to control for multiple testing (see section 2.3) (131). The same statistical analysis was employed in a set of control CNVs obtained from the AGP consortium (28) for comparison. The analysis pipeline for this study is described in Figure 3.2.

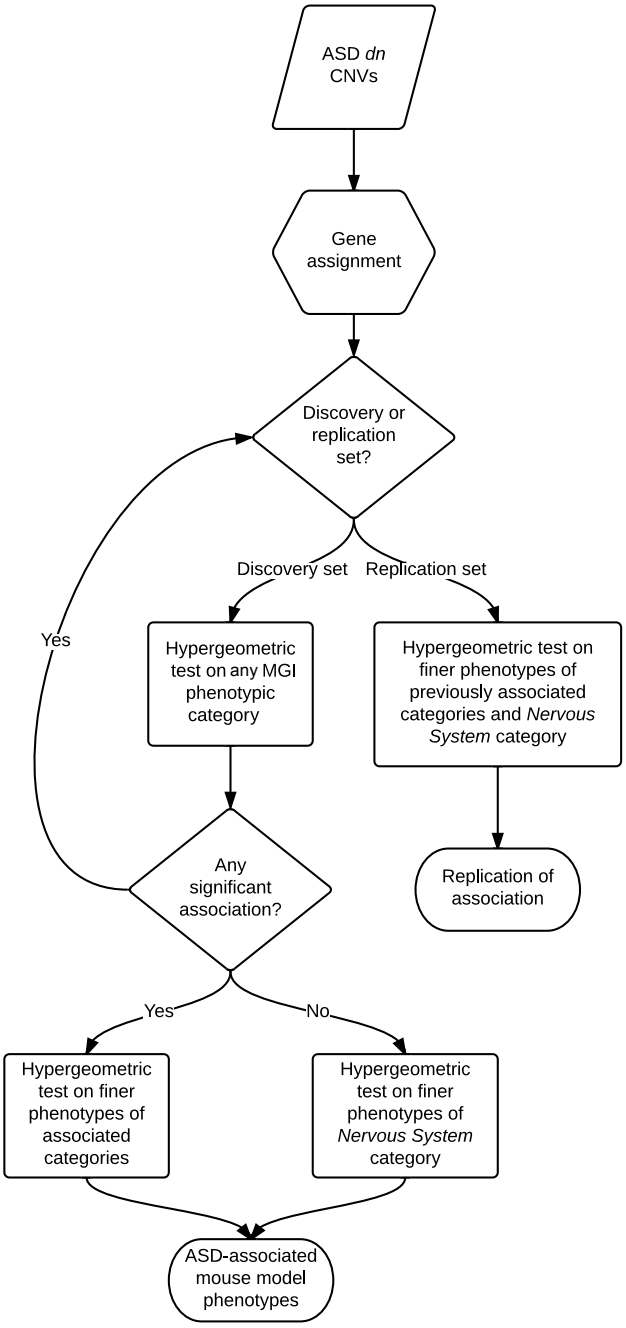


Figure 3.2: Flow diagram of ASD *dn* CNV analysis - 4 ASD *dn* CNV sets are employed in this study. The AGP ASD *dn* CNVs were used as a discovery set. Marshall *et al.*, Levy *et al.* and Sanders *et al.* sets were used as replication sets. Note that the same individuals were analysed in Levy *et al.* and Sanders *et al.* sets, while other sets are independent to one another.

3.4 Results

Previously, Webber *et al.* demonstrated the power of exploiting mouse knock-out models to better understand genetically complex neurological disorders (208). Leveraging this work, I sought to objectively identify autism-related mouse model phenotypes by identifying those knockout model phenotypes enriched among the mouse orthologues of ASD *dn* CNV genes (Figure 3.2). Furthermore, I will propose those genes that contribute to any enrichment as ASD candidate genes. Specifically, I first tested for significant enrichments of 30 broad phenotype categories among the ASD *dn* CNV genes, which then immediately provided the rationale to test the finer phenotypic terms within any significantly enriched phenotypic category. I also tested the finer phenotypic terms of the *Nervous System* category, regardless of the significance of that category's enrichment, based on the assumption that the behavioural presentations of ASD are likely the consequence of functional perturbations of the nervous system. Subsequently, I compared the phenotypic associations of ASD *dn* CNVs with those of an independent set of genes previously implicated in ASD (termed *ASD-implicated*), in order to further corroborate the findings. Finally, I proposed a causal hypothesis for ASD in the cohorts, based on the candidate genes identified through the significant mouse phenotypic enrichments.

3.4.1 Identifying mouse model phenotypes associated with ASD *dn* CNVs

3.4.1.1 Hypothesis-free approach

Among the 30 over-arching phenotypes of MP ontologies, only three categories (*Behavior/Neurological*, *Hearing/Vestibular/Ear* and *Lethality-Postnatal*) were significantly associated with the AGP ASD *dn* CNV genes (FDR<5%; Figure 3.3 and Appendix Table 3.8). Notably, these associations were specific to duplication CNVs (2.0-2.7-fold increases). I then tested finer phenotypic terms within each of the three associated categories (*Behavior/Neurological*, *Hearing/Vestibular/Ear* and *Lethality-Postnatal* categories; 162, 218 and 2 terms, respectively), and identified 23 behavioural, 27 hearing and 1 postnatal-lethality phenotypes as being significantly associated with the AGP ASD *dn* CNV genes (FDR<5%; Figures 3.4 and 3.5, and Appendix Tables 3.9 and 3.10). The 61

mouse model phenotypes enriched among the AGP *dn* CNV genes include specific behavioural phenotypes such as abnormal learning/conditioning, impaired coordination, absence and tonic-clonic seizures (Appendix Table 3.9), and sensorineural/peripheral hearing abnormalities (Figure 3.5 and Appendix Table 3.10).

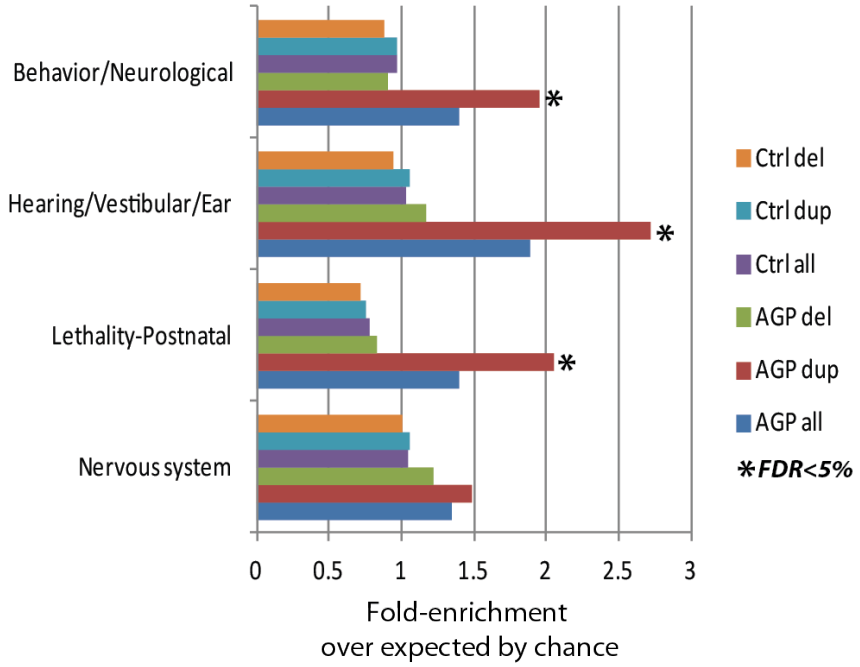


Figure 3.3: Enrichments of phenotypic categories among AGP ASD *dn* CNVs and AGP control CNVs - The figure shows enrichments of three significantly associated mouse phenotypic categories and *Nervous System* category among ASD AGP *dn* CNVs (shown as AGP). The x-axis represents the fold-enrichment of each phenotypic category over expected by chance. 0-fold enrichment indicates the case when none of the genes in a set are annotated with a testing category. Enrichments are shown for copy number duplications and deletions separately (shown as dup and del), and for their union (shown as all). AGP control CNVs (shown as Ctrl) are also shown for comparison. Enrichments that passed false discovery rate (FDR) cutoff at 5% is indicated with asterisk.

Corroboratively, 12 of 17 (71%) nominally significant behavioural phenotypic enrichments identified among Levy *et al.* ASD CNV genes (1.8-5.9-fold enrichments, p<5%, FDR>5%) had been observed previously for the AGP ASD *dn* CNV genes (Appendix Table 3.17). The behavioural phenotypic enrichments in the AGP ASD *dn* CNV genes are validated by these nominally significant associations from Levy *et al.* ASD CNV genes, as they ensure that the 'discovered' associations from the AGP

ASD *dn* CNV genes are likely due to true biological variation rather than some statistical artefacts. Also, all three nominally-associated behavioural phenotypes obtained from Sanders *et al.* ASD *dn* CNV genes had also been observed previously among the identified AGP phenotypic associations, again providing further validation (Appendix Table 3.21). Lastly, phenotypic associations from both Levy *et al.* and Sanders *et al.* ASD *dn* CNV genes showed a general trend also observed among AGP ASD *dn* CNVs whereby many behavioural mouse model phenotypic associations were observed only among duplication CNVs (Appendix Tables 3.9, 3.17 and 3.21, and Figure 3.4).

Importantly, the 61 mouse model phenotypic associations were identified without reference to prior information about human ASD indications. Therefore, these significantly associated behavioural and hearing mouse phenotypes were identified objectively from a hypothesis-free approach and these models and phenotypes may be relevant to the study of autism. It is notable that many of these model phenotypes are directly comparable to the behavioural presentations of human ASD (see Discussion), providing a link between the causal genetic variants and autistic symptoms through the ASD-related mouse models identified here.

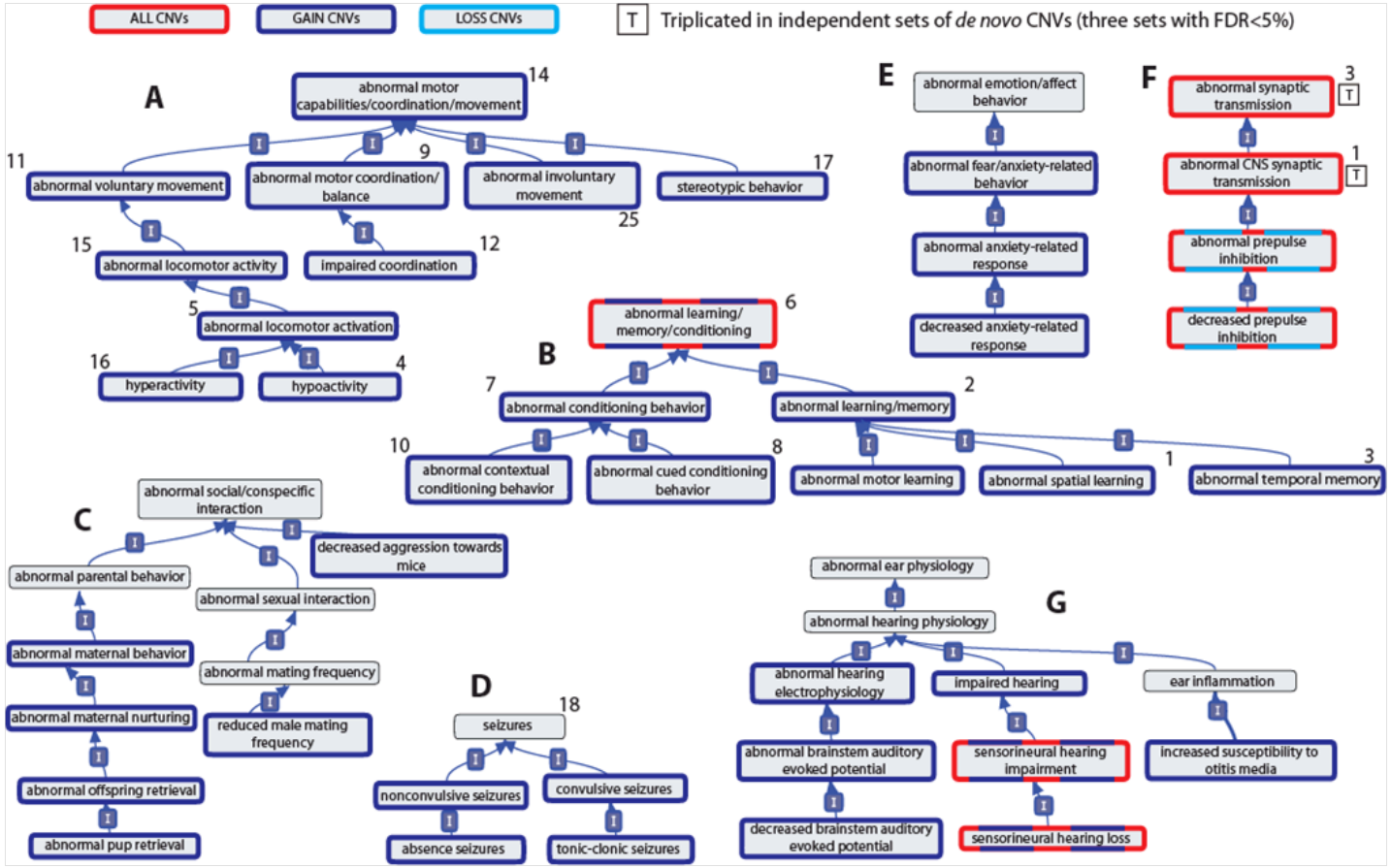


Figure 3.4: Relationships of mouse model phenotypic terms enriched among ASD *dn* CNV genes. - Those terms given within a colored border (red, dark and light blue) are significantly enriched (FDR<5%) among genes overlapped by at least one set of ASD *dn* CNV genes. An arrow between boxes shows the relation from a child-term to a parent-term within the Mammalian Phenotype Ontology. Associated phenotypes from the *Behaviour/Neurological* category are shown in panel A-E, while associated phenotypes from the *Nervous System* category and the *Hearing/Vestibular/Ear* category are shown in panel F and G, respectively. The rank of phenotypic associations among *ASD-implicated* genes is denoted by a number adjacent to phenotypic terms, where applicable (see section 3.4.2).

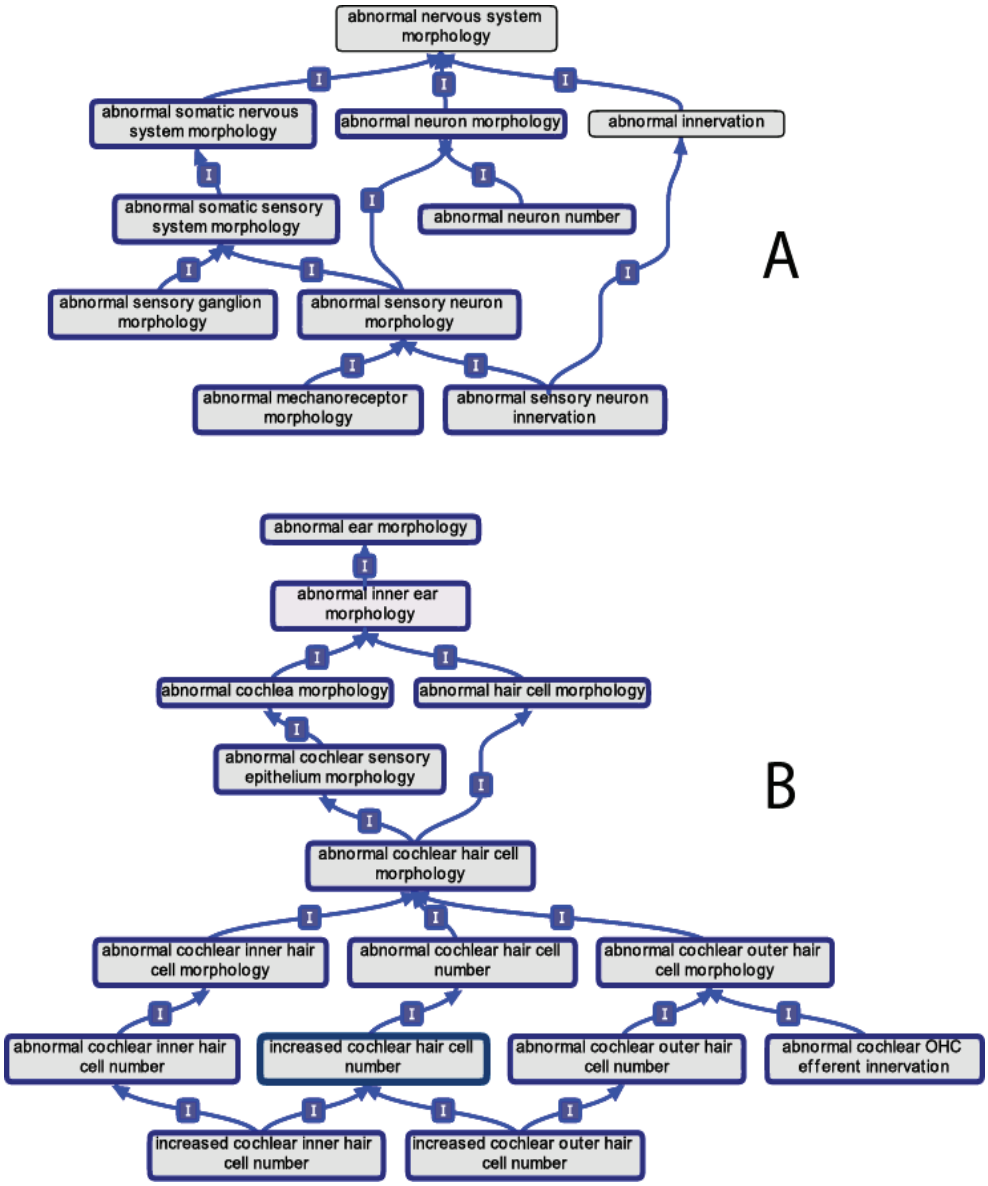


Figure 3.5: Enriched cochlear and mechanoreception-associated mouse model phenotypes among duplication AGP ASD *dn* CNV genes. - Relationships between the phenotypic terms within the Mammalian Phenotype Ontology are indicated by a blue arrow running from the child-term to the parent-term. Terms are significantly enriched (FDR<5%) if they are shown with a thick blue border.

3.4.1.2 Hypothesis-led approach: examining *Nervous System* phenotypes

The hypothesis-free approach, described in the previous section, gave rise to a number of phenotypic associations with human ASD presentations, thereby providing the objective mouse models for human ASD. Although the *Nervous System* phenotypic category did not yield a significant association with ASD *dn* CNV genes, I still examined the finer phenotypes of *Nervous System* category, based on the assumption that the behavioural presentation of ASD is probably due to an aberration of the nervous system.

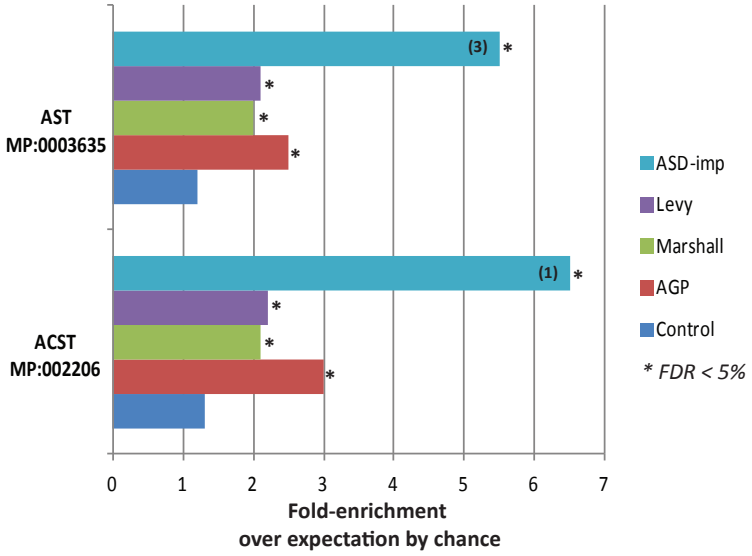


Figure 3.6: Enrichments of the *abnormal CNS synaptic transmission* phenotypes - The figure shows the two triplicated mouse model phenotype associations, the *Abnormal Synaptic Transmission (AST)* and the *Abnormal CNS Synaptic Transmission (ACST)*, with their mammalian phenotype ontology identifiers (MP). The x-axis represents the fold-enrichment of each phenotype over that expected by chance. 0-fold enrichment indicates the case when none of the genes in a set are annotated with a testing phenotype. Numbers in brackets indicate the rank of the mouse phenotypic enrichments within the phenotypic category examined among the *ASD-implicated* genes (see section 3.4.2).

Testing significant enrichments of finer phenotypes of the *Nervous System* category for the AGP ASD *dn* CNV genes identified four abnormal synaptic transmission phenotypes: *abnormal synaptic transmission, abnormal CNS synaptic transmission, abnormal prepulse inhibition* and *decreased prepulse inhibition* [2-10-fold enrichments; FDR<5%] and eight mechanoreception related phenotypes that were specific to the

duplication AGP *dn* CNVs (Figures 3.4 and 3.5 and Appendix Table 3.11).

Notably, the *abnormal synaptic transmission* and the *abnormal CNS synaptic transmission* phenotypic associations were independently replicated in the Marshall *et al.* ASD *dn* CNV genes and the Levy *et al.* ASD *dn* CNV genes. Together, the significant associations of the *abnormal (CNS) synaptic transmission* phenotypes with ASD *dn* CNVs were triplicated in three independent sets (AGP 3.0-fold enrichment, $p = 7 \times 10^{-5}$, FDR<5%; Marshall *et al.* 2.1-fold enrichment, $p = 1 \times 10^{-4}$, FDR<5%; Levy *et al.* 2.2-fold enrichment, $p = 5 \times 10^{-5}$, FDR<5%; Sanders *et al.* 1.6-fold enrichment, $p = 0.008$, FDR>5%; Figures 3.4 and 3.6, and Appendix Tables 3.11, 3.14, and 3.18). In contrast to the behavioural phenotypic associations that were specific to duplication ASD *dn* CNVs, the enrichments of *abnormal (CNS) synaptic transmission* phenotypes were significant for *All* CNVs as well as for deletion ASD *dn* CNVs (Figure 3.4 and Appendix Tables 3.11, 3.14, and 3.18).

3.4.2 Comparison with *ASD-implicated* genes

Examining the model phenotypes frequently observed among the disrupted mouse orthologues of human genes that are already implicated in ASD and comparing these mouse phenotypes with those enriched among ASD *dn* CNV genes may allow an evaluation of the approach employed here. The AGP consortium curated 36 genes that are known to be implicated in ASD (28). Of these, 14 genes, including 10 genes that do not have mouse phenotype information and 4 genes that are also overlapped by ASD *dn* CNVs, were excluded in order to generate a wholly independent set of 22 genes, herein termed *ASD-implicated* genes (Appendix Table 3.25).

Of the 30 over-arching phenotypic categories, 16 were found to be associated with *ASD-implicated* genes (FDR<5%; Appendix Table 3.26). Of these 16 categories, the most enriched were: *Nervous System* [1st], *Behaviour/Neurological* [2nd], *Hearing/Vestibular/Ear* [3rd], *Growth/Size* [4th], *Craniofacial* [5th], and *Lethality-Postnatal* [6th]. Testing finer phenotypes in the *Nervous System* category among *ASD-implicated* genes yielded both the *abnormal synaptic transmission* phenotype (ranked 3rd) and the *abnormal CNS synaptic transmission* phenotype (ranked 1st) as being significantly enriched. These two phenotypes were both associated with each of the three independent sets of ASD *dn* CNV genes (Appendix Tables 3.11, 3.14, 3.18, and 3.27). Moreover, the most 16 enriched finer behavioural phenotypes associated with *ASD-implicated* genes

had also been identified as being significant among ASD *dn* CNV genes (Figure 3.4). Lastly, the *postnatal lethality* phenotype under the *Lethality-Postnatal* category was found to be significantly associated with *ASD-implicated* genes, as it had been for the AGP ASD *dn* CNV genes (Appendix Table 3.29).

3.4.3 Model phenotype associations offer causal hypotheses and aetiological insights

In all, the study has identified 76 different mouse model phenotypes significantly associated with one or more of the three sets of ASD *dn* CNVs (AGP: 76 phenotypes, Marshall *et al.*: 2 phenotypes, Levy *et al.*: 2 phenotypes), while two *abnormal synaptic transmission* phenotypic associations were observed in three independent ASD *dn* CNV sets, and are thus triplicated (Figure 3.6).

The 78 genes that directly contribute to these associations thus become candidate genes whose copy number changes causally underlie ASD (AGP: 39 genes, Marshall *et al.*: 31 genes; Levy *et al.*: 31 genes; Appendix Tables 3.12, 3.15, and 3.19). In turn, these 78 genes offer hypotheses for the causal elements underlying the symptoms of 30% (48/160) of the patients with ASD *dn* CNVs employed here (AGP: 10/54 patients, Marshall *et al.*: 14/24 patients, Levy *et al.*: 24/82 patients; Appendix Tables 3.12, 3.15, and 3.19).

These candidate genes were found within 29% (56/195) of the ASD *dn* CNVs from three independent ASD *dn* CNV sets (AGP:16/73 ASD *dn* CNVs, Marshall *et al.*: 15/28 ASD *dn* CNVs, Levy *et al.*: 25/94 ASD *dn* CNVs; Appendix Tables 3.13, 3.16, and 3.20). The median number of candidate genes per CNV was 2 (ranges 1-8), raising the possibility of epistasis (Appendix Tables 3.13, 3.16, and 3.20). On average, each candidate gene was overlapped by only 1 CNV, with 19% (15/78) of the candidate genes overlapped more than once (Appendix Tables 3.12, 3.15, and 3.19). Note that as all of the individuals from Sanders *et al.* set were also analysed in Levy *et al.* ASD *dn* CNV set, I only considered the larger Levy *et al.* set for the analyses of causal hypothesis and CNV statistics, in order to avoid representing the identical patients or ASD *dn* CNVs more than once. Finally, there was no significant difference in the number of candidate genes overlapped by duplication CNVs compared to deletion CNVs (Kolmogorov-Smirnov test, $p > 0.05$).

3.5 Discussion

3.5.1 Similarities between associated mouse phenotypes and human ASD

Despite the identification of associated mouse model phenotypes in the complete absence of prior assumption of ASD symptoms, the mouse phenotypic associations and human ASD presentations correspond strikingly well.

Considering the three categories of the current diagnosis of ASD (140):

(A) "Qualitative impairment in social interaction" can be compared with the abnormal social/conspecific interaction phenotypes of Figure 3.4C;

(B) "Qualitative impairments in communication" may also be reflected in the abnormal social/conspecific interaction phenotypes of Figure 3.4C;

(C) "Repetitive and stereotyped patterns of behaviour and restricted interest and activities" might be manifested as stereotypic behaviour, abnormal voluntary movement and/or abnormal memory/learning phenotypes of Figures 3.4A and 3.4B.

In addition to the primary ASD symptoms, many of the associated mouse model phenotypes resemble co-morbid symptoms. For example, abnormal learning/ memory phenotypic associations may not only correspond to ASD diagnosis, but also to intellectual disability (ID) that has been observed among ∼55% of patients with ASD (209). Indeed, among 45/54 (83%) individuals from the AGP cohort, who have a record of IQ scores, 23/45 (51%) individuals with ASD also manifest ID. Yet, those individuals with ID were depleted among the ASD individuals whose CNVs yield abnormal learning/memory phenotypic associations (Fisher's exact test, 41% under-expected, $p = 0.2$). This suggests that these findings are more specific to ASD than they are to ID (Appendix Table 3.9).

In addition to ID, more specific behavioural model phenotypes that are analogous to ASD's co-morbid symptoms are associated with the ASD *dn* CNVs:

(A) "Impaired coordination" phenotypes are significantly associated with the AGP ASD *dn* CNV genes (Figure 3.4A and Appendix Table 3.9). ∼80% of young

individuals with ASD have been found to exhibit impaired coordination and motor deficits (210)(211)(212).

(B) "Seizures" phenotypes, specifically absence seizures and tonic-clonic seizures are significantly associated with the AGP ASD *dn* CNVs (Figure 3.4D and Appendix Table 3.9). Correspondingly, it has been estimated that up to ∼40% of ASD individuals suffer seizures (213). Furthermore, absence and tonic-clonic seizures are the predominant types among individuals with ASD (214)(215).

(C) "Anxiety" phenotypes are significantly associated with the AGP ASD *dn* CNVs (Figure 3.4A and Appendix Table 3.9). Again correspondingly, between 42-55% of children with ASD have been shown to present anxiety disorders (216).

(D) "Reduced mating frequency" phenotypes are significantly associated with the AGP ASD *dn* CNVs (Appendix Table 3.9). This phenotypic association is concordant with the notable reduction in number of offsprings observed among autistic individuals (217).

Abnormal hearing phenotypes are also strongly associated with the AGP ASD *dn* CNVs. This result corresponds well to co-morbid hearing abnormalities observed among 33-44% of ASD individuals (218)(219). In particular, the "increased susceptibility to otitis media" phenotypic association is surprisingly similar to the equivalent condition in human that has been shown to manifest in one quarter of ASD cases (220). Many hearing-related phenotypic associations are also relevant to cochlea and mechanoreception (Figure 3.5 and Appendix Table 3.10). This result suggests that the autism community may need to investigate inner ear mechanoreception in ASD anew, despite the mainstream studies on hearing abnormalities in ASD being focused on auditory processing (219). This notion is also supported by the mutations in several genes implicated in ASD (including *CHD7, NIPBL, PTPN11* and *TBX1*) that can cause inner ear abnormalities in humans (221)(222)(223)(224). Moreover, hearing improvements following cochlear implants among ASD individuals with impaired hearing have been reported (225).

The concordance among the associated model phenotypes that are identified from the ASD *dn* CNV genes and the presentation of ASD is further corroborated by the arresting correspondence to the model phenotypes identified as being enriched among the

genes that were previously implicated in ASD: all of the top 16 behavioural phenotypes that are significantly associated with the *ASD-implicated* genes were also associated with ASD *dn* CNV genes (Figure 3.4, Appendix Tables 3.9, 3.11, 3.27, and 3.28). In particular, given the complete separation of the genes identified here and the *ASD-implicated* genes, the independent identification of similar phenotypic associations in both gene sets is striking.

3.5.2 Mouse model phenotypic associations provide insights into the aetiology of ASD

While behavioural and hearing phenotypic associations resemble the primary and secondary symptoms of ASD, the associated nervous system phenotypes may provide an insight into the disorder's aetiology that concurs with the emerging role for synaptic dysfunction in autism (226). While an association of *abnormal synaptic transmission* phenotype with ASD CNVs has been identified previously (227), its replication has not been shown. Here, I triplicated *abnormal synaptic transmission* phenotypic associations and identified more specific phenotypes. For example, ASD *dn* CNV genes considered here were significantly associated with deficits in sensorimotor gating (decreased prepulse inhibition in Figure 3.4F), which has been observed in some mouse models for ASD (228).

Mouse phenotypic associations also provide insights into the genetic architecture of ASD. All genes that form phenotypic associations with ASD co-morbid symptoms (e.g. hearing abnormalities) can also yield phenotypic associations that are relevant to the primary ASD presentation. This implies that copy number change of these genes is affecting more than one phenotype per single gene (pleiotropy). Furthermore, the observation that the median number of candidate genes per CNV is 2 (ranging from 1 to 8) raises the possibility of combinatorial multi-gene effects (epistasis).

Lastly, although there is a considerable genomic overlap among the four sets of ASD *dn* CNVs employed here (8-88%; Table 3.3), the candidate genes are mostly private to an individual patient, as the average number of CNVs that overlap each candidate gene is only 1 (Appendix Tables 3.12, 3.15, 3.19, and 3.23). This is consistent with the assumption that individually rare variants may contribute to ASD by disrupting a common biological process (see section 1.2.4).

3.5.3 Linking mouse genotype to ASD phenotype

Despite the obvious resemblance between the human ASD presentations and the associated mouse behavioural phenotypes, it is important to note that most mouse model phenotypes result from gene losses whereas behavioural phenotypic associations were observed among duplication ASD *dn* CNVs. Although partial duplications of a gene may cause a gene disruption, this scenario is not applicable here because the significant associations identified for the duplication ASD *dn* CNV genes remains when considering only those genes that are wholly overlapped, and are thus duplicated, by these CNVs (data not shown).

In many clinical and experimental settings, overlaps between phenotypes arisen by deletion and those by duplication have been observed. For example, there are a number of reciprocal microduplication syndromes that frequently affect the same organ systems and manifest similar behavioural phenotypes. Both 15q11-q13 deletion involving *UBE3A* (Angelman syndrome/Prader-Willi syndrome) and its duplication (15q11-q13 duplication syndrome) exhibit similar behavioural symptoms and are strongly associated with ASD. Likewise, 17p11.2 deletion involving *RAI1* (Smith-Magenis syndrome) and duplication of the same region (Potocki-Lupski syndrome) are associated with ASD (see Introduction). In addition, two recent genetic mouse models for ASD highlight gene dosage dependent phenotypes. Smith *et al.* have shown that increased gene dosage of *UBE3A* results in autism traits and decreased glutamate synaptic transmission in mice (229), and Horev *et al.* have shown that both deletion and duplication of 16p11.2 loci affect behaviours in mice (230).

Such observations may be compatible with a possible scenario, the so-called "Goldilocks effect of gene dosage". In this scenario, the optimum dosage of critical gene products follow peaked distributions rather than flat distributions, hence an increase or decrease in the level of gene products is harmful. This is also consistent with that the density of many proteins within the synapse appear to be stringently controlled (231). As multiprotein complexes are likely to maintain the stoichiometry of their subunits to perform their biological function (232), an imbalanced stoichiometry of a protein complex caused by either excess or depletion of its subunits may lead to deleterious effects. This possibility is also supported by the *UBE3A* duplication mouse models for ASD

(229), in which the protein product of *UBE3A* is known to form ubiquitin complex in order to perform its molecular function (233)(234).

While gene dosage imbalance may be one possibility, the issue arising from the presence of either duplication or deletion among ASD patients will be more directly addressed in Chapter 4. In any case, if there were no correlations between the phenotypes resulting from deletion of mouse genes and those resulting from duplication of the human orthologues, I ought not to have been able to detect phenotypic associations, let alone ones that resemble ASD in such detail.

3.5.4 Post hoc assessment of methodology

3.5.4.1 Handling incomplete annotation of mouse phenotype

MGI phenotypes that are annotated with too few genes are less informative in identifying ASD-associated phenotypes. Thus, in order to avoid extreme cases of incomplete phenotyping and to improve statistical power, the study considers only those phenotypes that have been annotated with a reasonable number of genes (i.e. >1% of all the genes annotated with an over-arching category, e.g. behavioural phenotypes with more than 19 associated genes; see Materials and Methods). However, there were some relevant mouse model phenotypes that were annotated with too few genes to be considered by this approach. For example, when all the behavioural phenotypes were tested for the *ASD-implicated* genes, the *abnormal vocalisation* phenotype was additionally identified as a significant phenotypic association. The *abnormal vocalisation* phenotypes are typically identified by an increase in frequency of the pup's ultrasonic vocalisations after separation from the litter (199). In fact, the *abnormal maternal nurturing* and the *pup retrieval* phenotypes were also significantly associated with the ASD *dn* CNV genes (Figure 3.4) and this may be a consequence of abnormal pup vocalisations (235).

3.5.4.2 Handling effect of gene and CNV size

Considering that the size of ASD *dn* CNVs analysed in this study can be as small as 30kb, it is possible that a CNV falls wholly within an intron, hence potentially leaving the spliced transcript product unaffected. Moreover, even if a CNV overlaps exonic regions, the affected sequence may be alternatively spliced and expressed in a subset of transcripts for that gene. In particular, neuronal and tissue-specific genes

often contain many large introns and contain a large number of alternatively spliced forms (236)(237). Therefore, in order to ensure that the expressed sequence of a gene is affected by a given CNV, at least one coding exon of every known transcript of that gene was required to be overlapped by a CNV. Only then was that gene considered to have been affected by a CNV in this study (Table 3.2).

Another caveat in gene assignment is that among genes that are overlapped by CNVs a higher proportion of brain-expressed genes may be expected by chance. This is because brain-expressed genes are often much longer than genes that are not brain-expressed (medians 37.1Kb vs. 25.8Kb respectively, Mann-Whitney U-test $p = 8 \times 10^{-4}$; brain expression defined as in (208)). However, under the CNV gene-assignment approach in this study, the length of protein-coding sequence was found not to be significantly different between brain-expressed genes and those that are not brain-expressed (medians 455 vs. 433 residues, respectively, Mann-Whitney U-test $p = 0.56$). Therefore, it is unlikely that the gene assignment in this study was biased by longer brain-expressed genes.

3.5.4.3 The null of mouse phenotypic association test

The noisy distribution of the p-values obtained from the hypergeometric test on the randomised sets of genomic intervals that match with the ASD *dn* CNVs in size (Appendix Figure 3.7) indicates how unlikely the phenotypic association can be detected simply by chance. If there are no artefactual biases in a given test, the null (i.e. no association) should provide uniformly distributed p-values. If the null provides an enrichment of low p-values, it indicates that there is a higher chance of false-positive results. The null of the test employed in this study provided a contrary distribution, showing an unexpectedly large fraction of p-values approaching 1 (Appendix Figure 3.7). Such p-value distributions diminish the concerns of a high false-positive rate, and thus illustrate how gene-enriched the ASD *dn* CNVs are for the associated phenotypic categories.

3.5.4.4 Limitations

Whilst the approach employed in this study is objective and robust, the identified phenotypic associations are limited in these respects:

- the mouse models currently available (17),
- the extent of behaviours examinable and examined (238), and
- the available phenotypic classification within the MP ontology (207).

More specifically, hemizygous mouse models would have been more relevant than homozygous mouse models when comparing patients' genotypes. Also, a finer classification of both mouse and human phenotypes would have allowed the comparison of individual patient's symptoms to the mouse models where the genic disruption is comparable. However, each of these limitations is the subject of ongoing research, and hence likely subject to further improvement. Therefore, a frequent update of the set of ASD-associated mouse models employing the same methodological framework, taking advantage of evolving mouse phenotype database, would likely identify additional ASD-related mouse models and phenotypes.

3.5.5 Conclusions

The approach employed in this study objectively identifies mouse phenotypic associations to ASD, a complex behavioural human disorder, based on genetic commonalities. This method generates novel hypotheses including many novel candidate genes, a role for mechanoreception in ASD associated hearing abnormalities, and potential connections between abnormal vocalisation and abnormal pup retrieval or nurturing phenotypes, as well as supporting the emerging roles of pleiotropy, epistasis and abnormal synaptic transmission in ASD. This approach may be revisited as the human genetic and mouse phenotype data increase, in order to account and evaluate the contribution of available genetic mouse models to the understanding of ASD. Also, a similar approach may be applied to other complex behavioural disorders whose underlying genetics is complicated.

3.6 Appendices

3.6.1 Confirming the null

In order to confirm that the gene assignment method and statistical test which were employed in this study, are unbiased, it needs to be examined whether the expected p-values form the uniform distribution under the null hypothesis (i.e. no association) (239). Gene sets assigned from 500 sets of randomly-distributed genomic regions that match in size and number to the ASD *dn* Copy Number Variant Regions (CNVRs) were tested for associations with particular MGI phenotypic categories, i.e. Behavior/Neurological, Nervous System, and Hearing/Vestibular/Ear categories. While the consequent distributions of the randomised p-values fluctuate (Appendix Figure 3.7, modified from (239)), there is no evident excess of low p-values that might indicate that phenotypic associations with ASD *dn* CNVs that were detected in this study could have been a chance observation. In fact, there was an increase in frequency of p-values approaching 1, in all three phenotypic categories with both duplication and deletion CNVRs (Appendix Figure 3.7).

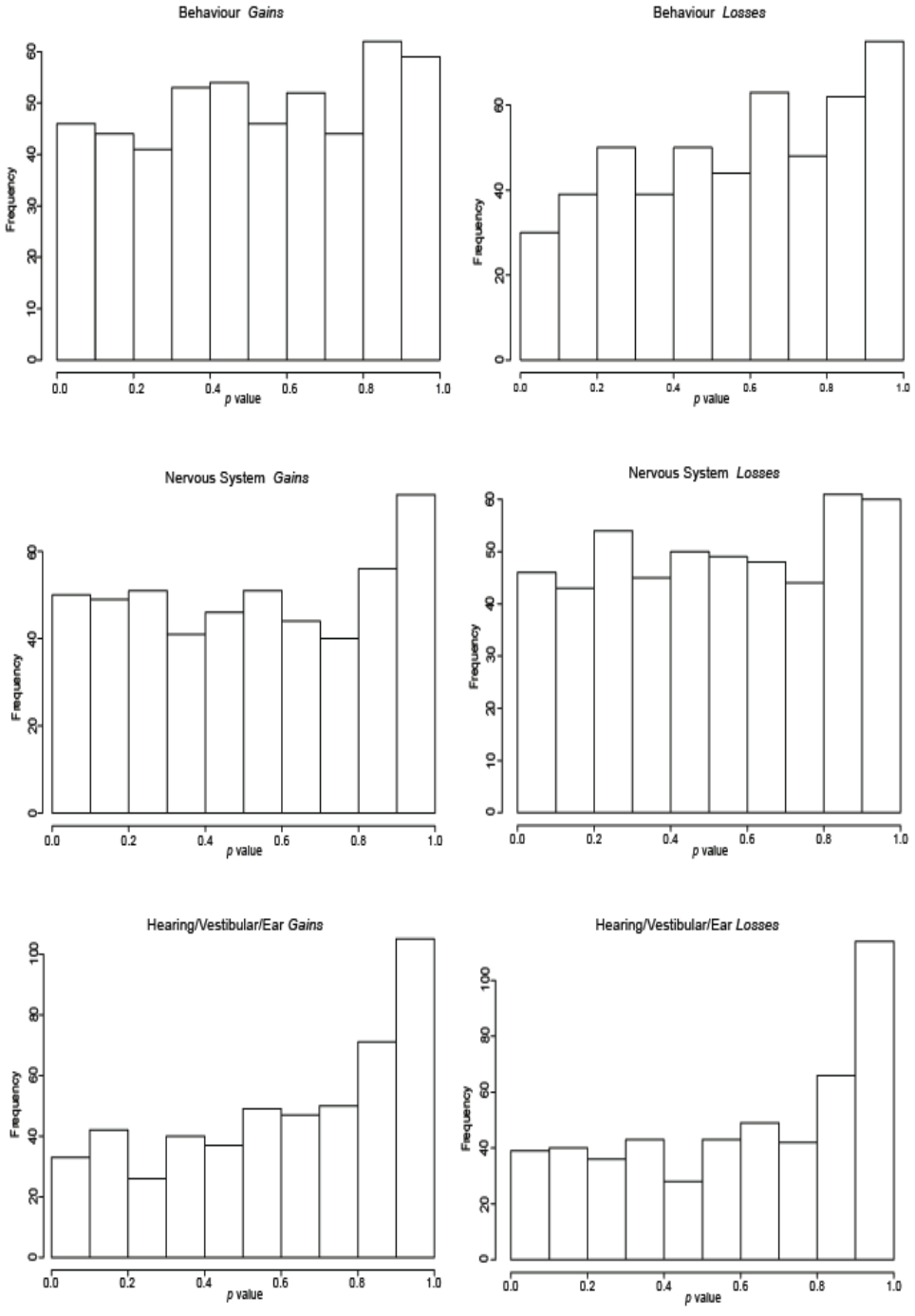


Figure 3.7: Distributions of p-values for the surfeits of genes associated with 3 phenotypic categories obtained from 500 randomised sets of case-matched CNVRs - 500 sets of genomic regions matched in size and number to each of the duplication and the deletion CNVRs were obtained. For each set, the likelihoods for the observed surfeit of genes associated with a particular phenotypic category, namely Behavior(/Neurological), Nervous System and Hearing/Vestibular/Ear are shown. Adapted from (239).

3.6.2 ASD *dn* CNV sets

Table 3.4: 73 *De novo* CNVs acquired from the AGP consortium (hg18)

Chr	Start	End	Copy Change	Inheritance
1	1803260	2029579	Gain	Potential_DeNovo
1	2320050	2497275	Gain	Potential_DeNovo
1	3567650	3649602	Gain	Potential_DeNovo
1	44205616	44239193	Gain	Confirmed_DeNovo
1	193185872	193314559	Gain	Potential_DeNovo
2	50493827	50677835	Gain	Confirmed_DeNovo
2	50539877	50730546	Loss	Confirmed_DeNovo
2	50990306	51222043	Loss	Confirmed_DeNovo
2	51002576	51157742	Loss	Confirmed_DeNovo
2	213551117	213740232	Loss	Potential_DeNovo
3	19921061	20096832	Loss	Confirmed_DeNovo
4	932439	1016866	Gain	Potential_DeNovo
4	1711332	1779448	Gain	Potential_DeNovo
5	130112368	130176516	Loss	Potential_DeNovo
5	180499295	180540489	Gain	Confirmed_DeNovo
6	26256979	26416296	Loss	Potential_DeNovo
6	33399849	33512042	Loss	Confirmed_DeNovo
6	66342679	66383666	Gain	Potential_DeNovo
7	1289102	1573785	Gain	Potential_DeNovo
7	44216390	44301067	Gain	Potential_DeNovo
7	61100583	62082083	Gain	Potential_DeNovo
7	100094164	100212023	Gain	Potential_DeNovo
7	102699832	102798745	Loss	Confirmed_DeNovo
7	124270458	124342396	Loss	Potential_DeNovo
7	143501355	143706682	Loss	Confirmed_DeNovo
7	143501355	143706682	Loss	Confirmed_DeNovo
7	143501355	143706682	Loss	Potential_DeNovo
7	143501355	143706682	Loss	Potential_DeNovo
7	151632121	151701091	Gain	Potential_DeNovo
7	153775586	153844747	Loss	Confirmed_DeNovo
8	704383	1521910	Gain	Confirmed_DeNovo
8	48631388	48802529	Gain	Confirmed_DeNovo
8	145938543	146118298	Gain	Confirmed_DeNovo
8	146206368	146264218	Gain	Potential_DeNovo
9	98998	3682923	Loss	Confirmed_DeNovo
9	98998	334508	Loss	Confirmed_DeNovo
9	138505259	139336068	Gain	Potential_DeNovo
10	6127800	6170083	Loss	Confirmed_DeNovo
10	22622727	22679909	Loss	Confirmed_DeNovo
10	22626925	22719183	Loss	Confirmed_DeNovo
11	70119917	70187872	Loss	Confirmed_DeNovo
11	70154458	70220632	Loss	Confirmed_DeNovo
12	7626398	7690853	Loss	Potential_DeNovo

Continued on next page

Chr	Start	End	Copy Change	Inheritance
12	7644562	7686003	Loss	Potential_DeNovo
12	9772366	9836626	Loss	Potential_DeNovo
12	9772366	9823337	Loss	Potential_DeNovo
12	28394989	28481418	Loss	Potential_DeNovo
12	31925819	32031365	Gain	Potential_DeNovo
12	103369862	103476369	Gain	Potential_DeNovo
13	99476752	99526202	Gain	Potential_DeNovo
14	20279711	20345174	Gain	Confirmed_DeNovo
14	106180071	106358708	Gain	Potential_DeNovo
15	21190624	26203954	Gain	Confirmed_DeNovo
15	42350178	42385934	Loss	Confirmed_DeNovo
15	42741567	42805781	Loss	Potential_DeNovo
15	91200007	91283004	Loss	Confirmed_DeNovo
16	29502984	30127026	Loss	Confirmed_DeNovo
16	29502984	30210849	Gain	Confirmed_DeNovo
16	29554843	30195224	Loss	Confirmed_DeNovo
16	31909325	33520331	Gain	Confirmed_DeNovo
16	32396948	33382661	Loss	Confirmed_DeNovo
16	54355871	54390068	Loss	Confirmed_DeNovo
16	54355871	54390068	Loss	Potential_DeNovo
17	34612208	34732327	Gain	Confirmed_DeNovo
17	76953064	77782267	Gain	Confirmed_DeNovo
17	77785939	77849717	Loss	Confirmed_DeNovo
19	24183904	24421125	Gain	Confirmed_DeNovo
20	14545734	14948785	Loss	Confirmed_DeNovo
21	33043726	33086703	Loss	Potential_DeNovo
21	43042997	43326347	Gain	Confirmed_DeNovo
22	17241748	19819918	Loss	Confirmed_DeNovo
X	55600108	55654095	Gain	Potential_DeNovo
X	103056385	103195953	Gain	Potential_DeNovo

Table 3.5: 28 *De novo* CNVs acquired from Marshall *et al.* (hg18)

Chr	Start	End	Copy Change
1	97332167	98424667	loss
2	186556739	186653869	loss
3	15125800	16535400	loss
3	65286300	70633200	loss
5	81949	13882933	loss
5	9275811	12705200	loss
7	108393666	119417172	loss
7	113528285	129015006	loss
8	710491	1501580	gain
12	40584198	41007040	loss
12	114191663	132287723	gain
15	18376200	30298800	gain
15	18427100	30298847	gain

Continued on next page

Chr	Start	End	Copy Change
15	69601300	73890800	loss
16	29559989	30085308	gain
16	29559989	30235818	loss
16	29652700	30085300	gain
16	87800593	88066260	loss
18	55756601	76115600	loss
20	60949339	62377000	gain
21	42974148	43328084	gain
22	17020346	19791646	loss
22	17270946	21552208	gain
22	17270946	21552208	gain
22	46335545	49565822	loss
22	49299969	49576671	loss
X	34419	6009994	loss
X	48202290	52850670	gain

Table 3.6: 94 *De novo* CNVs acquired from Levy *et al.* (hg18)

Chr	Start	End	Copy Change
1	30898119	36080200	gain
1	238596122	238912325	gain
1	240710016	240976706	gain
2	47680808	47695811	loss
2	50683648	50727454	loss
2	110091402	110500910	loss
3	4133021	4854563	loss
3	5013460	5026532	gain
3	37253000	37428450	loss
3	67219468	72341775	loss
3	80657349	80688814	loss
3	106452066	106531952	loss
3	159967432	159983514	loss
3	165652624	165673098	loss
3	187095345	187213250	loss
3	197208363	198838422	loss
5	11403621	11493124	loss
7	66529731	66739034	gain
7	72357744	73781556	gain
7	72377168	73780165	gain
8	3965290	4351494	loss
8	39293378	39301771	loss
9	192721	3376926	gain
9	13291055	13310257	loss
9	30917196	30942603	loss
9	94827643	96115846	gain
9	129907917	130512360	gain
9	133765650	133931148	gain

Continued on next page

Chr	Start	End	Copy Change
9	139800008	140140081	loss
10	52697241	54411314	gain
10	68176528	68240025	loss
10	120741108	120971953	gain
11	71711547	71723201	loss
11	84231213	84264089	gain
12	3526622	3538037	loss
12	16907527	29882113	loss
12	121182783	121212600	loss
13	104561136	104615019	gain
15	20093109	20621734	gain
15	20310487	20640026	loss
15	21224486	26313261	gain
15	28863993	30241335	gain
15	66886601	71939036	loss
16	21530	1255851	loss
16	8716070	9209667	gain
16	8806159	9153955	gain
16	15490860	16291179	gain
16	29557094	30107502	loss
16	29557094	30084175	gain
16	29557094	30107502	gain
16	29559250	30111196	gain
16	29559250	30107502	loss
16	29559250	30191661	loss
16	29563679	30107502	loss
16	29572667	30209382	loss
16	29572667	30111196	loss
16	29581119	30112681	gain
16	76500379	76591088	loss
16	79735268	84751023	loss
16	88324486	88336614	loss
17	6032329	7294541	gain
17	31808637	33295813	loss
17	41042187	41523831	loss
17	42135186	42476630	gain
17	76632298	76681110	loss
18	48171087	48214763	gain
18	53608263	53637885	loss
19	42369734	42488525	loss
22	17249508	18693261	loss
22	19037910	19795854	loss
22	26917065	26941353	loss
22	38825382	39101009	loss
X	22854451	23212135	loss
X	70296173	70329989	loss
X	153229884	153433101	gain

Continued on next page

Chr	Start	End	Copy Change
2	123891130	123938079	loss
2	131193793	131651161	gain
3	2426703	2464661	loss
3	182690570	182808377	gain
3	184618001	184691966	gain
4	94017702	94063088	loss
5	119960275	120033947	loss
6	98759619	98777549	gain
6	162942422	162986755	loss
7	11701698	11824943	loss
9	23639426	23673145	loss
10	30091165	30115310	gain
10	52697241	54411314	gain
13	101034390	101077585	loss
16	15419888	16200651	gain
17	55435679	57633705	gain
18	6477860	7903233	gain
19	559002	879090	gain
19	42606909	42648735	loss

Table 3.7: 67 *De novo* CNVs acquired from Sanders *et al.* (hg18)

Chr	Start	End	Copy Change
chr1	30897868	36080484	gain
chr1	144612035	146325568	gain
chr1	144838175	146324832	gain
chr1	238597914	238908329	gain
chr1	240718979	240968200	gain
chr2	61390488	61873147	gain
chr2	103264154	108718935	loss
chr2	110050606	111021091	loss
chr3	4133580	4852348	loss
chr3	37257074	37432212	loss
chr3	67223272	72356021	loss
chr3	106456026	106535466	loss
chr3	187095966	187210274	loss
chr3	197161073	198851029	loss
chr4	35410	3511385	gain
chr4	184857633	184952586	gain
chr5	11403359	11491117	loss
chr6	57325295	58188685	gain
chr6	61942912	62748639	gain
chr7	72300351	73782113	gain
chr7	72344426	73782113	gain
chr7	72355583	73782113	gain
chr7	72411506	73811186	gain
chr7	151632121	151695916	loss

Continued on next page

Chr	Start	End	Copy Change
chr9	175632	3373495	gain
chr9	129909379	130509108	gain
chr9	133779666	133924807	gain
chr9	139799894	140192015	loss
chr10	52699516	54408816	gain
chr10	120748083	120937654	gain
chr11	120769289	122402940	loss
chr12	16906534	29889066	loss
chr13	40002610	40810340	gain
chr15	19767013	26134114	gain
chr15	28714502	30348312	loss
chr15	28723577	30238780	gain
chr15	66887071	71944199	loss
chr16	835	1253638	loss
chr16	3416079	3443739	loss
chr16	8720164	9203038	gain
chr16	8828382	9147487	gain
chr16	15352729	16370751	gain
chr16	29499858	30123122	loss
chr16	29499858	30112803	loss
chr16	29499858	30107306	loss
chr16	29499858	30107306	loss
chr16	29502984	30239704	gain
chr16	29502984	30147029	loss
chr16	29502984	30107306	gain
chr16	29554843	30123122	loss
chr16	29554843	30107306	loss
chr16	29554843	30170811	gain
chr16	29563365	30130862	gain
chr16	68529466	71494580	loss
chr16	76502154	76593063	loss
chr16	79738681	84748965	loss
chr16	81796275	81830296	loss
chr17	6058839	7292887	gain
chr17	31889664	33570026	loss
chr19	14910717	15747565	gain
chr19	60124341	60265445	loss
chr20	14616243	14751454	loss
chr20	28251057	35143867	gain
chr20	40621732	40719457	loss
chr22	17257787	19793730	loss
chr22	38826602	39095056	loss
chrX	22860224	23214712	loss

3.6.3 Mouse phenotypic enrichments and associated genes

Table 3.8: Phenotypic category enrichments for AGP ASD *dn* CNVs and control CNVs. - Category, MGI phenotypic category; dup, duplication; del, deletion; all, union of dup and del; Ctrl, AGP control CNVs; Fold-change, fold enrichment of number of genes associated with a phenotype over that expected by chance.

Category	CNVs	Fold-change	P-value
Behavior/Neurological	AGP all	1.39	1.07×10^{-2}
	AGP dup	1.95	9.55×10^{-5}
	AGP del	0.91	7.13×10^{-1}
	Ctrl all	0.97	7.32×10^{-1}
	Ctrl dup	0.97	6.81×10^{-1}
	Ctrl del	0.88	8.64×10^{-1}
Hearing/Vestibular/Ear	AGP all	1.89	1.11×10^{-2}
	AGP dup	2.72	2.58×10^{-3}
	AGP del	1.17	4.24×10^{-1}
	Ctrl all	1.04	4.02×10^{-1}
	Ctrl dup	1.06	3.47×10^{-1}
	Ctrl del	0.94	6.35×10^{-1}
Lethality-Postnatal	AGP all	1.40	5.63×10^{-2}
	AGP dup	2.06	2.94×10^{-3}
	AGP del	0.83	7.68×10^{-1}
	Ctrl all	0.78	9.98×10^{-1}
	Ctrl dup	0.76	9.98×10^{-1}
	Ctrl del	0.72	9.71×10^{-1}
Nervous System	AGP all	1.35	8.54×10^{-3}
	AGP dup	1.49	1.36×10^{-2}
	AGP del	1.23	1.48×10^{-1}
	Ctrl all	1.04	2.10×10^{-1}
	Ctrl dup	1.05	1.83×10^{-1}
	Ctrl del	1.01	5.07×10^{-1}

In the following phenotypic enrichments tables, PhenotypeName (Abbreviation: Total Genes in Genome with Phenotype) provides the phenotype name. Within the brackets, the phenotype abbreviation is given first, while the second number gives the total number of human genes with a mouse orthologue that yields this phenotype when disrupted. Copy Change gives the direction of copy change of the CNVs whose overlapping genes possess this enrichment. Fold change is the fold-enrichment of genes associated with this phenotype over that expected from the genome by chance (see Materials and Methods). P-value is the associated probability for this enrichment (hypergeometric probability, FDR<5%). # Genes forming Enrichment is the number of overlapped genes that contribute to this enrichment. Total Genes in CNV set gives the number of genes with this CNV set for which phenotypic information is available. # ASD implicated genes hit gives the number of genes within this enrichment that also belong to a set of 26 genes that have both been previously implicated in ASD as defined by the AGP Consortium and for which phenotypic information is available.

Table 3.9: Significantly enriched mouse model behavioural phenotypes among the orthologues of AGP *de novo* CNVs genes

PhenotypeID	PhenotypeName (Abbreviation: Total Genes in Genome with Phenotype)	Copy change	Fold change	P-value	# Genes forming Enrichment	Total Genes in CNV set	# ASD implicated genes hit
MP:0002066	abnormal motor capabilities/ coordination/ movement (BehAMC/C/M:1219)	gain	2.57	5.20×10^{-6}	21	38	1
MP:0003491	abnormal voluntary movement (BehAVM:940)	gain	2.86	9.23×10^{-6}	18	38	1
MP:0004924	abnormal behavior (BehAB:1766)	gain	2.03	4.50×10^{-5}	24	38	1
MP:0002063	abnormal learning/ memory/ conditioning (BehAL/M/C:375)	all	2.95	7.15×10^{-5}	16	82	1
MP:0001392	abnormal locomotor activity (BehALAy:877)	gain	2.72	7.19×10^{-5}	16	38	1
MP:0001405	impaired coordination (BehIC:261)	gain	4.57	2.65×10^{-4}	8	38	1
MP:0003313	abnormal locomotor activation (BehALAn:603)	gain	2.97	3.74×10^{-4}	12	38	1
MP:0001463	abnormal spatial learning (BehASL:150)	gain	5.97	4.25×10^{-4}	6	38	1
MP:0003216	absence seizures (BehnsAS:25)	gain	17.91	5.77×10^{-4}	3	38	1
MP:0002804	abnormal motor learning (BehAML:25)	gain	17.91	5.77×10^{-4}	3	38	1
MP:0002063	abnormal learning/ memory/ conditioning (BehAL/M/C:375)	gain	3.58	6.41×10^{-4}	9	38	1
MP:0001513	limb grasping (BehLG:115)	gain	6.49	9.24×10^{-4}	5	38	1
MP:0000948	nonconvulsive seizures (BehnsNS:30)	gain	14.92	9.95×10^{-4}	3	38	1
MP:0001449	abnormal learning/ memory (BehAL/M:324)	gain	3.68	1.13×10^{-3}	8	38	1
MP:0002062	abnormal conditioning behavior (BehACB:185)	gain	4.84	1.28×10^{-3}	6	38	1

Continued on next page

PhenotypeID	PhenotypeName (Abbreviation: Total Genes in Genome with Phenotype)	Copy change	Fold change	P-value	# Genes forming Enrichment	Total Genes in CNV set	# ASD implicated genes hit
MP:0003312	abnormal locomotor coordination (BehALC:445)	gain	3.02	2.17×10^{-3}	9	38	1
MP:0002067	abnormal sensory capabilities/ reflexes/ nociception (BehASC/R/N:491)	gain	2.74	4.24×10^{-3}	9	38	1
MP:0001386	abnormal maternal nurturing (BehAMN:105)	gain	5.69	5.05×10^{-3}	4	38	0
MP:0001402	hypoactivity (BehH:247)	gain	3.63	5.49×10^{-3}	6	38	1
MP:0002573	behavioral despair (BehBD:17)	gain	17.56	5.58×10^{-3}	2	38	0
MP:0001380	reduced male mating frequency (BehRMF:17)	gain	17.56	5.58×10^{-3}	2	38	0
MP:0008779	abnormal maternal behavior (BehAMB:109)	gain	5.48	5.77×10^{-3}	4	38	0
MP:0009731	abnormal offspring retrieval (BehAOR:18)	gain	16.58	6.25×10^{-3}	2	38	0
MP:0001384	abnormal pup retrieval (BehAPR:18)	gain	16.58	6.25×10^{-3}	2	38	0

Table 3.10: Significantly enriched mouse model hearing/vestibular/ear phenotypes among the orthologues of AGP *de novo* CNV genes

PhenotypeID	PhenotypeName (Abbreviation: Total Genes in Genome with Phenotype)	Copy change	Fold change	P-value	# Genes forming Enrichment	Total Genes in CNV set	# ASD implicated genes hit
MP:0002622	abnormal cochlear hair cell morphology (hvensACHCM:132)	gain	7.91	2.18×10^{-5}	7	38	0
MP:0006329	sensorineural hearing impairment (hveSHI:26)	gain	22.96	2.31×10^{-5}	4	38	0
MP:0004740	sensorineural hearing loss (hveSHL:26)	gain	22.96	2.31×10^{-5}	4	38	0

Continued on next page

PhenotypeID	PhenotypeName (Abbreviation: Total Genes in Genome with Phenotype)	Copy change	Fold change	P-value	# Genes forming Enrichment	Total Genes in CNV set	# ASD implicated genes hit
MP:0006329	sensorineural hearing impairment (hveSHI:26)	all	13.30	2.90×10^{-5}	5	82	1
MP:0004740	sensorineural hearing loss (hveSHL:26)	all	13.30	2.90×10^{-5}	5	82	1
MP:0003308	abnormal cochlear sensory epithelium morphology (hveACSEM:138)	gain	7.57	2.92×10^{-5}	7	38	0
MP:0000042	abnormal organ of Corti (hveAOC:139)	gain	7.52	3.05×10^{-5}	7	38	0
MP:0000045	abnormal hair cell morphology (hvensAHCM:139)	gain	7.52	3.05×10^{-5}	7	38	0
MP:0004632	abnormal cochlear OHC efferent innervation (hveACOOEI:10)	gain	44.77	3.23×10^{-5}	3	38	0
MP:0004399	abnormal cochlear outer hair cell morphology (hvensACOHCM:107)	gain	8.37	6.62×10^{-5}	6	38	0
MP:0003169	abnormal scala media morphology (hveASMM:161)	gain	6.49	7.84×10^{-5}	7	38	0
MP:0004288	abnormal spiral ligament morphology (hveASLM:15)	gain	29.85	1.20×10^{-4}	3	38	0
MP:0000031	abnormal cochlea morphology (hveACM:175)	gain	5.97	1.33×10^{-4}	7	38	0
MP:0004401	increased cochlear outer hair cell number (hveICOHCN:17)	gain	26.34	1.77×10^{-4}	3	38	0
MP:0004632	abnormal cochlear OHC efferent innervation (hveACOOEI:10)	all	20.75	3.25×10^{-4}	3	82	0
MP:0004490	type IV spiral ligament fibrocyte degeneration (hveT4SLFD:5)	gain	59.69	4.32×10^{-4}	2	38	0
MP:0000026	abnormal inner ear morphology (hveAIEM:213)	gain	4.90	4.46×10^{-4}	7	38	0
MP:0004407	increased cochlear hair cell number (hvensICHCN:25)	gain	17.91	5.77×10^{-4}	3	38	0

Continued on next page

PhenotypeID	PhenotypeName (Abbreviation: Total Genes in Genome with Phenotype)	Copy change	Fold change	P-value	# Genes forming Enrichment	Total Genes in CNV set	# ASD implicated genes hit
MP:0004296	abnormal type IV spiral ligament fibrocytes (hveAT4SLF:6)	gain	49.75	6.45×10^{-4}	2	38	0
MP:0004765	decreased brainstem auditory evoked potential (hveDBAEP:107)	gain	6.97	6.66×10^{-4}	5	38	0
MP:0006325	impaired hearing (hveIH:125)	gain	5.97	1.35×10^{-3}	5	38	0
MP:0004400	abnormal cochlear outer hair cell number (hvensACOHCN:36)	gain	12.44	1.70×10^{-3}	3	38	0
MP:0002102	abnormal ear morphology (hveAEM:352)	gain	3.39	1.92×10^{-3}	8	38	0
MP:0004738	abnormal brainstem auditory evoked potential (hveABAEP:136)	gain	5.49	1.96×10^{-3}	5	38	0
MP:0004292	abnormal spiral ligament fibrocyte morphology (hveASLFM:11)	gain	27.13	2.31×10^{-3}	2	38	0
MP:0004393	abnormal cochlear inner hair cell morphology (hvensACIHCM:88)	gain	6.78	2.67×10^{-3}	4	38	0
MP:0006335	abnormal hearing electrophysiology (hveAHE:152)	gain	4.91	3.18×10^{-3}	5	38	0
MP:0001850	increased susceptibility to otitis media (hveISOM:14)	gain	21.32	3.78×10^{-3}	2	38	0
MP:0004406	abnormal cochlear hair cell number (hveACHCN:51)	gain	8.78	4.63×10^{-3}	3	38	0
MP:0004395	increased cochlear inner hair cell number (hveICIHCN:16)	gain	18.65	4.95×10^{-3}	2	38	0

Table 3.11: Significantly enriched mouse model nervous phenotypes among the orthologues of AGP *de novo* CNV genes

PhenotypeID	PhenotypeName (Abbreviation: Total Genes in Genome with Phenotype)	Copy change	Fold change	P-value	# Genes forming Enrichment	Total Genes in CNV set	# ASD implicated genes hit
MP:0003088	abnormal prepulse inhibition (nsAPI:48)	all	8.64	5.74×10^{-5}	6	82	1
MP:0009142	decreased prepulse inhibition (nsDPI:30)	all	11.53	6.00×10^{-5}	5	82	1
MP:0002206	abnormal CNS synaptic transmission (nsACST:373)	all	2.97	6.71×10^{-5}	16	82	2
MP:0003635	abnormal synaptic transmission (nsAST:438)	all	2.53	4.34×10^{-4}	16	82	2
MP:0003088	abnormal prepulse inhibition (nsAPI:48)	loss	13.43	2.98×10^{-5}	5	44	1
MP:0009142	decreased prepulse inhibition (nsDPI:30)	loss	17.18	7.46×10^{-5}	4	44	1
MP:0002622	abnormal cochlear hair cell morphology (hvensACHCM:132)	gain	7.91	2.18×10^{-5}	7	38	0
MP:0000045	abnormal hair cell morphology (hvensAHCM:139)	gain	7.52	3.05×10^{-5}	7	38	0
MP:0004399	abnormal cochlear outer hair cell morphology (hvensACOHCM:107)	gain	8.37	6.62×10^{-5}	6	38	0
MP:0000972	abnormal mechanoreceptor morphology (nsAMM:157)	gain	6.65	6.68×10^{-5}	7	38	0
MP:0000965	abnormal sensory neuron morphology (nsASNM:384)	gain	3.89	1.53×10^{-4}	10	38	0
MP:0000959	abnormal somatic sensory system morphology (nsASSSM:481)	gain	3.41	2.12×10^{-4}	11	38	0
MP:0002882	abnormal neuron morphology (nsANM:958)	gain	2.49	2.12×10^{-4}	16	38	1
MP:0003216	absence seizures (BehnsAS:25)	gain	17.91	5.77×10^{-4}	3	38	1

Continued on next page

PhenotypeID	PhenotypeName (Abbreviation: Total Genes in Genome with Phenotype)	Copy change	Fold change	P-value	# Genes forming Enrichment	Total Genes in CNV set	# ASD implicated genes hit
MP:0000948	nonconvulsive seizures (BehnsNS:30)	gain	14.92	9.95×10^{-4}	3	38	1
MP:0000960	abnormal sensory ganglion morphology (nsASGM:189)	gain	4.74	1.44×10^{-3}	6	38	0
MP:0004400	abnormal cochlear outer hair cell number (hvensACOHCN:36)	gain	12.44	1.70×10^{-3}	3	38	0
MP:0008946	abnormal neuron number (nsANN:267)	gain	3.91	1.70×10^{-3}	7	38	0
MP:0000968	abnormal sensory neuron innervation (nsASNI:82)	gain	7.28	2.06×10^{-3}	4	38	0
MP:0002752	abnormal somatic nervous system morphology (nsASNSM:640)	gain	2.57	2.39×10^{-3}	11	38	0
MP:0004393	abnormal cochlear inner hair cell morphology (hvensACIHCM:88)	gain	6.78	2.67×10^{-3}	4	38	0

GENE	Number of CNVs	Patients
ACTG1	1	5444_3
AGPAT2	1	1946_301
ARHGDIA	1	5444_3
B4GALT2	1	5412_3
CAMK2B	1	3228_004
DGCR8	1	3183_007
DOC2A	2	5359_4,5068_3
FBXL20	1	5056_4
FGFR3	1	3228_004
FGFRL1	1	3228_004
FSCN2	1	5444_3
GABRA5	1	13050_593
GABRB3	1	13050_593
GABRD	1	3228_004
GNB1L	1	3183_007
HERC2	1	13050_593
HES5	1	3228_004
HGS	1	5444_3
IDUA	1	3228_004
MAFG	1	5444_3
MAFK	1	3228_004
MAPK3	2	5359_4,5068_3
NDN	1	13050_593
NOTCH1	1	1946_301
NPB	1	5444_3
OCA2	1	13050_593
PDE6G	1	5444_3
RAC3	1	5444_3
RTN4R	1	3183_007
SEPT5	1	3183_007
SEZ6L2	2	5359_4,5068_3
SLC6A9	1	5412_3
SNRPN	1	13050_593
SYNGAP1	1	5353_3
TACC3	1	3228_004
TBX1	1	3183_007
TP73	1	3228_004
UBE3A	1	13050_593
ZDHHC8	1	3183_007

Table 3.12: Candidate genes from AGP *de novo* CNVs - *GENE* is the gene name. Number of CNVs gives the number of CNVs in this set that candidate gene overlaps. Patients lists the patients possessing these CNVs. Patients 1946_301, 3183_007, 3228_004, 5068_3, 5359_4 have an IQ>70, whereas patients 5056_4, 5353_3 present IQ<70. IQ scores were not available for the other patients.

CNV	Number Candidate Genes hit
13050_593_chr15_gain_21190624	7
1946_301_chr9_gain_138505259	2
3183_007_chr22_loss_17241748	6
3228_004_chr1_gain_1803260	1
3228_004_chr1_gain_2320050	1
3228_004_chr1_gain_3567650	1
3228_004_chr4_gain_1711332	2
3228_004_chr4_gain_932439	2
3228_004_chr7_gain_1289102	1
3228_004_chr7_gain_44216390	1
5056_4_chr17_gain_34612208	1
5068_3_chr16_loss_29502984	3
5353_3_chr6_loss_33399849	1
5359_4_chr16_loss_29554843	3
5412_3_chr1_gain_44205616	2
5444_3_chr17_gain_76953064	8

Table 3.13: AGP *de novo* CNVs and number of candidate genes in each CNV - CNV gives the CNV from this set that is overlapped by a candidate gene. The CNV format is: [$cnv_identifier$]_chr[$chromosome$]_[$gain/loss$]_[$chromosome\ start\ position$]. Number of Candidate Genes hit is the number of genes that contribute to the significant enrichments observed for this set of CNVs that are overlapped by this particular CNV.

PhenotypeID	PhenotypeName (Total Genes in Genome with Phenotype)	Copy change	Fold change	P-value	Number of Genes forming Enrichment	Total Genes in CNV set	Number of ASD implicated genes hit
MP:0003635	abnormal synaptic transmission (nsAST: 438)	all	1.99	1.47×10^{-4}	31	202	2
MP:0002206	abnormal CNS synaptic transmission (nsACST: 373)	all	2.11	1.17×10^{-4}	28	202	2

Table 3.14: Significantly enriched mouse model nervous phenotypes among the orthologues of Marshall *et al. de novo* CNV genes.

GENE	Number of CNVs	Patients
APBA2	2	SK0245-005,SK0073-003
CADPS2	1	SK0131-003
CAMKK2	1	MM0278-003
CHRNA4	1	MM0109-003
CPLX3	1	SK0243-003
CTNND2	2	SK0205-004,SK0152-003
DGCR8	1	SK0119-003
DNAJC5	1	MM0109-003
DOC2A	1	SK0019-004
EEF1A2	1	MM0109-003
FOXP2	2	SK0083-003,SK0131-003
GABRA5	2	SK0245-005,SK0073-003
GABRB3	2	SK0245-005,SK0073-003
GNB1L	1	SK0119-003
KCND2	1	SK0131-003
LEP	1	SK0131-003
MAPK3	1	SK0019-004
MAPK8IP2	2	NA0039-000,MM0109-003
MBP	1	SK0218-003
NETO1	1	SK0218-003
NOS1	1	MM0278-003
OPRL1	1	MM0109-003
P2RX4	1	MM0278-003
PTPRZ1	1	SK0131-003
SEPT5	1	SK0119-003
SEZ6L2	1	SK0019-004
SLC6A3	1	SK0205-004
SYP	1	SK0306-004
TBX1	1	SK0119-003
UBE3A	2	SK0245-005,SK0073-003
ZDHHC8	1	SK0119-003

Table 3.15: Candidate genes from Marshall *et al.* *de novo* CNVs - see Appendix Table 3.12.

CNV	Number Candidate Genes hit
MM0109-003_chr20_gain_60949339	4
MM0109-003_chr22_loss_49299969	1
MM0278-003_chr12_gain_114191663	3
NA0039-000_chr22_loss_46335545	1
SK0019-004_chr16_loss_29559989	3
SK0073-003_chr15_gain_18376200	4
SK0083-003_chr7_loss_108393666	1
SK0119-003_chr22_loss_17020346	5
SK0131-003_chr7_loss_113528285	5
SK0152-003_chr5_loss_9275811	1
SK0205-004_chr5_loss_81949	2
SK0218-003_chr18_loss_55756601	2
SK0243-003_chr15_loss_69601300	1
SK0245-005_chr15_gain_18427100	4
SK0306-004_chrX_gain_48202290	1

Table 3.16: Marshall *et al.* *de novo* CNVs and number of candidate genes in each CNV - see Appendix Table 3.13.

Table 3.17: Nominally significantly enriched mouse model behavioural phenotypes among the orthologues of Levy *et al. de novo* CNV genes

PhenotypeID	PhenotypeName (Abbreviation: Total Genes in Genome with Phenotype)	Copy change	Fold change	P-value	# Genes forming Enrichment	Total Genes in CNV set	# ASD implicated genes hit
MP:0002063	abnormal learning/ memory/ conditioning (BehAL/M/C: 375)	all	1.88	1.86×10^{-3}	24	193	2
MP:0002062	abnormal conditioning behavior (BehACB: 185)	all	2.22	3.98×10^{-3}	14	193	2
MP:0001449	abnormal learning/ memory (BehAL/M: 324)	all	1.81	6.67×10^{-3}	20	193	2
MP:0001463	abnormal spatial learning (BehASL: 150)	all	2.35	4.83×10^{-3}	12	193	2
MP:0001454	abnormal cued conditioning behavior (BehACuCB: 63)	all	3.73	1.21×10^{-3}	8	193	1
MP:0001469	abnormal contextual conditioning behavior (BehACoCB: 83)	all	2.48	2.22×10^{-2}	7	193	2
MP:0001468	abnormal temporal memory (BehATM: 93)	all	2.21	3.83×10^{-2}	7	193	2
MP:0003863	decreased aggression towards mice (BehDATM: 19)	all	4.64	2.52×10^{-2}	3	193	0
MP:0000948	nonconvulsive seizures (BehNS: 30)	gain	5.73	1.49×10^{-2}	3	99	1
MP:0003216	absence seizures (BehAS: 25)	gain	6.87	8.98×10^{-3}	3	99	1
MP:0002063	abnormal learning/ memory/ conditioning (BehAL/M/C: 375)	gain	1.83	2.92×10^{-2}	12	99	1
MP:0002062	abnormal conditioning behavior (BehACB: 185)	gain	2.48	1.50×10^{-2}	8	99	1
MP:0001463	abnormal spatial learning (BehASL: 150)	gain	2.29	4.65×10^{-2}	6	99	1
MP:0001380	reduced male mating frequency (BehRMF: 17)	gain	6.74	3.46×10^{-2}	2	99	0

Continued on next page

PhenotypeID	PhenotypeName (Abbreviation: Total Genes in Genome with Phenotype)	Copy change	Fold change	P-value	# Genes forming Enrichment	Total Genes in CNV set	# ASD implicated genes hit
MP:0001454	abnormal cued conditioning behavior (BehACuCB: 63)	gain	5.46	7.39×10^{-4}	6	99	0
MP:0009745	abnormal behavioral response to xenobiotics (BehABRX: 193)	gain	2.37	1.89×10^{-2}	8	99	0
MP:0003863	decreased aggression towards mice (BehDATM: 19)	gain	6.03	4.25×10^{-2}	2	99	0
MP:0002063	abnormal learning/ memory/ conditioning (BehAL/M/C: 375)	loss	1.93	2.02×10^{-2}	12	94	1
MP:0001449	abnormal learning/ memory (BehAL/M: 324)	loss	1.86	4.10×10^{-2}	10	94	1
MP:0001463	abnormal spatial learning (BehASL: 150)	loss	2.41	3.76×10^{-2}	6	94	1
MP:0001364	decreased anxiety-related response (BehDARR: 62)	loss	3.89	1.90×10^{-2}	4	94	0

Table 3.18: Significantly enriched mouse model nervous phenotypes among the orthologues of Levy *et al. de novo* CNV genes

PhenotypeID	PhenotypeName (Abbreviation: Total Genes in Genome with Phenotype)	Copy change	Fold change	P-value	# Genes forming Enrichment	Total Genes in CNV set	# ASD implicated genes hit
MP:0003635	abnormal synaptic transmission (nsAST: 438)	all	2.08	6.12×10^{-5}	31	193	3
MP:0003635	abnormal synaptic transmission (nsAST: 438)	gain	2.09	3.47×10^{-3}	16	99	1
MP:0003635	abnormal synaptic transmission (nsAST: 438)	loss	2.07	5.19×10^{-3}	15	94	2
MP:0002206	abnormal CNS synaptic transmission (nsACST: 373)	all	2.21	5.17×10^{-5}	28	193	3
MP:0002206	abnormal CNS synaptic transmission (nsACST: 373)	gain	2.15	4.97×10^{-3}	14	99	1

Continued on next page

PhenotypeID	PhenotypeName (Abbreviation: Total Genes in Genome with Phenotype)	Copy change	Fold change	P-value	# Genes forming Enrichment	Total Genes in CNV set	# ASD implicated genes hit
MP:0002206	abnormal CNS synaptic transmission (nsACST: 373)	loss	2.26	3.08×10^{-3}	14	94	2
MP:0003088	abnormal prepulse inhibition (nsAPI: 48)	all	3.67	5.36×10^{-3}	6	193	1
MP:0003088	abnormal prepulse inhibition (nsAPI: 48)	loss	6.28	1.10×10^{-3}	5	94	1
MP:0009142	decreased prepulse inhibition (nsDPI: 30)	all	4.90	3.09×10^{-3}	5	193	1
MP:0009142	decreased prepulse inhibition (nsDPI: 30)	loss	8.04	1.39×10^{-3}	4	94	1

GENE	Number of CNVs	Patients
BAIAP2	1	11186
BBS4	1	11233
CACNA1B	1	12581
CACNA1H	1	11435
CHRNB1	1	11532
CRHR1	1	11454
CTNND2	1	12289
CYLN2	2	11154,11129
DGCR8	1	12239
DLG4	1	11532
DLGAP3	1	11080
DNM1	1	12235
DOC2A	6	11680,12451,12100,11433,11540,11090
FZD9	2	11154,11129
GABRA5	1	12007
GABRB3	1	12007
GDI1	1	11092
GNB1L	1	12239
GRK5	1	11718
GTF2IRD1	2	11154,11129
HCN2	1	12295
HCRTR1	1	11080
ITPR1	1	11046
KISS1R	1	12295
LIMK1	2	11154,11129
MAPK3	6	11680,12451,12100,11433,11540,11090
MAPT	1	11454
NCDN	1	11080
NDN	1	12007
NLGN2	1	11532
NLGN3	1	11689
PRKG1	1	11962
RTN4R	1	12239
SDC3	1	11080
SEPT5	1	12239
SEZ6L2	6	11680,12451,12100,11433,11540,11090
STX1A	2	11154,11129
TBX1	1	12239
UBE3A	1	12007
VLDLR	1	11435
ZDHHC8	1	12239

Table 3.19: Candidate genes from Levy *et al.* *de novo* CNVs - see Appendix Table 3.12.

CNV	Number Candidate Genes hit
11046_chr3_loss_4133021	1
11080_chr1_gain_30898119	4
11090_chr16_loss_29559250	3
11092_chrX_gain_153229884	1
11129_chr7_gain_72357744	5
11154_chr7_gain_72377168	5
11186_chr17_loss_76632298	1
11233_chr15_loss_66886601	1
11433_chr16_loss_29563679	3
11435_chr16_loss_21530	1
11435_chr9_gain_192721	1
11454_chr17_loss_41042187	2
11532_chr17_gain_6032329	3
11540_chr16_loss_29557094	3
11680_chr16_loss_29559250	3
11689_chrX_loss_70296173	1
11718_chr10_gain_120741108	1
11962_chr10_gain_52697241	1
12007_chr15_gain_21224486	4
12100_chr16_loss_29572667	3
12235_chr9_gain_129907917	1
12239_chr22_loss_17249508	6
12289_chr5_loss_11403621	1
12295_chr19_gain_559002	2
12451_chr16_loss_29572667	3
12581_chr9_loss_139800008	1

Table 3.20: Levy *et al. de novo* CNVs and number of candidate genes in each CNV - see Appendix Table 3.13.

PhenotypeID	PhenotypeName (Total Genes in Genome with Phenotype)	Copy change	Fold change	P-value	# Genes forming Enrichment	Total Genes in CNV set	# ASD implicated genes hit
MP:0002063	abnormal learning/ memory/ conditioning (BehAL/M/C: 375)	all	1.47	3.79×10^{-2}	23	236	1
MP:0002063	abnormal learning/ memory/ conditioning (BehAL/M/C: 375)	gain	1.68	3.77×10^{-2}	14	126	1
MP:0001463	abnormal spatial learning (BehASL: 150)	all	1.92	2.18×10^{-2}	12	236	1
MP:0001463	abnormal spatial learning (BehASL: 150)	gain	2.40	1.80×10^{-2}	8	126	1
MP:0001454	abnormal cued conditioning behavior (BehACuCB: 63)	all	2.29	4.59×10^{-2}	6	236	0
MP:0001454	abnormal cued conditioning behavior (BehACuCB: 63)	gain	3.57	1.26×10^{-2}	5	126	0

Table 3.21: Nominally significantly enriched mouse model behavioural phenotypes among the orthologues of Sanders *et al. de novo* CNV genes

PhenotypeID	PhenotypeName (Total Genes in Genome with Phenotype)	Copy change	Fold chage	P-value	# Genes forming Enrichment	Total Genes in CNV set	# ASD implicated genes hit
MP:0003635	abnormal synaptic transmission (nsAST: 438)	all	1.59	7.84×10^{-3}	29	236	2
MP:0003635	abnormal synaptic transmission (nsAST: 438)	loss	1.73	2.45×10^{-2}	15	112	1
MP:0002206	abnormal CNS synaptic transmission (nsACST: 373)	all	1.61	1.15×10^{-2}	25	236	2
MP:0002206	abnormal CNS synaptic transmission (nsACST: 373)	loss	1.76	3.14×10^{-2}	13	112	1
MP:0003088	abnormal prepulse inhibition (nsAPI: 48)	all	3.00	1.38×10^{-2}	6	236	1
MP:0003088	abnormal prepulse inhibition (nsAPI: 48)	loss	5.27	2.39×10^{-3}	5	112	1
MP:0009142	decreased prepulse inhibition (nsDPI: 30)	all	4.00	7.27×10^{-3}	5	236	1
MP:0009142	decreased prepulse inhibition (nsDPI: 30)	loss	6.75	2.66×10^{-3}	4	112	1

Table 3.22: Nominally significantly enriched mouse model nervous phenotypes among the orthologues of Sanders *et al. de novo* CNV genes

GENE	Number of CNVs	Patients
CACNA1B	1	12581.p1
CACNA1H	1	11435.p1
CALB2	1	12691.p1
CHRNA7	1	13301.p1
CHRNB1	1	11532.p1
CTNND2	1	12289.p1
CYLN2	4	11154.p1,11129.p1,12594.p1,12955.p1
DGCR8	1	12239.p1
DLG4	1	11532.p1
DLGAP3	1	11080.p1
DNM1	1	12235.p1
DOC2A	7	12100.p1,11090.p1,11680.p1,11433.p1,12451.p1,12308.p1,11540.p1
FZD9	4	11154.p1,11129.p1,12594.p1,12955.p1
GABRA5	1	12007.p1
GABRB3	1	12007.p1
GNB1L	1	12239.p1
GTF2IRD1	4	11154.p1,11129.p1,12594.p1,12955.p1
HTT	1	13367.p1
IDUA	1	13367.p1
ITPR1	1	11046.p1
LIMK1	4	11154.p1,11129.p1,12594.p1,12955.p1
LRPAP1	1	13367.p1
MAPK3	7	12100.p1,11090.p1,11680.p1,11433.p1,12451.p1,12308.p1,11540.p1
NCDN	1	11080.p1
NDN	1	12007.p1
NLGN2	1	11532.p1
PRKG1	1	11962.p1
RTN4R	1	12239.p1
SDC3	1	11080.p1
Sep-05	1	12239.p1
SEZ6L2	7	12100.p1,11090.p1,11680.p1,11433.p1,12451.p1,12308.p1,11540.p1
SLC5A7	1	12735.p1
SNTA1	1	13036.p1
STX1A	4	11154.p1,11129.p1,12594.p1,12955.p1
TBX1	1	12239.p1
UBE3A	1	12007.p1
VLDLR	1	11435.p1
ZDHHC8	1	12239.p1

Table 3.23: Candidate genes from Sanders *et al.* *de novo* CNVs - see Appendix Table 3.12.

CNV	Number Of Candidate Genes hit
11046.p1_chr3_loss_4133580	1
11080.p1_chr1_gain_30897868	3
11090.p1_chr16_loss_29499858	3
11129.p1_chr7_gain_72355583	5
11154.p1_chr7_gain_72300351	5
11433.p1_chr16_loss_29554843	3
11435.p1_chr16_loss_835	1
11435.p1_chr9_gain_175632	1
11532.p1_chr17_gain_6058839	3
11540.p1_chr16_loss_29502984	3
11680.p1_chr16_loss_29554843	3
11962.p1_chr10_gain_52699516	1
12007.p1_chr15_gain_19767013	4
12100.p1_chr16_loss_29499858	3
12235.p1_chr9_gain_129909379	1
12239.p1_chr22_loss_17257787	6
12289.p1_chr5_loss_11403359	1
12308.p1_chr16_loss_29499858	3
12451.p1_chr16_loss_29499858	3
12581.p1_chr9_loss_139799894	1
12594.p1_chr7_gain_72344426	5
12691.p1_chr16_loss_68529466	1
12735.p1_chr2_loss_103264154	1
12955.p1_chr7_gain_72411506	5
13036.p1_chr20_gain_28251057	1
13301.p1_chr15_loss_28714502	1
13367.p1_chr4_gain_35410	3

Table 3.24: Sanders *et al.* *de novo* CNVs and number of candidate genes in each CNV - see Appendix Table 3.13.

Gene Name
AFF2
AHI1
ATRX
CACNA1C
CHD7
CNTNAP2
CREBBP
DHCR7
DMD
DMPK
FMR1
IL1RAPL1
MECP2
NF1
NIPBL
NLGN3
NRXN1
NSD1
PTEN
PTPN11
TSC1
TSC2

Table 3.25: Twenty two human genes deemed previously to have been implicated in ASD (termed *ASD-implicated* genes) - These genes are a subset of 36 disease genes deemed causally implicated in ASD by the Autism Genome Project (28), and were formed by retaining only those genes whose mouse orthologues' disruption had been phenotyped and discarding those genes that were overlapped by ASD *dn* CNVs considered in this study.

Table 3.26: Significantly enriched mouse model phenotype categories among *ASD-implicated* genes - The probability is that obtained through the hypergeometric test (FDR<5%). Total Genes with Phenotype gives the total number of human genes whose mouse orthologue is associated with this phenotype. Set Genes with Phenotype is the total number of genes within the set whose mouse orthologue is associated with this phenotype. Fold change over expected is that obtained when compared with the genome as a whole sampled randomly.

Probability	Phenotype ID	Description	Total Genes with Phenotype	Set Genes with Phenotype	Fold change
6.12×10^{-6}	MP:0003631	nervous system phenotype	2208	19	2.22
6.26×10^{-4}	MP:0005386	behavior/neurological phenotype	1841	15	2.10
6.86×10^{-4}	MP:0005377	hearing/vestibular/ear phenotype	549	8	3.76
7.91×10^{-4}	MP:0005382	craniofacial phenotype	711	9	3.26
8.49×10^{-4}	MP:0005378	growth/size phenotype	2387	17	1.84
1.09×10^{-3}	MP:0005373	lethality-postnatal	1089	11	2.60
1.81×10^{-3}	MP:0005369	muscle phenotype	969	10	2.66
2.84×10^{-3}	MP:0005388	respiratory system phenotype	848	9	2.74
2.93×10^{-3}	MP:0005395	other phenotype	159	4	6.48
4.03×10^{-3}	MP:0005385	cardiovascular system phenotype	1475	12	2.10
6.58×10^{-3}	MP:0005371	limbs/digits/tail phenotype	612	7	2.95
8.20×10^{-3}	MP:0005379	endocrine/exocrine gland phenotype	1182	10	2.18
1.23×10^{-2}	MP:0005380	embryogenesis phenotype	1251	10	2.06
2.08×10^{-2}	MP:0005375	adipose tissue phenotype	424	5	3.04
2.26×10^{-2}	MP:0005367	renal/urinary system phenotype	772	7	2.34
2.31×10^{-2}	MP:0005372	life span-post-weaning/aging	1369	10	1.88

Table 3.27: Significantly enriched mouse model nervous phenotypes among *ASD-implicated* genes - see Appendix Table 3.26.

Probability	Phenotype ID	Description	Total Genes with Phenotype	Set Genes with Phenotype	Fold change
4.30×10^{-7}	MP:0002206	abnormal CNS synaptic transmission	373	10	6.91

Continued on next page

Probability	Phenotype ID	Description	Total Genes with Phenotype	Set Genes with Phenotype	Fold change
1.56×10^{-6}	MP:0003633	abnormal nervous system physiology	1154	15	3.35
1.89×10^{-6}	MP:0003635	abnormal synaptic transmission	438	10	5.89
1.51×10^{-5}	MP:0000771	abnormal brain size	148	6	10.45
3.07×10^{-5}	MP:0002152	abnormal brain morphology	1064	13	3.15
6.87×10^{-5}	MP:0000783	abnormal forebrain morphology	654	10	3.94
1.40×10^{-4}	MP:0005238	increased brain size	27	3	28.64
2.79×10^{-4}	MP:0003632	abnormal nervous system morphology	1723	15	2.24
3.06×10^{-4}	MP:0008143	abnormal dendrite morphology	35	3	22.09
3.44×10^{-4}	MP:0008540	abnormal cerebrum morphology	370	7	4.88
4.78×10^{-4}	MP:0000929	open neural tube	177	5	7.28
5.98×10^{-4}	MP:0002151	abnormal neural tube morphology/development	405	7	4.46
9.44×10^{-4}	MP:0000774	decreased brain size	117	4	8.81
1.66×10^{-3}	MP:0003634	abnormal glial cell morphology	233	5	5.53
1.73×10^{-3}	MP:0000807	abnormal hippocampus morphology	235	5	5.48
1.86×10^{-3}	MP:0000801	abnormal temporal lobe morphology	239	5	5.39
1.97×10^{-3}	MP:0000787	abnormal telencephalon morphology	495	7	3.65
2.08×10^{-3}	MP:0003720	abnormal neural tube closure	245	5	5.26
2.16×10^{-3}	MP:0003861	abnormal nervous system development	816	9	2.84
2.44×10^{-3}	MP:0005442	abnormal pyramidal neuron morphology	71	3	10.89
2.64×10^{-3}	MP:0000812	abnormal dentate gyrus morphology	73	3	10.59
3.03×10^{-3}	MP:0004166	abnormal limbic system morphology	267	5	4.83

Continued on next page

Probability	Phenotype ID	Description	Total Genes with Phenotype	Set Genes with Phenotype	Fold change
3.03×10^{-3}	MP:0008946	abnormal neuron number	267	5	4.83
3.18×10^{-3}	MP:0002064	seizures	270	5	4.77
3.50×10^{-3}	MP:0002207	abnormal long term potentiation	167	4	6.17
4.08×10^{-3}	MP:0000951	sporadic seizures	25	2	20.62

Table 3.28: Significantly enriched mouse model behavioural phenotypes among *ASD-implicated* genes - see Appendix Table 3.26.

Probability	Phenotype ID	Description	Total Genes with Phenotype	Set Genes with Phenotype	Fold change
1.63×10^{-5}	MP:0001463	abnormal spatial learning	150	6	10.31
1.64×10^{-5}	MP:0001449	abnormal learning/ memory	324	8	6.36
2.25×10^{-5}	MP:0001468	abnormal temporal memory	93	5	13.86
2.66×10^{-5}	MP:0001402	hypoactivity	247	7	7.31
3.38×10^{-5}	MP:0003313	abnormal locomotor activation	603	10	4.27
4.77×10^{-5}	MP:0002063	abnormal learning/ memory/ conditioning	375	8	5.50
5.36×10^{-5}	MP:0002062	abnormal conditioning behavior	185	6	8.36
8.71×10^{-5}	MP:0001454	abnormal cued conditioning behavior	63	4	16.37
1.73×10^{-4}	MP:0001516	abnormal motor coordination/ balance	582	9	3.99
2.55×10^{-4}	MP:0001469	abnormal contextual conditioning behavior	83	4	12.42
2.93×10^{-4}	MP:0003491	abnormal voluntary movement	940	11	3.02
3.59×10^{-4}	MP:0001405	impaired coordination	261	6	5.93
3.78×10^{-4}	MP:0004924	abnormal behavior	1766	15	2.19
6.84×10^{-4}	MP:0002066	abnormal motor capabilities/ coordination/ movement	1219	12	2.54
8.12×10^{-4}	MP:0001392	abnormal locomotor activity	877	10	2.94
9.78×10^{-4}	MP:0001399	hyperactivity	207	5	6.23

Continued on next page

Probability	Phenotype ID	Description	Total Genes with Phenotype	Set Genes with Phenotype	Fold change
1.28×10^{-3}	MP:0001408	stereotypic behavior	127	4	8.12
3.18×10^{-3}	MP:0002064	seizures	270	5	4.77
3.90×10^{-3}	MP:0004262	abnormal physical strength	172	4	5.99
4.08×10^{-3}	MP:0000951	sporadic seizures	25	2	20.62
4.64×10^{-3}	MP:0002733	abnormal thermal nociception	89	3	8.69
4.79×10^{-3}	MP:0000746	weakness	90	3	8.59
6.07×10^{-3}	MP:0001486	abnormal startle reflex	98	3	7.89
6.24×10^{-3}	MP:0001488	increased startle reflex	31	2	16.63
6.41×10^{-3}	MP:0003492	abnormal involuntary movement	456	6	3.39
7.40×10^{-3}	MP:0001961	abnormal reflex	329	5	3.92
7.91×10^{-3}	MP:0001360	abnormal social investigation	35	2	14.73

Table 3.29: Significantly enriched mouse model behavioural phenotypes among *ASD-implicated* genes - see Appendix Table 3.26.

Probability	Phenotype ID	Description	Total Genes with Phenotype	Set Genes with Phenotype	Fold change
7.47×10^{-4}	MP:0002082	postnatal lethality	1043	11	2.72

4

Functional genomics identifies convergent aetiologies for both deletions and duplications found in ASD

4.1 Abstract

Given the ever-increasing list of genes implicated in autism thus far, interpreting these genes' roles in the context of disease aetiology is being emphasised more than ever before. Employing ASD *dn* CNV genes that are associated with objective autism-relevant mouse model phenotypes (Chapter 3), I applied several other functional enrichment analyses, in order to achieve biological insights for these genes contributing to autism. The protein products of the mouse model phenotype-associated genes were found to form a large and significantly connected interacting network, which immediately allowed an identification of many additional CNV genes whose protein products directly interacted within this network. Despite the minimum genic overlap, the network is significantly connected to genes that were previously implicated in autism and to two other smaller networks recently reported in large-scale exome sequencing studies. Considering duplicated and deleted networked genes separately, I found that there were significant functional biases that distinguished duplications and deletions: disrupted genes are significantly enriched in positive regulators (40/95 genes, 2.3-fold change, corrected $p = 2 \times 10^{-5}$), whereas duplicated genes are significantly enriched in negative regulators (32/92 genes, 2.2-fold change, corrected $p = 0.005$). I interpret this functional differentiation separating duplications and deletions in the context of the identified protein interactome, illustrating several examples of protein complexes and autism-relevant signalling cascades that suggest convergent aetiology both for duplications and for deletions.

4.2 Introduction

4.2.1 Functional Enrichment Analysis (FEA)

While the conventional genetic approaches normally investigate one or a few genes at once, high-throughput techniques such as microarray and next generation sequencing allow us to examine genes under certain biological conditions at a genome-wide scale simultaneously. These high-throughput approaches often yield a long list of "interesting" genes that is typically determined using statistical methods whose suitability depends on the experimental design. The resulting list is of high importance in identifying genes that may play roles in a given condition or disease. However, a list of "interesting" genes frequently does not provide underlying biological insights, leaving the interpretation of the large set of "interesting" genes as a major further challenge for researchers.

Scientists have attempted to solve this problem by grouping a large set of genes into smaller sets of similar or related genes. Accordingly, various knowledge-based resources have been developed in order to aid the functional characterisation of genes. Such functional annotations include molecular functions, biological processes, cellular component, physical interactions between protein products, co-expression of genes, and model organism phenotypes, among others (see section 2.2.2).

These approaches to interpret gene list by investigating which genes are involved in which biological context may be collectively called *pathway analyses*. Early *pathway analyses* relied on a relatively short list of handpicked genes and relatively subjective interpretation of those genes based on a few anecdoted cases. An emerging, but already widely-used, Functional Enrichment Analysis (FEA) is a statistically rigorous analytic method that allows an interpretation of a long list of genes based on a large-scale functional knowledge-based resources. The general principle of FEA is that if a statistically significant number of genes from a known biological processes or functions are present in the gene list of interest, those over-represented biological functions may underlie the given condition or disease.

In this section, I will introduce the three main current FEA approaches (reviewed in (240)) and discuss previous FEA studies that have been conducted for autism.

4.2.1.1 Current FEA methods

Over-Representation Analysis (ORA) Over-Representation Analysis (ORA) is a FEA method that statistically assesses the proportion of a pre-defined gene set or a pathway among genes that are examined. The step-by-step procedure of ORA is as follows: First, an "interesting" gene set is determined from an experiment by setting a certain cutoff value. For instance, in gene expression microarray experiment, genes that are differentially over- or under-expressed at a false discovery rate (FDR) of 5% may be selected. Another example is to choose genes whose loci overlap with copy number variants found in patients with a certain disease, or genes within highly associated genomic intervals from genome-wide association studies (GWAS) for a given condition. Second, "interesting" genes that are a member of pre-defined gene sets or pathways are counted. The same procedure is also performed on appropriate background genes. Then, these counts from the "interesting" genes and background genes for every pathway are compared to test whether the pathway is over- or under-represented among the "interesting" genes (n), compared to expectation by random chance calculated from the background genes (N).

The most frequently used statistical methods for such a test are hypergeometric, binomial, and χ^2 tests. The exact null distribution for the probability of $\geq$x of n genes having a given annotation, given that G of a total gene population of N have that annotation, is equivalent to the hypergeometric distribution that describes sampling without replacement (also known as the one-tailed Fisher's exact test, see section 2.3). However, for $N \gg n$, the binomial distribution (sampling with replacement) is a good approximation, and hence may be used as an approximate null distribution in the case of large sample size (241). The (Pearson's) χ^2 test is similar to the Fisher's exact test, but relies on an approximation of the observed distribution referred to the theoretical χ^2 distribution, which permits much simpler calculation (241). However, the χ^2 test gives approximate probabilities that are valid only in a large sample size, thus an appropriate exact test is preferred, whenever possible.

Functional Class Scoring (FCS) The Functional Class Scoring (FCS) approach is based on the assumption that although genes with large changes may strongly

affect pathways, aggregation of weaker changes in related genes may also perturb pathways. The main steps of the FCS method is as follows: First, a gene-level statistic is computed based on measurements taken from an experiment. For example, in a gene expression array experiment, one can calculate differential expression level for individual genes. Commonly used statistics at this stage include phenotypic correlation (242), ANOVA (242), Q-statistic (243), signal-to-noise ratio (244), *t*-test (245), and Z-score (246). The choice of a gene-level statistic has been found to have negligible effect on final results (247).

Second, the computed gene-level statistics are combined for a pre-defined gene set or a pathway to produce a single pathway-level statistic. The statistics that are commonly used for this stage include the Kolmogorov-Smirnov statistic (244), sum, mean, or median of the gene-level statistic (248), the Wilcoxon rank sum (249), and the maxmean statistic (250). Regardless of the type of statistics, the power of a pathway-level statistic relies on the proportion of the given genes in a pathway, the size of the pathway, and the degree of correlation between the genes in a pathway (240).

Lastly, the pathway-level statistics are assessed for significance. The null hypotheses that are tested mainly fall into two categories: i) the competitive null hypothesis and ii) self-contained null hypothesis (251)(247)(245)(250). A competitive test compares a set of genes in a given pathway with the rest of genes from the whole genome (i.e. equivalent to the ORA methods), whereas a self-contained test neglects genes that are not in the pathway, and permutes phenotype labels of individuals among the genes annotated with the pathway. For example, to calculate the p-value for a self-contained test, all the preceding steps of FCS are repeated on randomised data where phenotype/subject labels are permuted, and the pathway-level statistic then is compared with the permuted statistics within a given pathway (see section 2.3). A self-contained test is advantageous in that it retains the correlation structure between genes as it permutes subjects rather than genes. In addition, a self-contained test has more power as the test is more restrictive (251). However, the high power of the test can be problematic if a given set of genes is very large, as almost all pathways may then be called as being significant (240).

Unlike ORA approaches, FCS does not employ an arbitrary cutoff. Instead, FCS approaches utilise continuous measurements or scores that result from an experiment, hence allowing the detection of aggregated changes of multiple genes in a coordinated manner.

Pathway Topology (PT) A growing number of knowledge-based resources are now providing information beyond a simple gene list involving pathways, such as interactions between gene products within a pathway, effects of interactions (e.g. activation and inhibition), and cellular locations where protein-protein interactions occur (252)(253)(254)(255). Pathway Topology (PT) is an emerging method that utilises this additional information to improve the functional characterisation of genes. The key difference of the PT method from other methods is that PT computes gene-level statistics based on the given genes' topology information within a given pathway. For instance, a PT-based method called ScorePAGE computes the correlation between each gene pair in a pathway (256). These gene pair-level statistics are then weighted by the number of reactions that are required to connect the gene pairs in the pathway. These weighted pairwise statistics are finally averaged to calculate pathway-level statistics. Other PT-based methods consider the dynamics of a pathway by embedding various types of pathway-related information such as the change in gene expression, effects of interactions, location of gene products, and number of interactions in a pathway (257)(258). Although PT-based approaches are the most advanced method among the current FEA approaches given its capability to account for multiple layers of information, there remains considerable room for PT-based methods to be improved in order for them to be used more broadly. One apparent pitfall of the PT-based method is that pathway topology can be substantially different within different cell types due to cell-specific gene expression level, for which information is currently largely lacking (259).

Among the three FEA approaches reviewed here, the ORA approach was the most suitable analysis method for my study considering its experimental design: As I sought to analyse rare *de novo* copy number variants found in autism patients (termed ASD *dn* CNVs), genes could only be divided into those that are overlapped by ASD *dn* CNVs and those that are not. Thus, it was impossible to compute gene-level statistics that

are required to perform FCS or PT-based methods. However, to complement the ORA approach that lacks pathway topology information, I subsequently overlaid the genes of interest with all known protein-protein interactions in order to assess their degree of connectivity.

4.2.1.2 FEA for studying ASD genetics

Considering the complexity and heterogeneity in the genetics of autism, to which the variants of several hundred genes might contribute, a statistically robust identification of autism-relevant biological processes among the long list of candidate autism-causing genes (260) is an extremely important goal. Several studies have applied FEA approaches to investigate the genetics underlying autism. These mainly differ in the types of knowledge-based resources they used. The functional resources employed to study autism have included: literature-mining (261)(262)(263), protein-protein interactions (264), mouse model phenotypes (265), gene co-expression (266)(267)(268), and functional linkage networks (269). Each of these FEA studies often reported a large and mostly novel gene dataset, thereby rapidly increasing the aggregate candidate gene list for autism (260). However, replication of these repeatedly significant signals in functional enrichments has rarely been attempted and/or reported (227)(270).

In this chapter, I describe how I applied various FEA techniques to the candidate genes determined from significant enrichments of autism-relevant mouse model phenotypes among four sets of ASD *dn* CNVs (see Chapter 3). As a result, I show that the protein products of these candidate genes form a large, highly interconnected protein-protein interaction (PPI) network. In addition, I demonstrate that deleted genes are significantly over-represented in positive regulators of biological processes, whereas duplicated genes are significantly enriched in negative regulators, which in turn suggests a convergent aetiology for both duplications and deletions in autism. In the Discussion section, I provide several examples of how the direction of copy number change of the genes within the PPI network may yield converging biological effects in the context of autism-relevant signalling cascades.

4.3 Materials and Methods

4.3.1 *De novo* CNVs identified in ASD patients

Four sets of *de novo* CNVs found in autism patients (ASD *dn* CNVs), namely, AGP, Marshall *et al.*, Levy *et al.*, and Sanders *et al.* were employed in this study. Details for each cohort are described in Chapter 3.

4.3.2 Gene Ontology (GO)

In order to investigate biological features of the candidate genes identified via mouse phenotypic enrichments (see Chapter 3), I employed the Gene Ontology (GO) database (124). The GO terms are structured under controlled vocabularies (ontologies) that describe gene products in terms of biological processes, cellular components and molecular functions in a species-independent manner (124). 14,281 GO terms (9,547 terms in biological processes; 3,569 terms in molecular functions; and 1,165 terms in cellular components) and human genes that are mapped to each GO term were extracted from the Ensembl 62 database via BioMart. Genes that belong to each GO term were then imputed to their parent terms to ensure that genes that are annotated with specific terms are annotated with every more general term. Without this imputation, genes annotated with a finer GO term will not be considered when testing for a functional enrichment of its parent term.

In order to summarise significantly enriched GO terms, I used REVIGO (REduce and VIsualise Gene Ontology) (271) with standard parameters (allowed similarity, 0.7; database with GO term sizes, whole UniProt; semantic similarity measure, SimRel), which also accounted for the p-values associated with each GO term.

4.3.3 Protein-Protein Interaction (PPI) network

I used DAPPLE (Disease Association Protein-Protein Link Evaluator) (272) to discover a protein-protein interaction (PPI) network among the examined genes. DAPPLE builds direct and indirect interaction networks from protein products of seed genes, based on the InWeb database (126)(125). InWeb consists of 169,810 high-confidence pair-wise interactions among 12,793 proteins including human proteins and their orthologues. To be confident of the quality of PPIs reported in the database, only those protein interactions that have been observed in multiple independent assays or reported

more frequently in low-throughput experiments are considered for InWeb database. Rossin *et al.* then overlaid mRNA expression data to confirm co-expression of binding partners to build reference PPI networks in DAPPLE (272). For visual presentation of the identified PPI networks, I used Cytoscape (273).

4.3.4 Statistical Analysis

To identify GO enrichments, I tested a null hypothesis that the frequency of the input genes that are associated with a GO term is no different from that expected by chance from the whole genome, from the hypergeometric distribution. A False Discovery Rate (FDR) $< 5\%$ was then applied to the nominally significantly associated GO terms to correct for multiple testing.

To test whether the number of protein-protein interactions identified among the gene products of interest are significantly more than that expected by random chance, I randomly sampled genes from the whole-genome (or 1,766 genes that are annotated with *abnormal behaviour* in Mouse Genome Informatics (MGI) database (17) where appropriate), and generated 500 randomised gene sets that match in size with a set of "interesting" genes. I then calculated permuted p-values and fold-changes by comparing the number of direct protein-protein interactions identified among the protein products of "interesting" genes with those from randomly permuted gene sets.

4.4 Results

4.4.1 Mouse functional genomics offers candidate genes for ASD

Functional Enrichment Analysis (FEA) of rare ASD *dn* CNVs using mouse orthologue-phenotype information yielded 23 significant behavioural, 21 nervous system, 27 hearing and 1 postnatal-lethality phenotypic associations with autism (see Chapter 3). Among these phenotypic enrichments, *abnormal CNS synaptic transmission* phenotype was independently associated with 3 (AGP, Marshall *et al.*, and Levy *et al.*) of 4 CNV sets, thereby triplicating the association (AGP 3.0-fold enrichment, $p = 7 \times 10^{-5}$, FDR$<$ 5%; Marshall *et al.* 2.1-fold enrichment, $p = 1 \times 10^{-4}$, FDR$<$ 5%; Levy *et al.* 2.2-fold enrichment, $p = 5 \times 10^{-5}$, FDR$<$5%; Sanders *et al.* 1.6-fold enrichment, $p = 0.008$, FDR$>$5%). Subsequently, genes directly contributing to the triplicated abnormal synaptic transmission phenotypic association serve as a confident candidate gene set for autism (termed *AST* (Abnormal Synaptic Transmission) genes, 59 genes; Appendix Table 4.4). In addition, genes that contribute to any discovered phenotypic association and those annotated with any of these associations from other ASD *dn* CNV cohorts form a larger set of candidates for autism (termed *AAP* (All Associated Phenotypes') genes, 185 genes; Appendix Table 4.4).

4.4.2 ASD candidate genes identified via functional mouse genomics form a significantly interconnected PPI network

Given that many separate ASD *dn* CNV genes contribute to significant enrichments of certain mouse phenotypes (Chapter 3), I hypothesised that the protein products of these associated ASD *dn* CNV genes (i.e. 59 *AST* genes and 185 *AAP* genes) would highly interact each other, functioning in common biological processes or pathways. Accordingly, I tested whether the number of protein interactions identified among these genes is higher than that expected by chance. In order to account for the increased likelihood for the protein products of behavioural genes to be interacting with each other, compared to the protein products of randomly selected genes, I calculated the expected number of protein interactions and p-values by generating 500 randomised gene sets of matched size sampled from 1,766 behavioural genes (see Materials and Methods). As a result, I found that the number of interactions among the protein products of seed genes was significantly increased for both *AST* genes and *AAP* genes (*AST* genes - 22

direct PPIs, permuted $p = 0.006$, fold-change 3.8; *APP* genes - 90 direct PPIs, permuted $p = 0.03$, fold-change 1.7; Table 4.1). 34 genes that were previously implicated in autism (*ASD-implicated* genes) also form a significantly connected PPI network (6 direct PPIs, permuted $p = 0.04$, fold-change 3; Table 4.1). Furthermore, when the *ASD-implicated* genes were added to each candidate set, the significance of the PPI network connectivity was enhanced (*AST* + *ASD-implicated* genes, 43 direct PPIs, permuted $p < 0.002$, fold-change 3.2; *AAP* + *ASD-implicated* genes, 141 direct PPIs, permuted $p < 0.002$, fold-change 2; Table 4.1; Figure 4.1), suggesting that many more interactions occur between protein products of *AST* genes or *APP* genes and proteins of known ASD causative genes.

Gene set	**Background**	**#Genes**	**#Direct PPIs**	**#Expected PPIs**	**P_{perm}**	**Fold change**
AST	MGI Beh	59	22	5.87	0.006	3.75
AAP	MGI Beh	184	90	53.25	0.03	1.69
ASD-implicated	MGI Beh	34	6	2.03	0.04	2.96
AST+ASD-implicated	MGI Beh	90	43	13.31	<0.002	3.23
AAP+ASD-implicated	MGI Beh	215	141	69.8	<0.002	2.02

Table 4.1: PPIs among protein products of *AST* and *AAP* genes - *AST*, ASD *dn* CNV genes that are associated with triplicated *abnormal synaptic transmission* phenotypes; *AAP*, ASD *dn* CNV genes that have mouse phenotypic associations; *ASD-implicated*, genes that have been previously implicated in autism; Background (MGI Beh), human genes whose mouse orthologues are annotated with abnormal behaviour in the Mouse Genome Informatics (MGI) database; # Genes, the number of genes that have PPI information in each set; # Direct PPIs, the number of protein-protein interactions that result from the direct binding of protein products of a gene set; # Expected PPIs, the number of protein-protein interactions that are expected by chance; P_{perm}, p-values obtained from 500 randomisations; Fold change, fold-change of the number of direct PPIs compared to those expected by random chance.

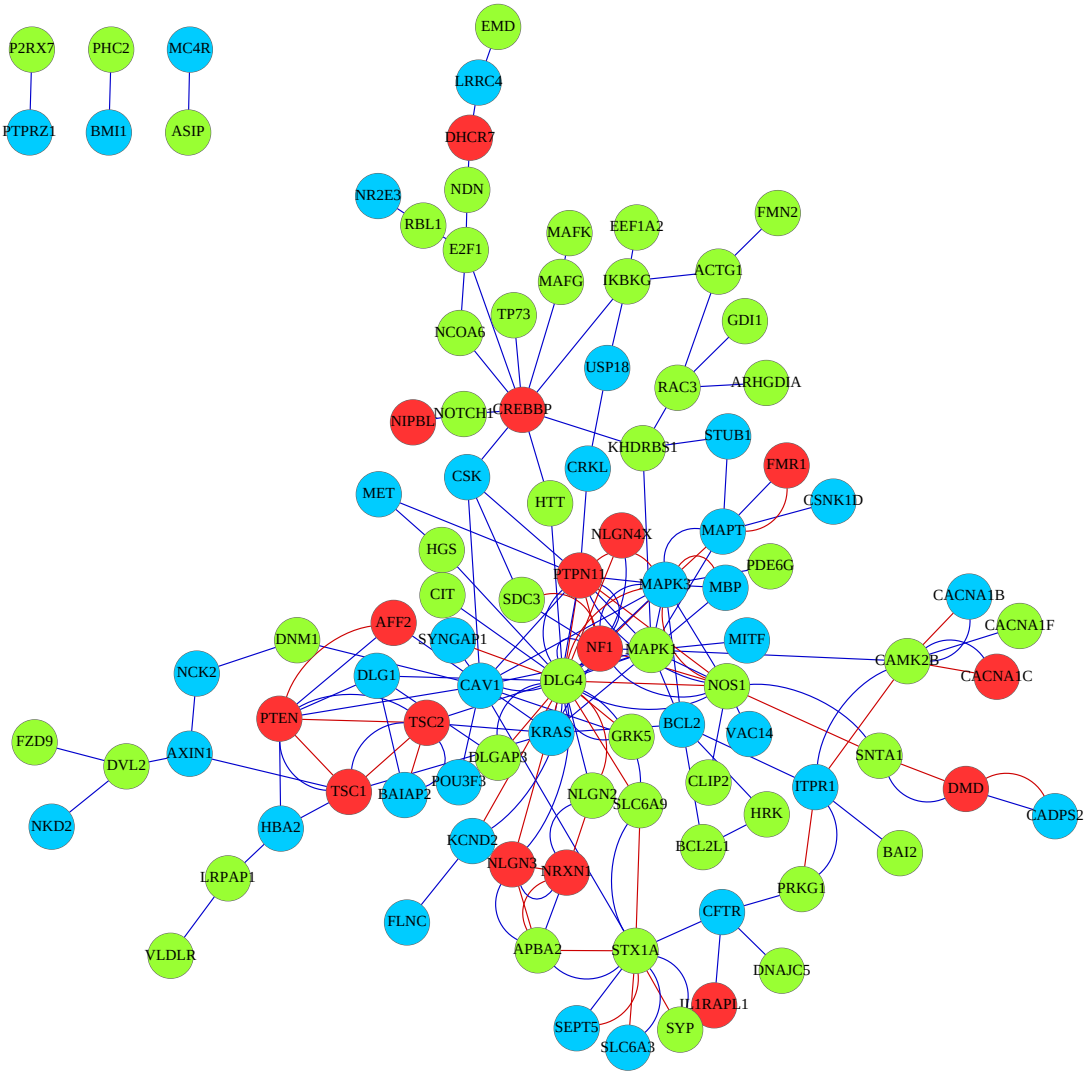


Figure 4.1: Protein-protein interaction network among *AST* and *AAP* genes - green nodes, protein products of genes from duplication CNVs; blue nodes, protein products of genes from deletion CNVs; red nodes, protein products of *ASD-implicated* genes; red edges, protein-protein interactions found for *AST* genes; blue edges, protein-protein interactions found only for *AAP* genes. See Table 4.1 for description of gene sets.

4.4.3 PPI analysis offers additional candidate genes for ASD

As mouse model phenotypic information is currently available only for the orthologues of less than a quarter of human genes (see Chapter 3), an FEA approach employing mouse phenotypic resources alone is likely to fail to identify the majority of genes within ASD *dn* CNVs that are causally associated with autism-relevant phenotypes. In order to overcome this limitation, I considered those candidate genes identified through mouse model phenotypes (*AST* or *AAP* genes) combined with the set of *ASD-implicated* as seed genes, and asked whether the protein products of these seed genes interact with the protein products of additional ASD *dn* CNV genes that, to date, are not annotated with these phenotypes.

Indeed, I identified a number of additional ASD *dn* CNV genes whose protein products directly bind to the proteins encoded by genes that were identified via mouse phenotype FEA: I identified 100 additional ASD *dn* CNV genes where proteins have direct PPIs with *AST* gene proteins, and 143 additional ASD *dn* CNV genes when *AST* and *ASD-implicated* genes were considered together. For the *AAP* gene set, PPI analysis identified an additional 218 ASD *dn* CNV genes; the analysis highlighted 231 further genes when the *AAP* and *ASD-implicated* genes were considered together. These expanded gene sets from the four original seed gene sets (i.e. *AST* & *AAP* gene sets, and each set combined with *ASD-implicated* genes) form highly connected PPI networks, yielding a large protein-protein interaction hub for autism (Table 4.2; Figure 4.2).

Gene set	Background	#Seed	#Add.Gene	#Direct PPIs	P_{perm}	Fold change
AST+dirPPI	MGI Beh	59	100	251	<0.002	6.28
AST+dirPPI*	MGI Beh	90	143	452	<0.002	5.4
AAP+dirPPI	MGI Beh	184	218	815	<0.002	3.59
AAP+dirPPI*	MGI Beh	215	231	944	<0.002	3.51

Table 4.2: PPIs among proteins of *AST* and *AAP* genes and their directly interacting partners encoded by ASD *dn* CNV genes - *, Additional *ASD-implicated* genes included; *dirPPI*, additional ASD *dn* CNV genes whose protein products directly interact with proteins of original candidate sets (*AST, AST*, AAP* or *AAP**) included; #Seed, the number of genes in original candidate sets; #Add.Gene, the number of *dirPPI* genes. See Table 4.1 for other column descriptions.

Surveying the functional GO annotations of these *AST-associated network* genes (i.e. *AST+ASD-implicated+dirPPI* genes) and *AAP-associated network* genes (i.e.

AAP+ASD-implicated+dirPPI genes) revealed significant (FDR<5%) enrichments of >550 GO terms in each network (*AST-associated network*, 565 GO terms including 427 biological processes, 73 cellular component, and 65 molecular function, Appendix Table 4.5; *AAP-associated network*, 579 GO terms including 479 biological processes, 68 cellular component, and 32 molecular function, Appendix Table 4.8). In order to summarise this long list of GO annotations, I used REVIGO (REduce and VIsualise Gene Ontology) (271). As expected, many of the summarised GO terms were shared between the *AST-associated* genes and the *AAP-associated* genes and had converging functional themes between these candidate genes. The converging GO biological processes in both gene sets include: regulation of cellular processes, cell-cell signalling, secretion, developmental process, behaviour, and cytoskeleton organisation (Appendix Figures 4.8 and 4.11). The network genes were also significantly enriched in GO cellular component including vesicle, ion channel complex, neuron projection, and synapse (Appendix Figures 4.9 and 4.12). The significantly enriched GO molecular functions include ion channel (regulator) activity, transporter activity, and GABA receptor activity (Appendix Figures 4.10 and 4.13). This functional convergence is consistent with the observation that 424/446 [95%] genes within the *AAP-associated* protein interaction network yield a single interconnected cluster (Figure 4.2). The large single interconnected cluster provides a mechanism through which cellular perturbations driven by separate mutations in different genes produce effects with aetiological convergence.

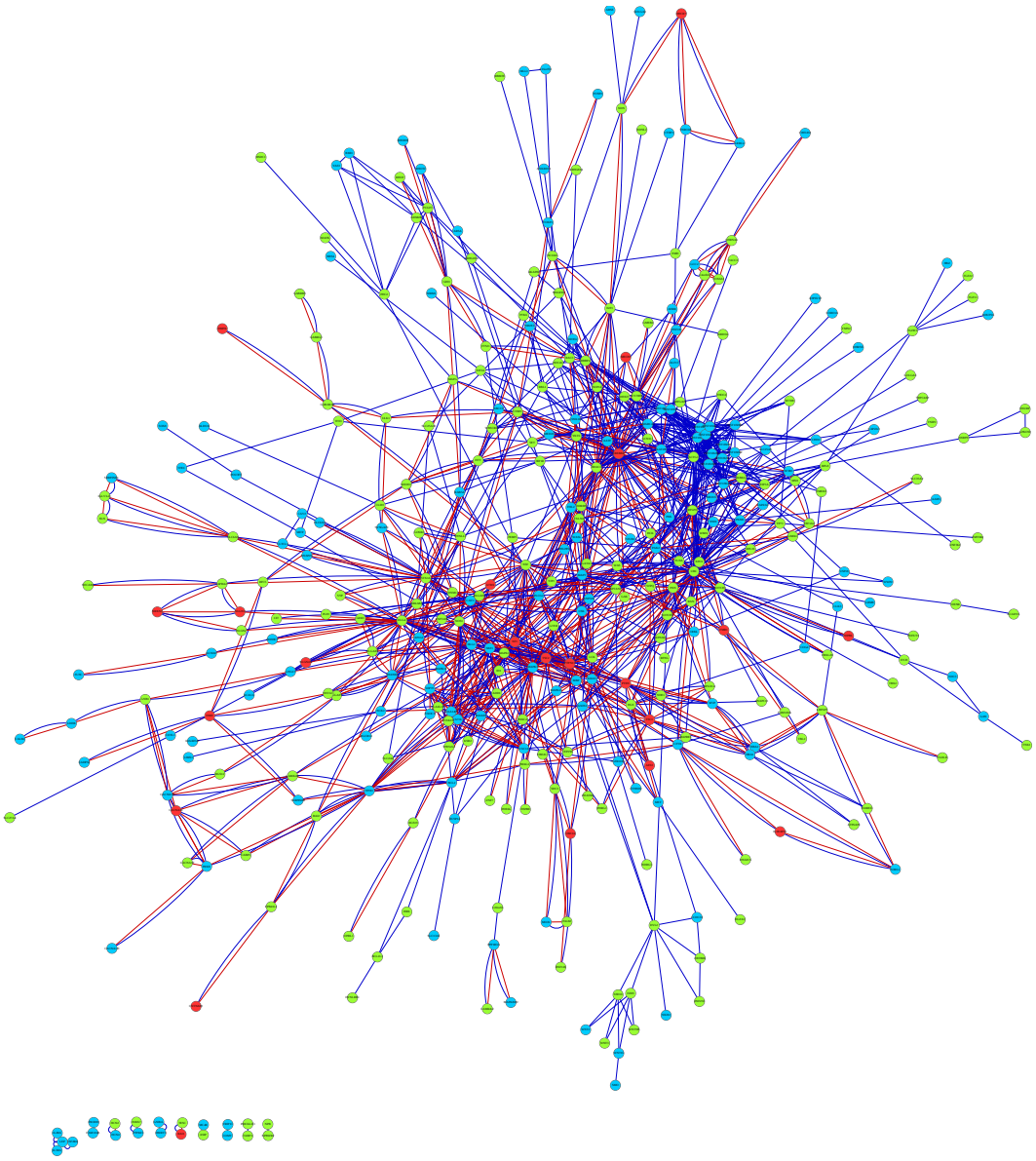


Figure 4.2: An extensive *AAP-associated* protein-protein interaction network - green nodes, protein products of genes from duplication CNVs; blue nodes, those of genes from deletion CNVs; red nodes, protein products of *ASD-implicated* genes; red edges, protein-protein interactions found in *AST-associated* network; blue edges, protein-protein interactions found only in the *AAP-associated* network.

4.4.4 Separate functional biases for deletions and duplications suggest convergent aetiologies for both directions of copy number change

In order to investigate how the direction of copy number change (duplication/deletion) might play a role in autism aetiology, I first separated these *AST-associated* and *AAP-associated* network genes according to the direction of copy number change of their CNVs of origin, and then analysed functional biases of the GO annotations assigned to each group that might reveal distinct effects arising from duplications or, separately, from deletions. While a large proportion of these GO functional biases are shared between duplicated and deleted genes in each network (*AST-associated* genes, Appendix Tables 4.6 and 4.7; *AAP-associated* genes, Appendix Tables 4.9 and 4.10), I noted a remarkable difference between the two groups. Deleted genes were significantly over-represented in genes that are positive regulators of biological processes (GO:0048518), whereas the duplicated genes were significantly over-represented in genes that are negative regulators of biological processes (GO:0048519) (Table 4.3).

Gene set	CN direction	GO terms	#Genes	P	Fold	FDR<5%
AST	Deletion	+ive regulation	10/27	1.51×10^{-2}	2.06	-
AST+dirPPI	Deletion	+ive regulation	25/64	5.98×10^{-5}	2.17	**
AST-associated	Deletion	+ive regulation	35/82	1.75×10^{-7}	2.37	**
AST	Duplication	-	-	-	-	-
AST+dirPPI	Duplication	-ive regulation	24/83	1.90×10^{-3}	1.82	-
AST-associated	Duplication	-ive regulation	34/105	1.98×10^{-5}	2.04	**
AAP	Deletion	+ive regulation	37/87	8.78×10^{-8}	2.36	**
AAP+dirPPI	Deletion	+ive regulation	59/170	1.31×10^{-7}	1.93	**
AAP-associated	Deletion	+ive regulation	61/174	5.30×10^{-8}	1.95	**
AAP	Duplication	-ive regulation	34/91	5.14×10^{-7}	2.35	**
AAP+dirPPI	Duplication	-ive regulation	62/202	8.78×10^{-8}	1.93	**
AAP-associated	Duplication	-ive regulation	64/211	8.92×10^{-8}	1.91	**

Table 4.3: Enrichment of GO regulation of biological processes among *AST-associated* and *AAP-associated* genes - dirPPI, ASD *dn* CNV genes that encode direct protein-protein interaction partners of AST-/AAP-gene products; *AST-associated, AST+ASD-implicated+dirPPI*; *AAP-associated, AAP+ASD-implicated+dirPPI*; CN direction, copy number direction; +ive regulation, GO:0048518; -ive regulation, GO:0048519; #Genes, [# genes contributing to the enrichment]/[# total genes in set]; P, probability; Fold, fold-change over expected by random chance; FDR, false discovery rate; **, passed FDR<5% cutoff.

Among the *AST-associated* network genes arising from deletion ASD *dn* CNVs, the *AST* genes alone were nominally significantly enriched in the GO term *positive*

regulation of biological processes ($p = 0.0015$, 2.06 fold; Table 4.3). When their protein products' direct interaction partners [within deletion ASD *dn* CNVs] were added, the signal became highly significant ($p = 5.98 \times 10^{-5}$, 2.17 fold change, FDR<5%; Table 4.3). This significance was further strengthened by adding *ASD-implicated* genes and their protein products' interacting partners [arising from deletion ASD *dn* CNVs] ($p = 1.75 \times 10^{-7}$, 2.37 fold change; Table 4.3). Among the *AST-associated* network genes arising from duplication ASD *dn* CNVs, the *AST* genes alone did not yield an enrichment of the GO term *negative regulation of biological processes* (Table 4.3). However, when their protein products' direct interaction partners [within duplication ASD *dn* CNVs] were added, a nominally significant enrichment of the GO term was detected ($p = 1.90 \times 10^{-3}$, 1.82 fold change; Table 4.3). In addition, the signal became highly significant once *ASD-implicated* genes and their protein products' interacting partners [arising from duplication ASD *dn* CNVs] were added ($p = 1.98 \times 10^{-5}$, 2.04 fold change; Table 4.3).

The same analysis on the gene sets from the larger *AAP-associated* network resulted in much stronger signals. The *AAP-associated* genes arising from deletion ASD *dn* CNVs were significantly enriched in the GO term *positive regulation of biological processes* ($p = 5.30 \times 10^{-8} - 1.31 \times 10^{-7}$, 1.93-2.36 fold change; FDR<5%; Table 4.3), while the *AAP-associated* genes arising from duplication ASD *dn* CNVs were significantly enriched in the GO term *negative regulation of biological processes* ($p = 8.78 \times 10^{-8} - 5.14 \times 10^{-7}$, 1.91-2.35 fold; Table 4.3).

These converse enrichments for duplications and deletions instantly suggest a convergent outcome for these copy number altered genes, in which the deletion of positive regulators or the duplication of negative regulators each have a similar effect on down-regulating the same target processes. Given this mechanistic model enriching gene duplications and deletions, those candidate genes that partially overlap duplication CNVs may be mechanistically equivalent to deletions in terms of their disruption of gene function. Convergently, a functional resemblance between partial duplications and deletions is to be expected. Indeed, when I considered those genes that are wholly overlapped by duplication CNVs only, the enrichment of duplicated genes among negative regulators became stronger for both *AST-associated* and *AAP-associated* genes (*AST-associated*, $p = 6.54 \times 10^{-6}$, 2.19 fold change; *AAP-associated* genes, $p = 3.4 \times 10^{-8}$,

1.99 fold change). Furthermore, these enrichments were accompanied by improved enrichments of positive regulators among the deleted genes with the partially duplicated genes (*AST-associated* genes, $p = 3.73 \times 10^{-8}$, 2.34 fold change; *AAP-associated* genes, $p = 4.61 \times 10^{-8}$, 1.9 fold change). The unexpectedly frequent deletion of positive regulators and duplication of negative regulators allow biologically-meaningful interpretations of the large protein interaction network reported here, through which most of the regulators were identified (see Discussion).

4.5 Discussion

Employing the genes identified from *(i)* the enrichment of genes whose mouse orthologues, when disrupted, yield an *abnormal synaptic transmission* phenotype among 3 independent sets of ASD *dn* CNVs, (and thus triplicated,) and *(ii)* the enrichment of genes that yield objective autism-relevant mouse model phenotypes (see Chapter 3), I have shown that *(i)* the protein products of these genes form a significantly highly connected protein interaction cluster among themselves, as well as with the products of genes that were previously implicated in autism, *(ii)* this protein interactome immediately extends the candidate genes, by identifying their direct binding partners as also being affected by ASD *dn* CNVs, *(iii)* the genes within this network have roles in synaptic function, organisation, amd maintenance, and *(iv)* among the network genes, those that are deleted frequently act to positively regulate biological processes whereas duplicated genes often negatively regulate biological processes.

4.5.1 *AST-associated* and *AAP-associated* genes offer an extensive interaction hub for ASD

The candidate *AST* genes and *AAP* genes formed significantly interconnected protein-protein interaction (PPI) networks ($p = 0.03 - 0.006$, fold-change 1.7-3.8; Table 4.1). Furthermore, when adding the genes that were previously implicated in autism (*ASD-implicated*), the significance of the connectivity was greatly enhanced ($p < 0.002$, fold-change 2.0-3.2; Table 4.1; Figure 4.1), demonstrating that the candidate genes and the *ASD-implicated* genes are highly connected within the identified protein interaction network. Finally, the ASD *dn* CNV genes whose protein products directly interact with the proteins encoded by the *AST*, *AAP*, or *ASD-implicated* genes formed large protein-protein interaction networks, in which most proteins reside within a single interconnected cluster, thereby further improving the connectivity of the protein-protein interaction networks ($p < 0.002$, 3.5-6.3 fold-change; Table 4.2; Figure 4.2).

I next examined whether the significant connectivity of the PPI networks identified from these candidate genes is a trivial outcome of being ASD *dn* CNV genes that are annotated with the MGI phenotype category *abnormal behavior* information (termed ASD *dn* CNV behavioural genes) or whether it does reflect a truly significant PPI connectivity. I sought to confirm that the PPIs among the protein products of the ASD

dn CNV behavioural genes do not yield a significantly connected PPI network. Indeed, when comparing the number of PPIs identified from the ASD *dn* CNV behavioural genes with all genes that have an MGI *abnormal behavior* phenotype category, the ASD *dn* CNV behavioural genes did not form a significantly connected PPI network (AGP, $p = 0.35$, 1.16 fold-change; Marshall *et al.*, $p = 0.59$, 0.94 fold-change; Levy *et al.*, $p = 0.58$, 0.94 fold-change; Sanders *et al.*, $p = 0.64$, 0.91 fold-change; 4 sets combined, $p = 0.26$, 1.1 fold-change), confirming that the *AST-associated* and *AAP-associated* networks are likely to reflect functional PPI network for autism.

Also, given that a recent study that constructed a protein interactome for autism using a yeast two-hybrid screen of a human complementary DNA library did not identify new protein-protein interactions (264), the *AST-associated* and *AAP-associated* networks, identified via known protein-protein interactions here, might be offering a protein interactome for autism that approaches being comprehensive.

4.5.2 Functions of identified candidate genes for ASD

Synaptic phenotypes in mice have previously been associated with CNVs that were identified among autism patients, but replication was not achieved (227). Here, despite the little genomic overlap among the three independent CNV sets employed (see Materials and Methods of Chapter 3), triplication of this association was achieved. These synaptic associations provide aetiological insights that agree with an emerging role for synaptic dysfunction in autism (226). Synaptic functions are further supported by the over-representation of GO functional annotations such as synapse, cell-cell signalling, secretion, vesicle transport, and ion channel complex (Appendix Tables 4.5 and 4.8 and Appendix Figures 4.8-4.13) among candidate genes identified via the protein interactome. In particular, many of these significantly over-represented GO functions are related to cellular processes such as cytoskeletal organisation and cellular developmental processes, thus directing the scope of synaptic dysfunctions in autism towards intracellular processes such as abnormalities in synaptic formation, structure and/or maintenance (Appendix Tables 4.5 and 4.8 and Appendix Figures 4.8-4.13) rather than extracellular signalling molecules. Indeed, this idea of a more intracellular role for these autism candidate genes' proteins is further corroborated by an observation that there is a significant deficit in signal peptides encoded by these genes (38% vs. 21% of all behavioural genes, $p = 0.03$).

4.5.3 *AST* genes yield behavioural phenotypes.

Almost all genes associated with an abnormal synaptic phenotype (85%; 50/59 *AST* genes) also yield behavioural phenotypes, thereby providing a direct link between *AST* genes and behavioural genes. Incomplete annotation or phenotyping of mouse models may explain those that do not (*CACNA1H, CHRNB1, CPLX3, FBXL20, GNB1L, P2RX4, PRKG1, SYP*, and *TBX1*). This strong overlap between behavioural genes and *AST* genes suggests that the identified synaptic genes (*AST* genes) and the identified behavioural genes (*AAP* genes) are aetiologically little different, rather, the disruption of these genes lead to the manifestation of autistic behaviours by affecting synaptic transmission.

4.5.4 Identified interactome is highly connected to two recently reported autism-networks

Recently, O'Roak *et al.* and Neale *et al.* published two large-scale exome-sequencing studies of autism patients that also identified excessive protein-protein interactions among genes possessing potential causal mutations (274)(275). Each study reported networks of 49 and 45 contributing genes, of which 3 genes and 2 genes, respectively, were also present in the interaction network identified here. Given the almost perfect exclusivity between published networks and the interactome identified here, I tested whether these networks interact with each other. After excluding common genes between the sets, I tested whether the number of connections between genes in each of the O'Roak *et al.* and Neale *et al.* networks to the genes in the network reported here is significantly greater than expected by random gene sets of equivalent size. Indeed, O'Roak *et al.* and Neale *et al.* networks each showed significantly enhanced connections to my network by 12-fold ($p < 0.002$) and 38-fold ($p < 0.002$), respectively (Appendix Table 4.11). This significant interconnectedness between these independently identified networks implies aetiological convergence.

4.5.5 Convergent aetiologies between deletions and duplications in ASD

A notable functional distinction between candidate genes derived from duplicated CNVs and those from deleted CNVs is that duplicated genes are enriched in *negative regu-*

lators of biological processes while deleted genes are enriched in *positive regulators of biological processes* (Table 4.3). This observation immediately suggests a model of convergent aetiologies, in which both deletions of positive regulators and duplications of negative regulators act to down-regulate autism-relevant biological processes. This concept could be proven by an experiment demonstrating that over-expression of duplicated genes in autism leads to the same effect as deletion of other candidate genes found in deleted CNVs. In the most straightforward case, over-expression of a gene (derived from duplicated CNVs) would negatively regulate a gene (from deleted CNVs) in autism. While the differential enrichments do not support a strict rule, there are a number of network interactions where such regulatory examples have already been proven experimentally and are described in the available scientific literature. In the following sections, I will discuss several examples of such network interactions and the roles of the participating genes in the context of brain function and autism.

4.5.5.1 Altered presynaptic exocytosis

Among our autism candidate genes, a protein encoded by *STX1A*, a candidate gene derived from duplication, interacts with four other proteins that are encoded by candidate genes derived from deletion (i.e. *CAV1, SLC6A3, SEPT5*, and *CFTR*). *STX1A* is implicated in the docking of synaptic vesicles with the presynaptic plasma membrane and its over-expression inhibits synaptic vesicle exocytosis (276). Copy number change of other candidate genes or *ASD-implicated* genes (i.e. *SYP* [duplication], *SNAP29* [deletion], *CACNA1B* [deletion], *CACNA1C* [*ASD-implicated*], and CACNA1H [deletion]) also supports a disruption of synaptic exocytosis in ASD (see below, Figure 4.3).

STX1A - CFTR / SLC6A3 Over-expression of STX1A (syntaxin-1A) decreases SLC6A3's dopamine transport activity (277), and reduces CFTR chloride current by inhibiting trafficking of CFTR to the surface (278). These findings suggest that the outcome of over-expression of STX1A would be similar to deletions of *SLC6A3* or of *CFTR*. This is consistent with my identification of the *STX1A* gene from a duplication CNV, while *CFTR* and *SLC6A3* from deletion CNVs. In addition, elevated expression of STX1A has been observed in high functioning ($IQ > 70$) autism (279), suggesting that this is an autism-specific mechanism, rather than for intellectual disability.

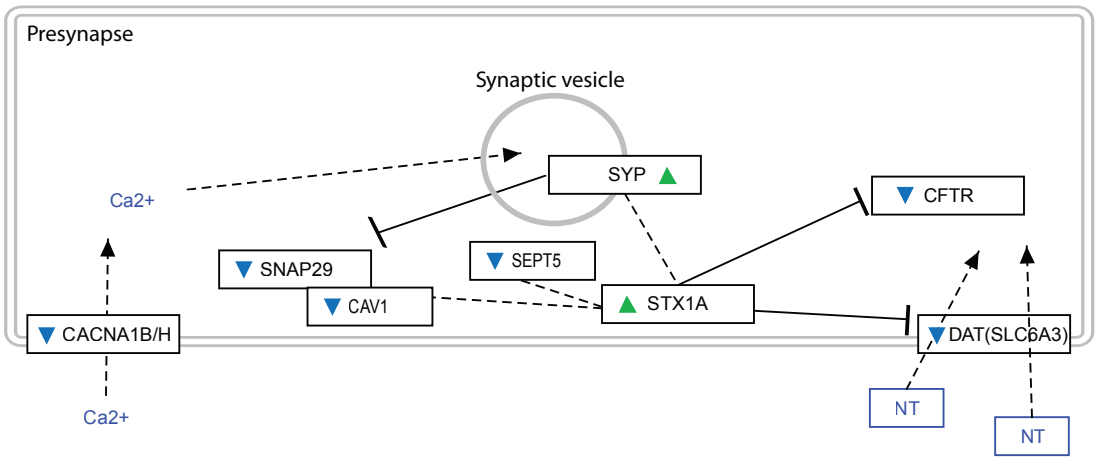


Figure 4.3: Copy number changes in candidate genes including *STX1A* and its protein interacting partners implicate presynaptic exocytosis in ASD - Green upwards arrows indicate duplicated candidate genes and blue downwards arrows indicate deleted candidate genes. NT represents neurotransmitters. Interactions between the proteins are shown in edges. For detailed edge and arrow type information, see the in-Figure legend to Figure 4.7.

STX1A - SEPT5 Due to high redundancy of Septins' function, the individual function of *SEPT5* (Septin 5) has remained elusive. However, it has been shown that *SEPT5* is predominantly expressed in neurons, localizes in presynaptic regions, interacts with STX1A, and regulates neurotransmitter release, thereby inhibiting exocytosis (280)(281). It has been suggested that *SEPT5* regulates the availability of soluble N-ethylmaleimide-sensitive factor attachment protein receptor (SNARE) proteins after exocytosis for observation that SEPT5 binds STX1A in the SNARE complex but not in the α-SNAP-SNARE complex (281). This observation implies that SEPT5 competes with α-SNAP (soluble NSF attachment protein) to bind to the SNARE complex (281). Considering these findings, it may be the case that both over-expression of *STX1A* and under-expression of *SEPT5* lead to abnormally excessive amount of α-SNAP-SNARE complex which plays an important role in the stimulation of exocytosis (282).

SNAP29 - CAV1 *SNAP29*, a member of *SNAP25* gene family, encodes a protein involved in multiple membrane trafficking steps (283). *CAV1* (caveolin 1) encodes the scaffolding protein that associates with lipid rafts and forms oligomers that leads to the local change in morphology of the membrane (284). DAPPLE has

identified that CAV1 directly associates with SNAP29 and STX1A. Also, it has been shown that SNAP proteins and CAV1 are present in neurons and colocalise in axonal varicosities (285)(286). Specifically, SNAP29 was shown to inhibit disassembly of the SNARE complex, whose disassembly is required after synaptic vesicle fusion to complete exocytosis (287). Hemizygous deletion of both *SNAP29* and *CAV1*, thus, might lead to under-expression of both proteins, thereby disinhibiting the SNARE complex disassembly step during synaptic exocytosis.

***STX1A - SYP* / voltage-dependent calcium channels** In addition, STX1A interacts with SYP, which negatively regulates SNAP proteins (288)(289). Therefore, again, over-expression of SYP (duplicated) and under-expression of SNAP29 (deletion) may have a similar effect in synaptic exocytosis. Together with an altered calcium entry via disrupted voltage-dependent calcium channels (*CACNA1B* [deletion], *CACNA1C* [*ASD-implicated*], and *CACNA1H* [deletion]) that affect the formation of synaptic vesicle (290), copy number changes of the candidate genes whose proteins locate at the presynaptic terminal may alter synaptic exocytosis processes in autism (Figure 4.3).

4.5.5.2 Reduced activity of MAPK pathway

Another example that supports the convergent aetiologies of duplications and deletions in autism is the interactions between *MAPK3* (deletion) and four other genes from duplicated CNVs (i.e. *NOS1, DLG4*, and *PDE6G*). MAPK3 is a member of the Mitogen-activated protein (MAP) kinase family, which acts in a MAPK signalling cascade that regulates various cellular processes including proliferation, differentiation, and cell cycle progression in response to a variety of extracellular signals. In addition to the direct binding partners of MAPK3, copy number changes of the genes that are involved in a MAPK3 signalling suggest that CREB-dependent transcription may be down-regulated in ASD (see below, Figure 4.4).

MAPK3 - NOS1 Due to functional compensation between *MAPK3* and *MAPK1*, many experiments were performed to measure *MAPK3* and *MAPK1* activities jointly. Hence it is difficult to determine whether a certain protein exclusively plays a role in regulating MAPK3 function. Nevertheless, it has been shown that

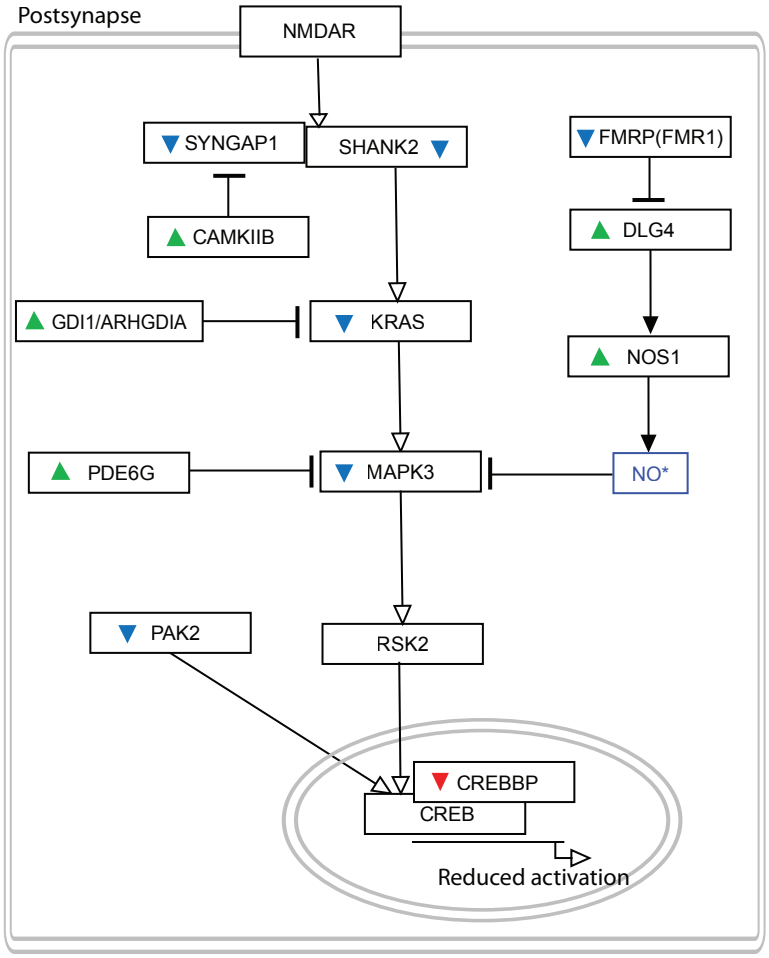


Figure 4.4: Copy number changes of candidate genes within the MAPK pathway implicate a reduced activation of MAPK-dependent transcription in ASD - *CREBBP* is marked with a red downwards arrow which indicates *ASD-implicated* genes that were known to be disrupted in autism. See Figure 4.3 and 4.7.

calcium ionophore stimulation of NOS1 leads to the inhibition of MAPK3/1 activity via NO* production (291). This indicates that excessive NO* produced by over-expression of NOS1 may reduce MAPK3 activity, which would be a comparable scenario to when *NOS1* gene is duplicated and *MAPK3* is deleted.

MAPK3 - DLG4 There has been no evidence showing that deletion of *MAPK3* directly affects the function of *DLG4*. However, over-expressing *DLG4* increases the amount of NOS1 at postsynaptic densities, recruiting NOS1 to its required locale (292). As DLG4 promotes the recruitment of NOS1, which in turn leads to the inhibition of MAPK3 activity, duplication of the *DLG4* gene may negatively regulate the function of MAPK3 via its interaction with NOS1.

MAPK3 - PDE6G Over-expression of a short cone *PDE6G* isoform was shown to reduce the EGF-dependent activation of MAPK3/1 (293).

Deletions of other candidate genes that contribute to the MAPK signalling pathway (*SYNGAP1, SHANK2, KRAS, PAK2* and *CREBBP* [*ASD-implicated*]) and duplications of their negative regulators (*FMR1, GDI1, ARHGDIA, CAMK2B*, and *CAMKK2*) also support convergent effects on the MAPK pathway, namely reduction of CREB-dependent transcription (294)(295)(296) (Figure 4.4).

4.5.5.3 Altered calcium homeostatis

Among the identified network genes, ITPR1 that was detected from ASD deletions interacts with other proteins from duplicated CNV genes including *PRKG1*, and *CAMK2B*. IP3 receptor type 1, ITPR1, acts as a calcium channel that is activated by IP3, and leads to the release of calcium from intracellular store sites (Figure 4.5). It has been shown that deletion of *ITPR1* underlies ataxia in mice and spinocerebellar ataxia type 15 in humans (297).

ITPR1 - CAMK2B Calcium/calmodulin-dependent protein kinase type II β chain (CAMK2B) is one of four different chains (α, β, γ, and δ) of the calcium/calmodulin dependent protein kinase that belongs to the serine/threonine protein kinase family. IP3 kinase A (that converts IP3 to IP4) is regulated by calcium both via direct interaction with calmodulin and phosphorylation by CaMK2 (298)(299)(300). Activation of IP3 kinase A by CAMK2 implies that more CAMK2 might enhance

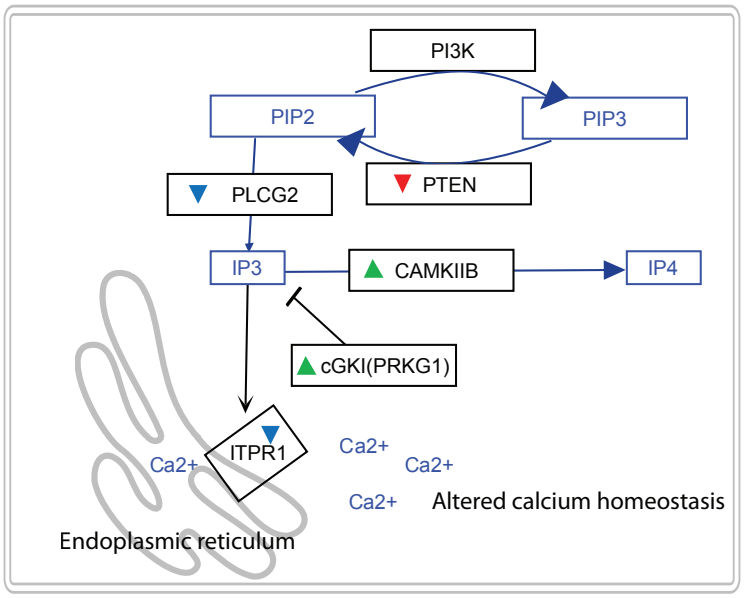


Figure 4.5: Copy number changes of candidate genes related to IP3 production implicates altered calcium homeostasis in ASD - Blue arrows indicate conversion of molecules. See Figure 4.7 for full legend.

the conversion from IP3 to IP4, resulting in less IP3 to act on ITPR1. Lacking ligands (IP3) to act on receptors (ITPR1) may perhaps yield a similar outcome to the lack of ITPR1. Furthermore, IP3 kinase A and CAMK2B co-localise in synaptically mature hippocampal cultures, associating with F-actin in dendritic spines (301), where calcium stores play key roles in modulating synaptic plasticity (302)(303).

ITPR1 - PRKG1 Similar to CAMK2B, activation of PRKG1 (cGMP-dependent protein kinase 1, α isozyme) suppresses calcium transient by inhibition of IP3 production (304). Accordingly, over-expression of PRKG1 would lead to less IP3 to act on ITPR1, which may produce a similar effect to the deletion of *ITPR1*.

In addition, copy number changes of other candidate genes also support altered calcium homeostasis in ASD. Deletion of *PLCG2*, which converts PIP2 to IP3 and DAG, leading to less cytosolic IP3, would exert a similar outcome to the duplication of *CAMK2B* or of *PRKG1*. Also, disruption of *PTEN* (a *ASD-implicated* gene), which converts PIP3 to PIP2 that eventually breaks down to IP3, might decrease the level of IP3, by producing

less of its precursor, PIP2 (Figure 4.5).

4.5.5.4 Up-regulation of *Wnt* signalling

Wnt signalling is a major player during development and its misregulation often leads to disease (305). Recent studies have predicted its role for autism (306) given the involvement of *Wnt* signalling in the formation of neural circuits (307) and regulation of dendrite morphogenesis (308)(309). Signals from both its canonical (DVL, AXIN1, β-catenin) and noncanonical signalling pathway (DVL1/2/3, Rac1, and JNK) converge and regulate actin morphogenesis. Within this well-known *Wnt* signalling pathway, *DVL2*, one of the candidate genes detected from duplications, interacts with other candidates from deletions, *AXIN1* and *NKD2*. Extensive literature has suggested that *DVL2* is a positive regulator of *Wnt* signalling, while *AXIN1* and *NKD2* are negative regulators of the same pathway (310) (see below, Figure 4.6).

DVL2 - AXIN1 DVL2 is a key component of the *Wnt* receptor complex, which inhibits a second complex of proteins including AXIN, GSK-3, and APC, when activated by *Wnt*. The AXIN/GSK-3/APC complex promotes the proteolytic degradation of β-catenin, an intracellular signalling molecule, hence is called the "β-catenin destruction complex" (311)(312). After this is inhibited, β-catenins can enter the nucleus and interact with transcription factors to promote expression of specific genes. As a result, DVL positively regulates Wnt signaling and gene expression by "dis-inhibiting" the inhibitory AXIN/GSK-3/APC complex (313). Thus, deletion of AXIN1 may enhance *Wnt* signalling, which would be a comparable outcome to the over-expression of *DVL2* by duplication.

DVL2 - NKD2 It has been shown that NKD1 and NKD2 are negative regulators of *Wnt* signalling during development. Current evidence suggests that NKDs degrade DVL proteins to antagonize the *Wnt* signalling (312)(314). Thus, a lack or reduction of NKD2 is likely to lead to more DVL protein, which would produce a similar effect to the duplication of *DVL* genes that, in turn, may up-regulate *Wnt* signalling.

Copy number changes of other proteins also support the up-regulation of *Wnt*-stimulated TCF/LEF-dependent transcription. Ubiquitination of AXIN1 and β-catenin

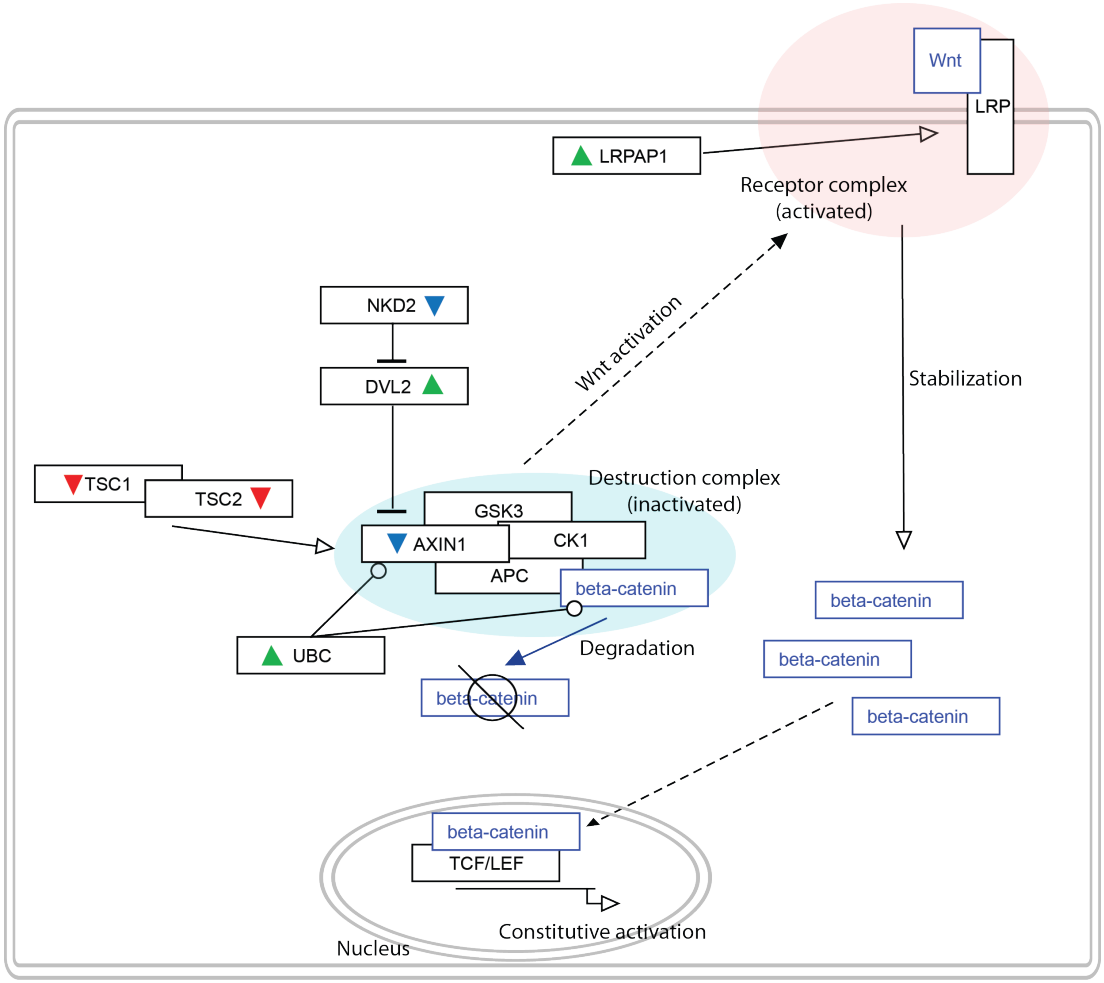


Figure 4.6: Copy number changes of candidate genes within *Wnt* signalling pathway implicates its constitutive activation in ASD - See Figure 4.7 for full legend.

by duplicated UBC (ubiquitin C) might enhance the degradation of the "β-catenin destruction complex" in the *Wnt*/β-catenin signalling pathway (315). Also, disruption of the *ASD-implicated* TSC1 and TSC2 that are known to activate the β-catenin destruction complex would also yield similar outcome to copy number changes of other candidate genes, inhibiting the destruction complex (316). Lastly, duplication of *LRPAP1*, which promotes the *Wnt*-activated receptor complex that stabilises β-catenin, might in turn increase the cellular level of β-catenin, exerting a similar effect to inhibition of β-catenin destruction complex (Figure 4.6). Taken together, the outcome of copy number changes or disruption of each of these genes is likely to the up-regulation of the *Wnt*-stimulated TCF/LEF-dependent transcription, a pathway whose down-regulation may have therapeutic benefits in autism (317)(318)(319).

4.5.5.5 Crosstalk between signalling cascades

Strikingly, crosstalk between these signalling cascades also supports the outcomes from the copy number changes of candidate genes in each of the biological processes and signalling pathways discussed above (Figure 4.7). The inhibitory molecule AKT within the PI3K pathway may be over-activated by increased PIP3 arising from *ASD-implicated* PTEN, and hence would over-inhibit the β-catenin destruction complex in the *Wnt* signalling pathway. In addition, a reduced production of DAG due to disrupted PTEN (a *ASD-implicated* gene) or PLCG2 (an ASD deleted gene) may affect the activation of KRAS (deleted) within the MAPK3 pathway, which, in turn, would down-regulate the MAPK pathway. Also, increased level of NO* by duplicated NOS1 not only may over-activate PRKG1 leading to abnormal calcium homeostasis, but may also inhibit MAPK3 within the MAPK3 pathway (291), where other proteins from candidate genes within the pathway also act towards reduction of CREB-dependent transcription of the MAPK pathway.

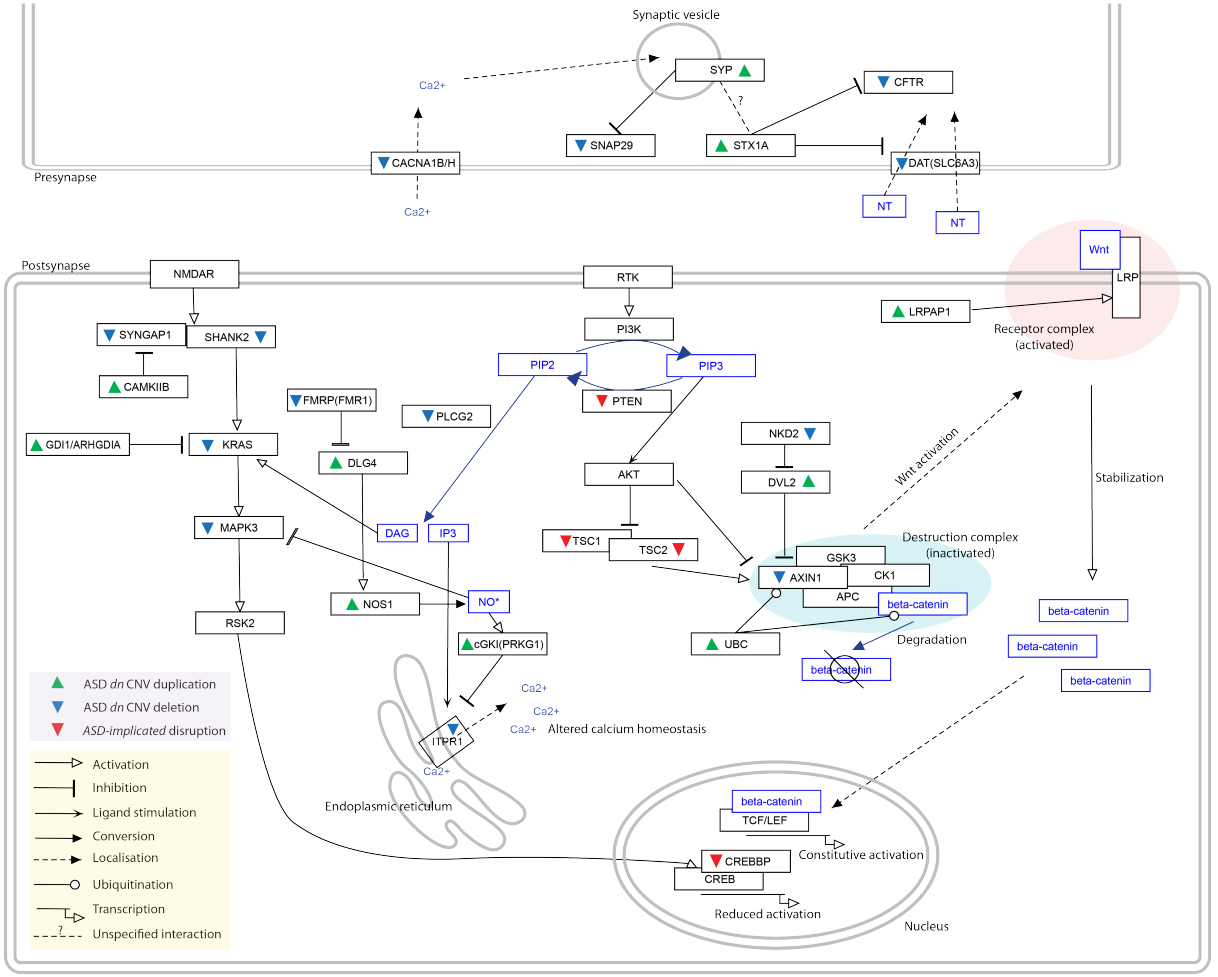


Figure 4.7: Copy number changes in the ASD protein-protein interaction network implicate MAPK pathway, *Wnt*/β-catenin signalling pathway and altered calcium homeostasis in ASD. - Candidate genes identified in this study contribute to signalling cascades and their direction of copy number changes suggests convergent effects on these pathways.

4.5.6 Conclusions

Considering the rapidly-growing list of genetic variants that are being implicated in autism, the forthcoming challenge will be to understand how these large number of variants play roles in the biological processes and pathways underlying autism. Here, I have identified a large protein interactome where the participating genes are found to be affected by copy number variants identified in autistic individuals. Importantly, I suggested a model showing how the direction of copy number changes can be used to interpret perturbations of autism-relevant pathways. Regulating these perturbed signaling cascades against the convergent outcome from copy number changes of candidate genes may provide potential therapeutic options for autism. In particular, pathological gene duplications may serve as attractive targets for knock-down treatments, as a way to alleviate the outcome of dysregulated cascades.

4.6 Appendices

4.6.1 Genes identified via mouse phenotypic enrichments

Table 4.4: Candidate gene sets derived from ASD *dn* CNVs via mouse functional genomics described in Chapter 3 - *AST* (Abnormal Synaptic Transmission) genes consist of 59 genes that contribute to the triplicated *abnormal synaptic transmission* phenotypic associations. *AAP* (All Associated Phenotypes) genes consist of 185 genes that are selected from all of the discovered phenotypic associations described in Chapter 3.

***AST* genes**	***AAP* genes**		
APBA2	*AARS*	*FBXL20*	*MSI1*
BAIAP2	*ACADS*	*FBXO45*	*NCDN*
CACNA1B	*ACADVL*	*FEZF1*	*NCK2*
CACNA1H	*ACTG1*	*FGFR3*	*NCOA6*
CADPS2	*AGPAT2*	*FGFRL1*	*NDN*
CALB2	*AIPL1*	*FLNC*	*NEO1*
CAMK2B	*APBA2*	*FMN2*	*NETO1*
CAMKK2	*ARHGDIA*	*FOXP2*	*NKD2*
CHRNA4	*ARSA*	*FOXP3*	*NLGN2*
CHRNA7	*ASIP*	*FOXS1*	*NLGN3*
CHRNB1	*AXIN1*	*FSCN2*	*NOS1*
CPLX3	*B4GALT2*	*FZD9*	*NOTCH1*
CTNND2	*BAI2*	*GABRA5*	*NPB*
DGCR8	*BAIAP2*	*GABRB3*	*NR2E3*
DLG4	*BBS4*	*GABRD*	*OCA2*
DLGAP3	*BCL2*	*GAN*	*OPRL1*
DNAJC5	*BCL2L1*	*GDI1*	*P2RX4*
DNM1	*BHLHE23*	*GNB1L*	*P2RX7*
DOC2A	*BHLHE41*	*GPR85*	*PDE6B*
EEF1A2	*BMI1*	*GRK5*	*PDE6G*
FBXL20	*BSX*	*GRM8*	*PHC2*
FOXP2	*CACNA1B*	*GTF2IRD1*	*PLXNA3*
GABRA5	*CACNA1F*	*HBA2*	*PLXNB2*
GABRB3	*CACNA1H*	*HCN2*	*POU3F3*
GABRD	*CADPS2*	*HCRTR1*	*PPM1F*
GDI1	*CALB2*	*HERC2*	*PRKAB1*
GNB1L	*CAMK2B*	*HES5*	*PRKG1*
GRK5	*CAMKK2*	*HEXA*	*PRODH*
HCN2	*CAPN7*	*HGS*	*PTHLH*
ITPR1	*CAV1*	*HPCA*	*PTPRZ1*
KCND2	*CFTR*	*HRK*	*RAC3*
LEP	*CHD2*	*HSD17B2*	*RBL1*
LIMK1	*CHRNA4*	*HTT*	*RTN4R*
MAPK3	*CHRNA7*	*HYDIN*	*SDC3*
MAPK8IP2	*CHRNB1*	*IDUA*	*SEPT5*
MAPT	*CIT*	*IKBKG*	*SEZ6L2*
MBP	*CLIP2*	*ISLR2*	*SLC12A7*

Continued on next page

***AST* genes**	***AAP* genes**		
NETO1	*COLQ*	*ITPR1*	*SLC13A1*
NLGN2	*COMT*	*ITPR2*	*SLC27A4*
NLGN3	*CPLX3*	*KCND2*	*SLC5A7*
NOS1	*CRHR1*	*KCNQ2*	*SLC6A3*
OPRL1	*CRKL*	*KHDRBS1*	*SLC6A9*
P2RX4	*CSK*	*KISS1R*	*SMO*
PRKG1	*CSNK1D*	*KRAS*	*SNRPN*
PTPRZ1	*CTNND2*	*LEP*	*SNTA1*
SDC3	*CYP11A1*	*LHX1*	*ST8SIA1*
SEPT5	*CYP1A2*	*LIMK1*	*STUB1*
SEZ6L2	*DGCR8*	*LMF1*	*STX1A*
SLC5A7	*DLG1*	*LRPAP1*	*SUMF1*
SLC6A3	*DLG4*	*LRRC4*	*SYNGAP1*
SLC6A9	*DLGAP3*	*MAFG*	*SYP*
SNTA1	*DNAH5*	*MAFK*	*TACC3*
STX1A	*DNAJC5*	*MAPK1*	*TBX1*
SYNGAP1	*DNM1*	*MAPK3*	*TP73*
SYP	*DOC2A*	*MAPK8IP2*	*TSHZ1*
TBX1	*DOK7*	*MAPT*	*TSPAN12*
UBE3A	*DVL2*	*MBP*	*TYMP*
VLDLR	*E2F1*	*MC4R*	*UBE3A*
ZDHHC8	*EEF1A2*	*MET*	*USP18*
	EMD	*MITF*	*VAC14*
	FABP3	*MMP24*	*VLDLR*
			ZDHHC8

4.6.2 GO functional enrichments

Table 4.5: GO terms significantly enriched among *AST-associated* genes - P, probability; N, total genes in the genome for each GO category (i.e. biological processes, molecular function, and cellular component); G, total genes annotated; n, number of input genes; x, number of annotated genes among the input genes; fold change, fold change of associated genes over expected by random chance; GO term, term description

GOID	P	N	G	n	x	fold change	GO term
32502	1.65×10^{-13}	15685	3852	217	101	1.90	developmental process
51234	2.49×10^{-13}	15685	3442	217	91	1.91	establishment of localization
19904	7.82×10^{-13}	17373	492	228	33	5.11	protein domain specific binding
50804	1.06×10^{-12}	15685	163	217	18	7.98	regulation of synaptic transmission
6468	1.28×10^{-12}	15685	1114	217	47	3.05	protein phosphorylation
31399	1.52×10^{-12}	15685	787	217	38	3.49	regulation of protein modification process
65007	2.36×10^{-12}	15685	8625	217	170	1.42	biological regulation
7268	3.06×10^{-12}	15685	454	217	41	6.53	synaptic transmission

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
45202	3.44×10^{-12}	17415	438	221	43	7.74	synapse
50890	3.88×10^{-12}	15685	144	217	19	9.54	cognition
7267	4.11×10^{-12}	15685	739	217	45	4.40	cell-cell signaling
43005	5.29×10^{-12}	17415	461	221	33	5.64	neuron projection
42592	5.91×10^{-12}	15685	923	217	41	3.21	homeostatic process
23052	6.09×10^{-12}	15685	1999	217	84	3.04	signaling
19725	7.48×10^{-12}	15685	539	217	32	4.29	cellular homeostasis
50789	7.96×10^{-12}	15685	8161	217	164	1.45	regulation of biological process
44463	8.89×10^{-12}	17415	464	221	28	4.76	cell projection part
6810	9.19×10^{-12}	15685	3401	217	91	1.93	transport
44057	9.71×10^{-12}	15685	385	217	29	5.44	regulation of system process
55082	1.06×10^{-11}	15685	448	217	29	4.68	cellular chemical homeostasis
3008	1.09×10^{-11}	15685	1544	217	57	2.67	system process
45211	1.11×10^{-11}	17415	162	221	20	9.73	postsynaptic membrane
65009	1.11×10^{-11}	15685	1417	217	53	2.70	regulation of molecular function
16310	1.18×10^{-11}	15685	1203	217	48	2.88	phosphorylation
6873	1.33×10^{-11}	15685	430	217	28	4.71	cellular ion homeostasis
23051	1.35×10^{-11}	15685	1544	217	61	2.86	regulation of signaling
43025	1.38×10^{-11}	17415	217	221	20	7.26	neuronal cell body
32501	1.47×10^{-11}	15685	4009	217	101	1.82	multicellular organismal process
7610	1.54×10^{-11}	15685	432	217	33	5.52	behavior
50801	1.59×10^{-11}	15685	475	217	29	4.41	ion homeostasis
5515	1.60×10^{-11}	17373	10841	228	197	1.38	protein binding
71840	1.72×10^{-11}	15685	3403	217	90	1.91	cellular component organization or biogenesis
7154	1.92×10^{-11}	15685	1722	217	67	2.81	cell communication
51239	1.95×10^{-11}	15685	1381	217	56	2.93	regulation of multicellular organismal process
42995	2.25×10^{-11}	17415	976	221	50	4.04	cell projection
65008	2.46×10^{-11}	15685	2092	217	81	2.80	regulation of biological quality
97060	2.47×10^{-11}	17415	188	221	22	9.22	synaptic membrane
30054	2.49×10^{-11}	17415	609	221	37	4.79	cell junction
14069	2.52×10^{-11}	17415	96	221	14	11.49	postsynaptic density
42391	2.57×10^{-11}	15685	117	217	17	10.50	regulation of membrane potential
48856	2.66×10^{-11}	15685	2336	217	71	2.20	anatomical structure development
19226	3.08×10^{-11}	15685	188	217	19	7.31	transmission of nerve impulse
35637	3.08×10^{-11}	15685	188	217	19	7.31	multicellular organismal signaling
16043	3.11×10^{-11}	15685	3372	217	89	1.91	cellular component organization

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
51969	3.53×10^{-11}	15685	175	217	18	7.43	regulation of transmission of nerve impulse
44297	3.83×10^{-11}	17415	232	221	20	6.79	cell body
267	3.93×10^{-11}	17415	1257	221	47	2.95	cell fraction
44456	4.03×10^{-11}	17415	320	221	35	8.62	synapse part
19717	4.40×10^{-11}	17415	121	221	15	9.77	synaptosome
7611	4.97×10^{-11}	15685	133	217	16	8.70	learning or memory
50794	5.87×10^{-11}	15685	7745	217	154	1.44	regulation of cellular process
31644	1.41×10^{-10}	15685	188	217	18	6.92	regulation of neurological system process
1932	1.54×10^{-10}	15685	626	217	32	3.69	regulation of protein phosphorylation
43234	1.60×10^{-10}	17415	3229	221	81	1.98	protein complex
6793	1.66×10^{-10}	15685	1379	217	50	2.62	phosphorus metabolic process
6796	1.66×10^{-10}	15685	1379	217	50	2.62	phosphate metabolic process
51899	1.91×10^{-10}	15685	55	217	11	14.46	membrane depolarization
42325	1.94×10^{-10}	15685	670	217	33	3.56	regulation of phosphorylation
5829	2.22×10^{-10}	17415	2092	221	61	2.30	cytosol
7010	2.29×10^{-10}	15685	601	217	31	3.73	cytoskeleton organization
5626	3.74×10^{-10}	17415	959	221	38	3.12	insoluble fraction
19220	4.21×10^{-10}	15685	690	217	33	3.46	regulation of phosphate metabolic process
51174	4.21×10^{-10}	15685	690	217	33	3.46	regulation of phosphorus metabolic process
45859	4.73×10^{-10}	15685	475	217	27	4.11	regulation of protein kinase activity
48878	4.86×10^{-10}	15685	618	217	31	3.63	chemical homeostasis
48519	4.89×10^{-10}	15685	2489	217	71	2.06	negative regulation of biological process
71841	6.67×10^{-10}	15685	2617	217	73	2.02	cellular component organization or biogenesis at cellular level
35176	6.96×10^{-10}	15685	24	217	8	24.09	social behavior
51179	9.96×10^{-10}	15685	1177	217	44	2.70	localization
71842	1.04×10^{-9}	15685	2587	217	72	2.01	cellular component organization at cellular level
43549	1.07×10^{-9}	15685	493	217	27	3.96	regulation of kinase activity
43405	1.20×10^{-9}	15685	216	217	18	6.02	regulation of MAP kinase activity
19222	1.39×10^{-9}	15685	3944	217	95	1.74	regulation of metabolic process
50877	1.71×10^{-9}	15685	1152	217	43	2.70	neurological system process
42221	1.72×10^{-9}	15685	2344	217	67	2.07	response to chemical stimulus
51338	2.06×10^{-9}	15685	508	217	27	3.84	regulation of transferase activity

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
32268	2.14×10^{-9}	15685	1028	217	40	2.81	regulation of cellular protein metabolic process
5856	2.58×10^{-9}	17415	1574	221	49	2.45	cytoskeleton
30425	2.69×10^{-9}	17415	190	221	16	6.64	dendrite
32991	3.14×10^{-9}	17415	3888	221	88	1.78	macromolecular complex
5886	3.40×10^{-9}	17415	4157	221	92	1.74	plasma membrane
46903	3.86×10^{-9}	15685	713	217	32	3.24	secretion
10646	4.04×10^{-9}	15685	1095	217	41	2.71	regulation of cell communication
50790	4.35×10^{-9}	15685	1143	217	42	2.66	regulation of catalytic activity
7612	4.63×10^{-9}	15685	73	217	11	10.89	learning
71900	5.16×10^{-9}	15685	295	217	20	4.90	regulation of protein serine/threonine kinase activity
51246	5.62×10^{-9}	15685	1153	217	42	2.63	regulation of protein metabolic process
30030	6.06×10^{-9}	15685	498	217	26	3.77	cell projection organization
5624	6.20×10^{-9}	17415	925	221	35	2.98	membrane fraction
5737	6.41×10^{-9}	17415	8582	221	151	1.39	cytoplasm
48523	9.62×10^{-9}	15685	2279	217	64	2.03	negative regulation of cellular process
43228	9.65×10^{-9}	17415	2837	221	70	1.94	non-membrane-bounded organelle
43232	9.65×10^{-9}	17415	2837	221	70	1.94	intracellular non-membrane-bounded organelle
51049	1.17×10^{-8}	15685	747	217	32	3.10	regulation of transport
50896	1.26×10^{-8}	15685	6717	217	134	1.44	response to stimulus
30424	1.33×10^{-8}	17415	242	221	17	5.54	axon
187	1.86×10^{-8}	15685	125	217	13	7.52	activation of MAPK activity
44422	2.07×10^{-8}	17415	5841	221	114	1.54	organelle part
34702	2.21×10^{-8}	17415	220	221	16	5.73	ion channel complex
30036	2.26×10^{-8}	15685	322	217	20	4.49	actin cytoskeleton organization
6816	2.29×10^{-8}	15685	203	217	16	5.70	calcium ion transport
30029	2.81×10^{-8}	15685	359	217	21	4.23	actin filament-based process
48167	4.01×10^{-8}	15685	70	217	10	10.33	regulation of synaptic plasticity
43406	4.10×10^{-8}	15685	158	217	14	6.40	positive regulation of MAP kinase activity
19899	4.35×10^{-8}	17373	786	228	31	3.01	enzyme binding
6996	5.08×10^{-8}	15685	1688	217	51	2.18	organelle organization
51705	6.20×10^{-8}	15685	40	217	8	14.46	behavioral interaction between organisms
48869	6.66×10^{-8}	15685	2069	217	58	2.03	cellular developmental process
9966	6.70×10^{-8}	15685	1305	217	43	2.38	regulation of signal transduction

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
32940	6.79×10^{-8}	15685	600	217	27	3.25	secretion by cell
60078	8.15×10^{-8}	15685	28	217	7	18.07	regulation of postsynaptic membrane potential
48518	1.09×10^{-7}	15685	2824	217	71	1.82	positive regulation of biological process
15276	1.32×10^{-7}	17373	129	228	12	7.09	ligand-gated ion channel activity
22834	1.32×10^{-7}	17373	129	228	12	7.09	ligand-gated channel activity
9653	1.44×10^{-7}	15685	1341	217	43	2.32	anatomical structure morphogenesis
44430	1.46×10^{-7}	17415	1100	221	36	2.58	cytoskeletal part
51649	1.80×10^{-7}	15685	1551	217	47	2.19	establishment of localization in cell
60255	1.89×10^{-7}	15685	3335	217	79	1.71	regulation of macromolecule metabolic process
32947	1.95×10^{-7}	17373	33	228	7	16.16	protein complex scaffold
43412	2.05×10^{-7}	15685	2357	217	62	1.90	macromolecule modification
22836	2.44×10^{-7}	17373	320	228	18	4.29	gated channel activity
30001	2.45×10^{-7}	15685	599	217	26	3.14	metal ion transport
71902	2.57×10^{-7}	15685	183	217	14	5.53	positive regulation of protein serine/threonine kinase activity
50793	2.73×10^{-7}	15685	1037	217	36	2.51	regulation of developmental process
70838	2.78×10^{-7}	15685	243	217	16	4.76	divalent metal ion transport
51641	2.79×10^{-7}	15685	603	217	26	3.12	cellular localization
6464	3.01×10^{-7}	15685	2271	217	60	1.91	protein modification process
72511	3.46×10^{-7}	15685	247	217	16	4.68	divalent inorganic cation transport
44093	3.54×10^{-7}	15685	822	217	31	2.73	positive regulation of molecular function
12506	3.56×10^{-7}	17415	236	221	15	5.01	vesicle membrane
31323	3.62×10^{-7}	15685	3568	217	82	1.66	regulation of cellular metabolic process
8283	3.62×10^{-7}	15685	1337	217	42	2.27	cell proliferation
44446	3.72×10^{-7}	17415	5762	221	109	1.49	intracellular organelle part
51704	3.83×10^{-7}	15685	825	217	31	2.72	multi-organism process
48169	4.06×10^{-7}	15685	22	217	6	19.71	regulation of long-term neuronal synaptic plasticity
5230	4.10×10^{-7}	17373	72	228	9	9.52	extracellular ligand-gated ion channel activity
43229	4.16×10^{-7}	17415	10129	221	164	1.28	intracellular organelle
10959	4.42×10^{-7}	15685	137	217	12	6.33	regulation of metal ion transport
43085	4.70×10^{-7}	15685	703	217	28	2.88	positive regulation of catalytic activity

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
6461	4.74×10^{-7}	15685	620	217	26	3.03	protein complex assembly
45860	4.89×10^{-7}	15685	319	217	18	4.08	positive regulation of protein kinase activity
22607	4.92×10^{-7}	15685	1303	217	41	2.27	cellular component assembly
9987	5.03×10^{-7}	15685	13399	217	208	1.12	cellular process
6811	5.30×10^{-7}	15685	973	217	34	2.53	ion transport
48011	5.34×10^{-7}	15685	224	217	15	4.84	nerve growth factor receptor signaling pathway
48812	5.34×10^{-7}	15685	224	217	15	4.84	neuron projection morphogenesis
43226	5.78×10^{-7}	17415	10167	221	164	1.27	organelle
51924	6.08×10^{-7}	15685	116	217	11	6.85	regulation of calcium ion transport
48513	6.54×10^{-7}	15685	982	217	34	2.50	organ development
44444	6.96×10^{-7}	17415	6195	221	114	1.45	cytoplasmic part
35556	7.36×10^{-7}	15685	1680	217	48	2.07	intracellular signal transduction
32879	7.51×10^{-7}	15685	988	217	34	2.49	regulation of localization
48583	7.75×10^{-7}	15685	1895	217	52	1.98	regulation of response to stimulus
60402	8.21×10^{-7}	15685	55	217	8	10.51	calcium ion transport into cytosol
33674	8.35×10^{-7}	15685	331	217	18	3.93	positive regulation of kinase activity
7167	8.36×10^{-7}	15685	767	217	29	2.73	enzyme linked receptor protein signaling pathway
7204	8.78×10^{-7}	15685	146	217	12	5.94	elevation of cytosolic calcium ion concentration
7015	9.11×10^{-7}	15685	174	217	13	5.40	actin filament organization
33043	9.11×10^{-7}	15685	333	217	18	3.91	regulation of organelle organization
60401	9.47×10^{-7}	15685	56	217	8	10.33	cytosolic calcium ion transport
32870	1.04×10^{-6}	15685	336	217	18	3.87	cellular response to hormone stimulus
6812	1.05×10^{-6}	15685	689	217	27	2.83	cation transport
48522	1.07×10^{-6}	15685	2527	217	63	1.80	positive regulation of cellular process
48858	1.09×10^{-6}	15685	237	217	15	4.57	cell projection morphogenesis
9725	1.09×10^{-6}	15685	648	217	26	2.90	response to hormone stimulus
80090	1.10×10^{-6}	15685	3537	217	80	1.63	regulation of primary metabolic process
7173	1.11×10^{-6}	15685	99	217	10	7.30	epidermal growth factor receptor signaling pathway
30659	1.14×10^{-6}	17415	225	221	14	4.90	cytoplasmic vesicle membrane

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
7169	1.19×10^{-6}	15685	568	217	24	3.05	transmembrane receptor protein tyrosine kinase signaling pathway
51347	1.23×10^{-6}	15685	340	217	18	3.83	positive regulation of transferase activity
51046	1.34×10^{-6}	15685	342	217	18	3.80	regulation of secretion
51716	1.36×10^{-6}	15685	5092	217	104	1.48	cellular response to stimulus
71822	1.37×10^{-6}	15685	742	217	28	2.73	protein complex subunit organization
6836	1.39×10^{-6}	15685	126	217	11	6.31	neurotransmitter transport
32990	1.42×10^{-6}	15685	242	217	15	4.48	cell part morphogenesis
44424	1.47×10^{-6}	17415	11891	221	182	1.21	intracellular part
33267	1.49×10^{-6}	17415	111	221	10	7.10	axon part
60537	1.51×10^{-6}	15685	127	217	11	6.26	muscle tissue development
8104	1.61×10^{-6}	15685	420	217	20	3.44	protein localization
44087	1.70×10^{-6}	15685	214	217	14	4.73	regulation of cellular component biogenesis
51480	1.78×10^{-6}	15685	156	217	12	5.56	cytosolic calcium ion homeostasis
16247	1.83×10^{-6}	17373	64	228	8	9.52	channel regulator activity
8306	1.86×10^{-6}	15685	43	217	7	11.77	associative learning
60341	1.90×10^{-6}	15685	463	217	21	3.28	regulation of cellular localization
16192	1.95×10^{-6}	15685	800	217	29	2.62	vesicle-mediated transport
21987	2.18×10^{-6}	15685	44	217	7	11.50	cerebral cortex development
71495	2.27×10^{-6}	15685	355	217	18	3.67	cellular response to endogenous stimulus
7519	2.36×10^{-6}	15685	84	217	9	7.74	skeletal muscle tissue development
4674	2.36×10^{-6}	17373	578	228	23	3.03	protein serine/threonine kinase activity
6874	2.48×10^{-6}	15685	221	217	14	4.58	cellular calcium ion homeostasis
44085	2.52×10^{-6}	15685	287	217	16	4.03	cellular component biogenesis
1666	2.60×10^{-6}	15685	191	217	13	4.92	response to hypoxia
44267	2.61×10^{-6}	15685	3064	217	71	1.67	cellular protein metabolic process
32989	2.75×10^{-6}	15685	435	217	20	3.32	cellular component morphogenesis
8542	2.94×10^{-6}	15685	30	217	6	14.46	visual learning
55074	3.23×10^{-6}	15685	226	217	14	4.48	calcium ion homeostasis
72503	3.58×10^{-6}	15685	228	217	14	4.44	cellular divalent inorganic cation homeostasis
6936	3.58×10^{-6}	15685	228	217	14	4.44	muscle contraction
31982	3.64×10^{-6}	17415	801	221	27	2.66	vesicle
43269	4.10×10^{-6}	15685	169	217	12	5.13	regulation of ion transport

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
23061	4.27×10^{-6}	15685	299	217	16	3.87	signal release
50878	4.34×10^{-6}	15685	570	217	23	2.92	regulation of body fluid levels
4672	4.36×10^{-6}	17373	644	228	24	2.84	protein kinase activity
90066	4.56×10^{-6}	15685	201	217	13	4.67	regulation of anatomical structure size
72507	4.84×10^{-6}	15685	234	217	14	4.32	divalent inorganic cation homeostasis
30534	5.06×10^{-6}	15685	92	217	9	7.07	adult behavior
14706	5.54×10^{-6}	15685	118	217	10	6.13	striated muscle tissue development
33036	5.76×10^{-6}	15685	457	217	20	3.16	macromolecule localization
31410	6.09×10^{-6}	17415	775	221	26	2.64	cytoplasmic vesicle
70482	6.29×10^{-6}	15685	207	217	13	4.54	response to oxygen levels
7632	6.37×10^{-6}	15685	34	217	6	12.76	visual behavior
9719	6.56×10^{-6}	15685	715	217	26	2.63	response to endogenous stimulus
42127	7.30×10^{-6}	15685	998	217	32	2.32	regulation of cell proliferation
51493	7.41×10^{-6}	15685	179	217	12	4.85	regulation of cytoskeleton organization
7420	7.41×10^{-6}	15685	179	217	12	4.85	brain development
16044	7.45×10^{-6}	15685	465	217	20	3.11	cellular membrane organization
165	7.63×10^{-6}	15685	313	217	16	3.69	MAPKKK cascade
2000026	7.66×10^{-6}	15685	812	217	28	2.49	regulation of multicellular organismal development
61024	7.94×10^{-6}	15685	467	217	20	3.10	membrane organization
5216	8.04×10^{-6}	17373	409	228	18	3.35	ion channel activity
5488	8.04×10^{-6}	17373	14540	228	213	1.12	binding
3012	8.20×10^{-6}	15685	245	217	14	4.13	muscle system process
44433	8.58×10^{-6}	17415	305	221	15	3.88	cytoplasmic vesicle part
19900	8.67×10^{-6}	17373	259	228	14	4.12	kinase binding
8022	9.27×10^{-6}	17373	161	228	11	5.21	protein C-terminus binding
35264	1.01×10^{-5}	15685	100	217	9	6.51	multicellular organism growth
42471	1.03×10^{-5}	15685	4	217	3	54.21	ear morphogenesis
22838	1.05×10^{-5}	17373	417	228	18	3.29	substrate-specific channel activity
40007	1.07×10^{-5}	15685	646	217	24	2.69	growth
51128	1.15×10^{-5}	15685	877	217	29	2.39	regulation of cellular component organization
10638	1.22×10^{-5}	15685	129	217	10	5.60	positive regulation of organelle organization
71375	1.23×10^{-5}	15685	254	217	14	3.98	cellular response to peptide hormone stimulus
55080	1.25×10^{-5}	15685	363	217	17	3.39	cation homeostasis
10522	1.25×10^{-5}	15685	38	217	6	11.41	regulation of calcium ion transport into cytosol

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
7507	1.29×10^{-5}	15685	189	217	12	4.59	heart development
10604	1.31×10^{-5}	15685	1126	217	34	2.18	positive regulation of macromolecule metabolic process
43434	1.34×10^{-5}	15685	365	217	17	3.37	response to peptide hormone stimulus
7165	1.52×10^{-5}	15685	4481	217	91	1.47	signal transduction
44419	1.54×10^{-5}	15685	408	217	18	3.19	interspecies interaction between organisms
22603	1.60×10^{-5}	15685	409	217	18	3.18	regulation of anatomical structure morphogenesis
15267	1.63×10^{-5}	17373	431	228	18	3.18	channel activity
10033	1.65×10^{-5}	15685	1447	217	40	2.00	response to organic substance
60079	1.66×10^{-5}	15685	24	217	5	15.06	regulation of excitatory postsynaptic membrane potential
22803	1.74×10^{-5}	17373	433	228	18	3.17	passive transmembrane transporter activity
19233	1.83×10^{-5}	15685	60	217	7	8.43	sensory perception of pain
7596	2.02×10^{-5}	15685	498	217	20	2.90	blood coagulation
50817	2.20×10^{-5}	15685	501	217	20	2.89	coagulation
16020	2.22×10^{-5}	17415	7834	221	130	1.31	membrane
48168	2.26×10^{-5}	15685	42	217	6	10.33	regulation of neuronal synaptic plasticity
7599	2.27×10^{-5}	15685	502	217	20	2.88	hemostasis
48745	2.31×10^{-5}	15685	13	217	4	22.24	smooth muscle tissue development
5622	2.33×10^{-5}	17415	12324	221	183	1.17	intracellular
31325	2.52×10^{-5}	15685	1162	217	34	2.11	positive regulation of cellular metabolic process
1505	2.52×10^{-5}	15685	112	217	9	5.81	regulation of neurotransmitter levels
1952	2.60×10^{-5}	15685	43	217	6	10.09	regulation of cell-matrix adhesion
15075	2.70×10^{-5}	17373	767	228	25	2.48	ion transmembrane transporter activity
71844	2.72×10^{-5}	15685	967	217	30	2.24	cellular component assembly at cellular level
45595	2.94×10^{-5}	15685	733	217	25	2.47	regulation of cell differentiation
8219	3.02×10^{-5}	15685	1593	217	42	1.91	cell death
48041	3.05×10^{-5}	15685	27	217	5	13.39	focal adhesion assembly
16265	3.11×10^{-5}	15685	1595	217	42	1.90	death
7411	3.13×10^{-5}	15685	351	217	16	3.29	axon guidance
7049	3.15×10^{-5}	15685	1124	217	33	2.12	cell cycle
8285	3.18×10^{-5}	15685	472	217	19	2.91	negative regulation of cell proliferation
44464	3.28×10^{-5}	17415	16105	221	218	1.07	cell part

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
5623	3.31×10^{-5}	17415	16106	221	218	1.07	cell
7409	3.43×10^{-5}	15685	176	217	11	4.52	axonogenesis
4707	3.52×10^{-5}	17373	15	228	4	20.32	MAP kinase activity
10035	3.77×10^{-5}	15685	318	217	15	3.41	response to inorganic substance
30900	3.79×10^{-5}	15685	67	217	7	7.55	forebrain development
33555	3.86×10^{-5}	15685	46	217	6	9.43	multicellular organismal response to stress
9893	4.04×10^{-5}	15685	1241	217	35	2.04	positive regulation of metabolic process
35094	4.38×10^{-5}	15685	29	217	5	12.46	response to nicotine
48731	4.57×10^{-5}	15685	945	217	29	2.22	system development
7264	4.62×10^{-5}	15685	528	217	20	2.74	small GTPase mediated signal transduction
44306	4.79×10^{-5}	17415	32	221	5	12.31	neuron projection terminus
9887	4.85×10^{-5}	15685	487	217	19	2.82	organ morphogenesis
8134	5.00×10^{-5}	17373	343	228	15	3.33	transcription factor binding
32868	5.25×10^{-5}	15685	253	217	13	3.71	response to insulin stimulus
32844	5.36×10^{-5}	15685	153	217	10	4.72	regulation of homeostatic process
3682	5.41×10^{-5}	17373	230	228	12	3.98	chromatin binding
30003	5.74×10^{-5}	15685	330	217	15	3.29	cellular cation homeostasis
31175	6.29×10^{-5}	15685	222	217	12	3.91	neuron projection development
42802	6.29×10^{-5}	17373	808	228	25	2.36	identical protein binding
43254	6.39×10^{-5}	15685	126	217	9	5.16	regulation of protein complex assembly
16773	6.61×10^{-5}	17373	761	228	24	2.40	phosphotransferase activity alcohol group as acceptor
9791	6.62×10^{-5}	15685	73	217	7	6.93	post-embryonic development
2000021	6.62×10^{-5}	15685	73	217	7	6.93	regulation of ion homeostasis
8092	6.91×10^{-5}	17373	572	228	20	2.66	cytoskeletal protein binding
40011	7.14×10^{-5}	15685	1019	217	30	2.13	locomotion
45893	7.36×10^{-5}	15685	872	217	27	2.24	positive regulation of transcription DNA-dependent
19538	7.50×10^{-5}	15685	3622	217	75	1.50	protein metabolic process
4890	7.65×10^{-5}	17373	18	228	4	16.93	GABA-A receptor activity
16301	7.80×10^{-5}	17373	819	228	25	2.33	kinase activity
50851	7.81×10^{-5}	15685	160	217	10	4.52	antigen receptor-mediated signaling pathway
17157	7.83×10^{-5}	15685	52	217	6	8.34	regulation of exocytosis
9628	7.86×10^{-5}	15685	505	217	19	2.72	response to abiotic stimulus
71310	7.88×10^{-5}	15685	779	217	25	2.32	cellular response to organic substance
48589	7.93×10^{-5}	15685	193	217	11	4.12	developmental growth
1775	7.95×10^{-5}	15685	732	217	24	2.37	cell activation

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
6875	7.96×10^{-5}	15685	301	217	14	3.36	cellular metal ion homeostasis
30154	8.00×10^{-5}	15685	1717	217	43	1.81	cell differentiation
7626	8.15×10^{-5}	15685	130	217	9	5.00	locomotory behavior
32535	8.48×10^{-5}	15685	102	217	8	5.67	regulation of cellular component size
7613	8.73×10^{-5}	15685	53	217	6	8.18	memory
71109	8.77×10^{-5}	15685	7	217	3	30.98	superior temporal gyrus development
6887	9.20×10^{-5}	15685	231	217	12	3.75	exocytosis
65003	9.25×10^{-5}	15685	884	217	27	2.21	macromolecular complex assembly
33198	9.36×10^{-5}	15685	18	217	4	16.06	response to ATP
5261	9.42×10^{-5}	17373	282	228	13	3.51	cation channel activity
45121	9.86×10^{-5}	17415	179	221	10	4.40	membrane raft
15031	9.86×10^{-5}	15685	1140	217	32	2.03	protein transport
8324	1.01×10^{-4}	17373	588	228	20	2.59	cation transmembrane transporter activity
149	1.01×10^{-4}	17373	36	228	5	10.58	SNARE binding
10810	1.01×10^{-4}	15685	78	217	7	6.49	regulation of cell-substrate adhesion
166	1.04×10^{-4}	17373	2428	228	53	1.66	nucleotide binding
16023	1.06×10^{-4}	17415	515	221	18	2.75	cytoplasmic membrane-bounded vesicle
34330	1.06×10^{-4}	15685	166	217	10	4.35	cell junction organization
33554	1.08×10^{-4}	15685	942	217	28	2.15	cellular response to stress
7269	1.10×10^{-4}	15685	79	217	7	6.40	neurotransmitter secretion
51588	1.12×10^{-4}	15685	35	217	5	10.33	regulation of neurotransmitter transport
55065	1.13×10^{-4}	15685	311	217	14	3.25	metal ion homeostasis
48468	1.13×10^{-4}	15685	609	217	21	2.49	cell development
31401	1.16×10^{-4}	15685	312	217	14	3.24	positive regulation of protein modification process
1662	1.17×10^{-4}	15685	19	217	4	15.22	behavioral fear response
9416	1.23×10^{-4}	15685	169	217	10	4.28	response to light stimulus
2682	1.25×10^{-4}	15685	851	217	26	2.21	regulation of immune system process
7243	1.27×10^{-4}	15685	614	217	21	2.47	intracellular protein kinase cascade
50852	1.29×10^{-4}	15685	138	217	9	4.71	T cell receptor signaling pathway
32869	1.30×10^{-4}	15685	204	217	11	3.90	cellular response to insulin stimulus
5215	1.34×10^{-4}	17373	1274	228	33	1.97	transporter activity
45184	1.34×10^{-4}	15685	1159	217	32	2.00	establishment of protein localization

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
2429	1.36×10^{-4}	15685	171	217	10	4.23	immune response-activating cell surface receptor signaling pathway
31988	1.38×10^{-4}	17415	526	221	18	2.70	membrane-bounded vesicle
7172	1.39×10^{-4}	15685	8	217	3	27.11	signal complex assembly
10467	1.40×10^{-4}	15685	2817	217	61	1.57	gene expression
19901	1.43×10^{-4}	17373	217	228	11	3.86	protein kinase binding
2209	1.45×10^{-4}	15685	20	217	4	14.46	behavioral defense response
40014	1.45×10^{-4}	15685	58	217	6	7.48	regulation of multicellular organism growth
70887	1.46×10^{-4}	15685	959	217	28	2.11	cellular response to chemical stimulus
44092	1.48×10^{-4}	15685	486	217	18	2.68	negative regulation of molecular function
32555	1.48×10^{-4}	17373	1974	228	45	1.74	purine ribonucleotide binding
32553	1.50×10^{-4}	17373	1975	228	45	1.74	ribonucleotide binding
51607	1.50×10^{-4}	15685	83	217	7	6.10	defense response to virus
18105	1.50×10^{-4}	15685	83	217	7	6.10	peptidyl-serine phosphorylation
10468	1.56×10^{-4}	15685	2525	217	56	1.60	regulation of gene expression
2768	1.56×10^{-4}	15685	174	217	10	4.15	immune response-regulating cell surface receptor signaling pathway
32271	1.62×10^{-4}	15685	84	217	7	6.02	regulation of protein polymerization
44237	1.63×10^{-4}	15685	8126	217	139	1.24	cellular metabolic process
17076	1.70×10^{-4}	17373	1986	228	45	1.73	purine nucleotide binding
7271	1.77×10^{-4}	15685	21	217	4	13.77	synaptic transmission cholinergic
9790	1.81×10^{-4}	15685	366	217	15	2.96	embryo development
5891	1.85×10^{-4}	17415	23	221	4	13.70	voltage-gated calcium channel complex
71702	1.92×10^{-4}	15685	541	217	19	2.54	organic substance transport
60627	1.98×10^{-4}	15685	146	217	9	4.46	regulation of vesicle-mediated transport
5654	2.03×10^{-4}	17415	1407	221	34	1.90	nucleoplasm
9889	2.03×10^{-4}	15685	2671	217	58	1.57	regulation of biosynthetic process
32956	2.08×10^{-4}	15685	116	217	8	4.98	regulation of actin cytoskeleton organization
55001	2.11×10^{-4}	15685	62	217	6	6.99	muscle cell development
3016	2.14×10^{-4}	15685	22	217	4	13.14	respiratory system process
2757	2.16×10^{-4}	15685	253	217	12	3.43	immune response-activating signal transduction
44428	2.18×10^{-4}	17415	2268	221	48	1.67	nuclear part
46982	2.21×10^{-4}	17373	267	228	12	3.42	protein heterodimerization activity

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
7399	2.24×10^{-4}	15685	640	217	21	2.37	nervous system development
51403	2.30×10^{-4}	15685	63	217	6	6.88	stress-activated MAPK cascade
2764	2.40×10^{-4}	15685	256	217	12	3.39	immune response-regulating signaling pathway
9892	2.41×10^{-4}	15685	887	217	26	2.12	negative regulation of metabolic process
12501	2.43×10^{-4}	15685	1468	217	37	1.82	programmed cell death
43933	2.49×10^{-4}	15685	1042	217	29	2.01	macromolecular complex subunit organization
6935	2.53×10^{-4}	15685	553	217	19	2.48	chemotaxis
42330	2.53×10^{-4}	15685	553	217	19	2.48	taxis
72384	2.57×10^{-4}	15685	23	217	4	12.57	organelle transport along microtubule
42596	2.57×10^{-4}	15685	23	217	4	12.57	fear response
51130	2.60×10^{-4}	15685	297	217	13	3.16	positive regulation of cellular component organization
32970	2.63×10^{-4}	15685	120	217	8	4.82	regulation of actin filament-based process
22857	2.64×10^{-4}	17373	992	228	27	2.07	transmembrane transporter activity
2684	2.71×10^{-4}	15685	556	217	19	2.47	positive regulation of immune system process
10741	2.74×10^{-4}	15685	65	217	6	6.67	negative regulation of intracellular protein kinase cascade
31326	2.76×10^{-4}	15685	2641	217	57	1.56	regulation of cellular biosynthetic process
44260	2.78×10^{-4}	15685	6044	217	109	1.30	cellular macromolecule metabolic process
43197	2.81×10^{-4}	17415	71	221	6	6.66	dendritic spine
44309	2.81×10^{-4}	17415	71	221	6	6.66	neuron spine
50808	2.85×10^{-4}	15685	92	217	7	5.50	synapse organization
5326	2.94×10^{-4}	17373	25	228	4	12.19	neurotransmitter transporter activity
10941	2.96×10^{-4}	15685	1158	217	31	1.94	regulation of cell death
7044	3.04×10^{-4}	15685	43	217	5	8.40	cell-substrate junction assembly
3013	3.04×10^{-4}	15685	225	217	11	3.53	circulatory system process
2028	3.05×10^{-4}	15685	24	217	4	12.05	regulation of sodium ion transport
43170	3.06×10^{-4}	15685	6690	217	118	1.27	macromolecule metabolic process
43623	3.07×10^{-4}	15685	263	217	12	3.30	cellular protein complex assembly
5912	3.13×10^{-4}	17415	169	221	9	4.20	adherens junction

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
43086	3.20×10^{-4}	15685	386	217	15	2.81	negative regulation of catalytic activity
51495	3.23×10^{-4}	15685	67	217	6	6.47	positive regulation of cytoskeleton organization
9605	3.24×10^{-4}	15685	1111	217	30	1.95	response to external stimulus
22008	3.25×10^{-4}	15685	304	217	13	3.09	neurogenesis
8016	3.25×10^{-4}	15685	94	217	7	5.38	regulation of heart contraction
22892	3.28×10^{-4}	17373	1060	228	28	2.01	substrate-specific transporter activity
9314	3.29×10^{-4}	15685	265	217	12	3.27	response to radiation
51928	3.39×10^{-4}	15685	44	217	5	8.21	positive regulation of calcium ion transport
43274	3.41×10^{-4}	17373	11	228	3	20.78	phospholipase binding
16917	3.44×10^{-4}	17373	26	228	4	11.72	GABA receptor activity
8152	3.45×10^{-4}	15685	9213	217	152	1.19	metabolic process
18209	3.47×10^{-4}	15685	95	217	7	5.33	peptidyl-serine modification
45664	3.54×10^{-4}	15685	229	217	11	3.47	regulation of neuron differentiation
10524	3.59×10^{-4}	15685	25	217	4	11.57	positive regulation of calcium ion transport into cytosol
10605	3.62×10^{-4}	15685	810	217	24	2.14	negative regulation of macromolecule metabolic process
32270	3.67×10^{-4}	15685	391	217	15	2.77	positive regulation of cellular protein metabolic process
902	3.69×10^{-4}	15685	308	217	13	3.05	cell morphogenesis
51960	3.69×10^{-4}	15685	308	217	13	3.05	regulation of nervous system development
40008	3.74×10^{-4}	15685	435	217	16	2.66	regulation of growth
22891	3.74×10^{-4}	17373	907	228	25	2.10	substrate-specific transmembrane transporter activity
51291	3.80×10^{-4}	15685	69	217	6	6.29	protein heterooligomerization
16772	3.84×10^{-4}	17373	962	228	26	2.06	transferase activity transferring phosphorus-containing groups
8154	3.94×10^{-4}	15685	97	217	7	5.22	actin polymerization or depolymerization
5245	4.00×10^{-4}	17373	27	228	4	11.29	voltage-gated calcium channel activity
5231	4.03×10^{-4}	17373	48	228	5	7.94	excitatory extracellular ligand-gated ion channel activity
44445	4.05×10^{-4}	17415	175	221	9	4.05	cytosolic part
6915	4.29×10^{-4}	15685	1456	217	36	1.79	apoptosis
42493	4.43×10^{-4}	15685	314	217	13	2.99	response to drug
1568	4.43×10^{-4}	15685	71	217	6	6.11	blood vessel development

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
60047	4.75×10^{-4}	15685	100	217	7	5.06	heart contraction
3015	4.75×10^{-4}	15685	100	217	7	5.06	heart process
50767	4.75×10^{-4}	15685	276	217	12	3.14	regulation of neurogenesis
33596	4.77×10^{-4}	17415	3	221	2	52.53	TSC1-TSC2 complex
40018	4.87×10^{-4}	15685	27	217	4	10.71	positive regulation of multicellular organism growth
51726	5.03×10^{-4}	15685	632	217	20	2.29	regulation of cell cycle
2087	5.24×10^{-4}	15685	12	217	3	18.07	regulation of respiratory gaseous exchange by neurological system process
51893	5.24×10^{-4}	15685	12	217	3	18.07	regulation of focal adhesion assembly
90109	5.24×10^{-4}	15685	12	217	3	18.07	regulation of cell-substrate junction assembly
30551	5.30×10^{-4}	17373	29	228	4	10.51	cyclic nucleotide binding
30168	5.44×10^{-4}	15685	241	217	11	3.30	platelet activation
18076	5.66×10^{-4}	15685	3	217	2	48.19	N-terminal peptidyl-lysine acetylation
6913	5.72×10^{-4}	15685	408	217	15	2.66	nucleocytoplasmic transport
48585	5.78×10^{-4}	15685	498	217	17	2.47	negative regulation of response to stimulus
19002	5.79×10^{-4}	17373	13	228	3	17.58	GMP binding
30072	5.81×10^{-4}	15685	169	217	9	3.85	peptide hormone secretion
51259	5.83×10^{-4}	15685	243	217	11	3.27	protein oligomerization
7166	5.94×10^{-4}	15685	2659	217	56	1.52	cell surface receptor linked signaling pathway
30165	5.99×10^{-4}	17373	79	228	6	5.79	PDZ domain binding
10942	5.99×10^{-4}	15685	593	217	19	2.32	positive regulation of cell death
51169	6.01×10^{-4}	15685	410	217	15	2.64	nuclear transport
7265	6.03×10^{-4}	15685	244	217	11	3.26	Ras protein signal transduction
19905	6.06×10^{-4}	17373	30	228	4	10.16	syntaxin binding
34704	6.07×10^{-4}	17415	31	221	4	10.17	calcium channel complex
51258	6.10×10^{-4}	15685	136	217	8	4.25	protein polymerization
2000112	6.29×10^{-4}	15685	2361	217	51	1.56	regulation of cellular macromolecule biosynthetic process
70161	6.30×10^{-4}	17415	186	221	9	3.81	anchoring junction
2260	6.45×10^{-4}	15685	29	217	4	9.97	lymphocyte homeostasis
46777	6.74×10^{-4}	15685	106	217	7	4.77	protein autophosphorylation
16079	6.74×10^{-4}	15685	13	217	3	16.68	synaptic vesicle exocytosis
60081	6.74×10^{-4}	15685	13	217	3	16.68	membrane hyperpolarization
44065	6.74×10^{-4}	15685	13	217	3	16.68	regulation of respiratory system process
10559	6.74×10^{-4}	15685	13	217	3	16.68	regulation of glycoprotein biosynthetic process

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
30041	6.85×10^{-4}	15685	77	217	6	5.63	actin filament polymerization
51090	7.13×10^{-4}	15685	249	217	11	3.19	regulation of sequence-specific DNA binding transcription factor activity
2790	7.17×10^{-4}	15685	174	217	9	3.74	peptide secretion
1934	7.17×10^{-4}	15685	174	217	9	3.74	positive regulation of protein phosphorylation
31324	7.17×10^{-4}	15685	799	217	23	2.08	negative regulation of cellular metabolic process
10556	7.25×10^{-4}	15685	2436	217	52	1.54	regulation of macromolecule biosynthetic process
46928	7.36×10^{-4}	15685	30	217	4	9.64	regulation of neurotransmitter secretion
1508	7.36×10^{-4}	15685	30	217	4	9.64	regulation of action potential
10975	7.39×10^{-4}	15685	140	217	8	4.13	regulation of neuron projection development
44459	7.47×10^{-4}	17415	1945	221	41	1.66	plasma membrane part
9888	7.47×10^{-4}	15685	604	217	19	2.27	tissue development
46983	7.53×10^{-4}	17373	687	228	20	2.22	protein dimerization activity
45927	7.53×10^{-4}	15685	108	217	7	4.68	positive regulation of growth
35639	7.65×10^{-4}	17373	1942	228	42	1.65	purine ribonucleoside triphosphate binding
10038	7.80×10^{-4}	15685	213	217	10	3.39	response to metal ion
6921	8.09×10^{-4}	15685	53	217	5	6.82	cellular component disassembly involved in apoptosis
30055	8.12×10^{-4}	17415	119	221	7	4.64	cell-substrate junction
10557	8.27×10^{-4}	15685	707	217	21	2.15	positive regulation of macromolecule biosynthetic process
31056	8.36×10^{-4}	15685	31	217	4	9.33	regulation of histone modification
1776	8.36×10^{-4}	15685	31	217	4	9.33	leukocyte homeostasis
42136	8.49×10^{-4}	15685	14	217	3	15.49	neurotransmitter biosynthetic process
3712	8.52×10^{-4}	17373	400	228	14	2.67	transcription cofactor activity
122	8.68×10^{-4}	15685	425	217	15	2.55	negative regulation of transcription from RNA polymerase II promoter
43195	8.68×10^{-4}	17415	34	221	4	9.27	terminal button
50688	8.82×10^{-4}	15685	54	217	5	6.69	regulation of defense response to virus
989	8.94×10^{-4}	17373	402	228	14	2.65	transcription factor binding transcription factor activity
51171	8.99×10^{-4}	15685	2704	217	56	1.50	regulation of nitrogen compound metabolic process
14070	8.99×10^{-4}	15685	217	217	10	3.33	response to organic cyclic compound

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
34707	9.02×10^{-4}	17415	59	221	5	6.68	chloride channel complex
43198	9.71×10^{-4}	17415	35	221	4	9.01	dendritic shaft
51336	9.72×10^{-4}	15685	522	217	17	2.35	regulation of hydrolase activity
6950	9.73×10^{-4}	15685	2285	217	49	1.55	response to stress
6355	9.92×10^{-4}	15685	2777	217	57	1.48	regulation of transcription DNA-dependent
5634	9.93×10^{-4}	17415	5650	221	94	1.31	nucleus
17081	1.01×10^{-3}	17373	4	228	2	38.10	chloride channel regulator activity
34329	1.02×10^{-3}	15685	147	217	8	3.93	cell junction assembly
10628	1.02×10^{-3}	15685	620	217	19	2.22	positive regulation of gene expression
30135	1.03×10^{-3}	17415	124	221	7	4.45	coated vesicle
7585	1.04×10^{-3}	15685	56	217	5	6.45	respiratory gaseous exchange
5938	1.06×10^{-3}	17415	161	221	8	3.92	cell cortex
31098	1.07×10^{-3}	15685	184	217	9	3.54	stress-activated protein kinase signaling cascade
43067	1.09×10^{-3}	15685	1142	217	29	1.84	regulation of programmed cell death
42803	1.10×10^{-3}	17373	458	228	15	2.50	protein homodimerization activity
4889	1.10×10^{-3}	17373	16	228	3	14.29	nicotinic acetylcholine-activated cation-selective channel activity
6357	1.11×10^{-3}	15685	877	217	24	1.98	regulation of transcription from RNA polymerase II promoter
50673	1.11×10^{-3}	15685	185	217	9	3.52	epithelial cell proliferation
6089	1.12×10^{-3}	15685	4	217	2	36.14	lactate metabolic process
6184	1.17×10^{-3}	15685	306	217	12	2.83	GTP catabolic process
988	1.18×10^{-3}	17373	414	228	14	2.58	protein binding transcription factor activity
3713	1.19×10^{-3}	17373	237	228	10	3.22	transcription coactivator activity
30705	1.19×10^{-3}	15685	34	217	4	8.50	cytoskeleton-dependent intracellular transport
10970	1.19×10^{-3}	15685	34	217	4	8.50	microtubule-based transport
33138	1.19×10^{-3}	15685	34	217	4	8.50	positive regulation of peptidyl-serine phosphorylation
5892	1.20×10^{-3}	17415	17	221	3	13.91	nicotinic acetylcholine-gated receptor-channel complex
71705	1.22×10^{-3}	15685	226	217	10	3.20	nitrogen compound transport
46649	1.23×10^{-3}	15685	440	217	15	2.46	lymphocyte activation
30073	1.26×10^{-3}	15685	152	217	8	3.80	insulin secretion

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
9968	1.26×10^{-3}	15685	396	217	14	2.56	negative regulation of signal transduction
1726	1.31×10^{-3}	17415	64	221	5	6.16	ruffle
43231	1.32×10^{-3}	17415	9016	221	137	1.20	intracellular membrane-bounded organelle
43270	1.32×10^{-3}	15685	59	217	5	6.13	positive regulation of ion transport
51017	1.32×10^{-3}	15685	59	217	5	6.13	actin filament bundle assembly
5246	1.32×10^{-3}	17373	17	228	3	13.45	calcium channel regulator activity
43227	1.36×10^{-3}	17415	9022	221	137	1.20	membrane-bounded organelle
32412	1.42×10^{-3}	15685	60	217	5	6.02	regulation of ion transmembrane transporter activity
31334	1.42×10^{-3}	15685	60	217	5	6.02	positive regulation of protein complex assembly
42110	1.46×10^{-3}	15685	314	217	12	2.76	T cell activation
31234	1.47×10^{-3}	17415	39	221	4	8.08	extrinsic to internal side of plasma membrane
10332	1.48×10^{-3}	15685	36	217	4	8.03	response to gamma radiation
15837	1.49×10^{-3}	15685	193	217	9	3.37	amine transport
50678	1.49×10^{-3}	15685	156	217	8	3.71	regulation of epithelial cell proliferation
30665	1.50×10^{-3}	17415	66	221	5	5.97	clathrin coated vesicle membrane
48872	1.53×10^{-3}	15685	61	217	5	5.92	homeostasis of number of cells
42692	1.54×10^{-3}	15685	122	217	7	4.15	muscle cell differentiation
71514	1.54×10^{-3}	15685	17	217	3	12.76	genetic imprinting
42327	1.54×10^{-3}	15685	194	217	9	3.35	positive regulation of phosphorylation
5217	1.57×10^{-3}	17373	18	228	3	12.70	intracellular ligand-gated ion channel activity
15464	1.57×10^{-3}	17373	18	228	3	12.70	acetylcholine receptor activity
46039	1.58×10^{-3}	15685	317	217	12	2.74	GTP metabolic process
42383	1.60×10^{-3}	17415	67	221	5	5.88	sarcolemma
48646	1.64×10^{-3}	15685	548	217	17	2.24	anatomical structure formation involved in morphogenesis
10627	1.67×10^{-3}	15685	408	217	14	2.48	regulation of intracellular protein kinase cascade
60090	1.72×10^{-3}	17373	131	228	7	4.07	binding bridging
278	1.74×10^{-3}	15685	551	217	17	2.23	mitotic cell cycle
51247	1.74×10^{-3}	15685	456	217	15	2.38	positive regulation of protein metabolic process

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
7417	1.77×10^{-3}	15685	125	217	7	4.05	central nervous system development
45937	1.77×10^{-3}	15685	198	217	9	3.29	positive regulation of phosphate metabolic process
10562	1.77×10^{-3}	15685	198	217	9	3.29	positive regulation of phosphorus metabolic process
22898	1.77×10^{-3}	15685	63	217	5	5.74	regulation of transmembrane transporter activity
5625	1.77×10^{-3}	17415	398	221	13	2.57	soluble fraction
8286	1.81×10^{-3}	15685	161	217	8	3.59	insulin receptor signaling pathway
32273	1.82×10^{-3}	15685	38	217	4	7.61	positive regulation of protein polymerization
43029	1.83×10^{-3}	15685	18	217	3	12.05	T cell homeostasis
48873	1.83×10^{-3}	15685	18	217	3	12.05	homeostasis of number of cells within a tissue
43576	1.83×10^{-3}	15685	18	217	3	12.05	regulation of respiratory gaseous exchange
15833	1.83×10^{-3}	15685	199	217	9	3.27	peptide transport
42983	1.85×10^{-3}	15685	5	217	2	28.91	amyloid precursor protein biosynthetic process
10523	1.85×10^{-3}	15685	5	217	2	28.91	negative regulation of calcium ion transport into cytosol
60501	1.85×10^{-3}	15685	5	217	2	28.91	positive regulation of epithelial cell proliferation involved in lung morphogenesis
1956	1.85×10^{-3}	15685	5	217	2	28.91	positive regulation of neurotransmitter secretion
2000794	1.85×10^{-3}	15685	5	217	2	28.91	regulation of epithelial cell proliferation involved in lung morphogenesis
60017	1.85×10^{-3}	15685	5	217	2	28.91	parathyroid gland development
51589	1.85×10^{-3}	15685	5	217	2	28.91	negative regulation of neurotransmitter transport
42984	1.85×10^{-3}	15685	5	217	2	28.91	regulation of amyloid precursor protein biosynthetic process
31327	1.88×10^{-3}	15685	555	217	17	2.21	negative regulation of cellular biosynthetic process
50680	1.90×10^{-3}	15685	64	217	5	5.65	negative regulation of epithelial cell proliferation
50778	1.91×10^{-3}	15685	414	217	14	2.44	positive regulation of immune response
42981	1.94×10^{-3}	15685	1131	217	28	1.79	regulation of apoptosis

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
47485	1.94×10^{-3}	17373	99	228	6	4.62	protein N-terminus binding
7275	1.95×10^{-3}	15685	1584	217	36	1.64	multicellular organismal development
226	1.96×10^{-3}	15685	201	217	9	3.24	microtubule cytoskeleton organization
32559	1.97×10^{-3}	17373	1607	228	35	1.66	adenyl ribonucleotide binding

Table 4.6: GO terms significantly enriched among *AST-associated* genes from duplication CNVs - P, probability; N, total genes in the genome for each GO category (i.e. biological processes, molecular function, and cellular component); G, total genes annotated; n, number of input genes; x, number of annotated genes among the input genes; fold change, fold change over expected by random chance; GO term, term description

GOID	P	N	G	n	x	fold change	GO term
23052	2.16×10^{-13}	15685	1999	105	46	3.44	signaling
7268	4.46×10^{-12}	15685	454	105	22	7.24	synaptic transmission
7267	8.97×10^{-12}	15685	739	105	25	5.05	cell-cell signaling
45202	1.12×10^{-11}	17415	438	106	21	7.88	synapse
42995	1.55×10^{-11}	17415	976	106	28	4.71	cell projection
7154	1.67×10^{-11}	15685	1722	105	38	3.30	cell communication
44456	4.48×10^{-11}	17415	320	106	17	8.73	synapse part
97060	1.41×10^{-10}	17415	188	106	13	11.36	synaptic membrane
45211	3.20×10^{-10}	17415	162	106	12	12.17	postsynaptic membrane
43005	3.32×10^{-10}	17415	461	106	18	6.41	neuron projection
30054	6.72×10^{-10}	17415	609	106	20	5.40	cell junction
23051	1.30×10^{-8}	15685	1544	105	31	3.00	regulation of signaling
7010	1.54×10^{-8}	15685	601	105	19	4.72	cytoskeleton organization
5230	1.71×10^{-8}	17373	72	110	8	17.55	extracellular ligand-gated ion channel activity
43234	3.35×10^{-8}	17415	3229	106	44	2.24	protein complex
5829	3.95×10^{-8}	17415	2092	106	34	2.67	cytosol
19904	8.09×10^{-8}	17373	492	110	16	5.14	protein domain specific binding
43025	8.61×10^{-8}	17415	217	106	11	8.33	neuronal cell body
65008	1.17×10^{-7}	15685	2092	105	35	2.50	regulation of biological quality
44297	1.69×10^{-7}	17415	232	106	11	7.79	cell body
5515	2.73×10^{-7}	17373	10841	110	93	1.35	protein binding
32991	3.72×10^{-7}	17415	3888	106	47	1.99	macromolecular complex
16043	6.81×10^{-7}	15685	3372	105	45	1.99	cellular component organization
19717	7.35×10^{-7}	17415	121	106	8	10.86	synaptosome
44463	8.42×10^{-7}	17415	464	106	14	4.96	cell projection part

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
71840	8.91×10^{-7}	15685	3403	105	45	1.98	cellular component organization or biogenesis
6873	1.04×10^{-6}	15685	430	105	14	4.86	cellular ion homeostasis
6996	1.23×10^{-6}	15685	1688	105	29	2.57	organelle organization
50794	1.28×10^{-6}	15685	7745	105	76	1.47	regulation of cellular process
15276	1.61×10^{-6}	17373	129	110	8	9.79	ligand-gated ion channel activity
22834	1.61×10^{-6}	17373	129	110	8	9.79	ligand-gated channel activity
30036	1.62×10^{-6}	15685	322	105	12	5.57	actin cytoskeleton organization
55082	1.68×10^{-6}	15685	448	105	14	4.67	cellular chemical homeostasis
5737	2.14×10^{-6}	17415	8582	106	76	1.45	cytoplasm
50789	2.30×10^{-6}	15685	8161	105	78	1.43	regulation of biological process
71842	2.31×10^{-6}	15685	2587	105	37	2.14	cellular component organization at cellular level
19725	2.91×10^{-6}	15685	539	105	15	4.16	cellular homeostasis
71841	3.06×10^{-6}	15685	2617	105	37	2.11	cellular component organization or biogenesis at cellular level
50801	3.33×10^{-6}	15685	475	105	14	4.40	ion homeostasis
4890	4.35×10^{-6}	17373	18	110	4	35.10	GABA-A receptor activity
5626	4.69×10^{-6}	17415	959	106	19	3.26	insoluble fraction
30029	4.96×10^{-6}	15685	359	105	12	4.99	actin filament-based process
267	5.56×10^{-6}	17415	1257	106	22	2.88	cell fraction
32501	5.89×10^{-6}	15685	4009	105	48	1.79	multicellular organismal process
5856	6.26×10^{-6}	17415	1574	106	25	2.61	cytoskeleton
51716	7.54×10^{-6}	15685	5092	105	56	1.64	cellular response to stimulus
44444	7.62×10^{-6}	17415	6195	106	60	1.59	cytoplasmic part
10646	9.82×10^{-6}	15685	1095	105	21	2.86	regulation of cell communication
16310	1.23×10^{-5}	15685	1203	105	22	2.73	phosphorylation
6468	1.28×10^{-5}	15685	1114	105	21	2.82	protein phosphorylation
42391	1.34×10^{-5}	15685	117	105	7	8.94	regulation of membrane potential
9966	1.36×10^{-5}	15685	1305	105	23	2.63	regulation of signal transduction
50804	1.36×10^{-5}	15685	163	105	8	7.33	regulation of synaptic transmission
6810	1.61×10^{-5}	15685	3401	105	42	1.84	transport
43228	1.68×10^{-5}	17415	2837	106	35	2.03	non-membrane-bounded organelle
43232	1.68×10^{-5}	17415	2837	106	35	2.03	intracellular non-membrane-bounded organelle
5886	1.79×10^{-5}	17415	4157	106	45	1.78	plasma membrane

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
48519	1.98×10^{-5}	15685	2489	105	34	2.04	negative regulation of biological process
16917	2.04×10^{-5}	17373	26	110	4	24.30	GABA receptor activity
187	2.06×10^{-5}	15685	125	105	7	8.37	activation of MAPK activity
50877	2.12×10^{-5}	15685	1152	105	21	2.72	neurological system process
51234	2.19×10^{-5}	15685	3442	105	42	1.82	establishment of localization
48523	2.22×10^{-5}	15685	2279	105	32	2.10	negative regulation of cellular process
51969	2.27×10^{-5}	15685	175	105	8	6.83	regulation of transmission of nerve impulse
3008	2.28×10^{-5}	15685	1544	105	25	2.42	system process
7165	2.60×10^{-5}	15685	4481	105	50	1.67	signal transduction
44430	3.24×10^{-5}	17415	1100	106	19	2.84	cytoskeletal part
60402	3.26×10^{-5}	15685	55	105	5	13.58	calcium ion transport into cytosol
65007	3.36×10^{-5}	15685	8625	105	78	1.35	biological regulation
60401	3.56×10^{-5}	15685	56	105	5	13.34	cytosolic calcium ion transport
22836	3.60×10^{-5}	17373	320	110	10	4.94	gated channel activity
42592	3.77×10^{-5}	15685	923	105	18	2.91	homeostatic process
19226	3.80×10^{-5}	15685	188	105	8	6.36	transmission of nerve impulse
31644	3.80×10^{-5}	15685	188	105	8	6.36	regulation of neurological system process
35637	3.80×10^{-5}	15685	188	105	8	6.36	multicellular organismal signaling
5624	4.14×10^{-5}	17415	925	106	17	3.02	membrane fraction
19233	4.98×10^{-5}	15685	60	105	5	12.45	sensory perception of pain
65009	5.04×10^{-5}	15685	1417	105	23	2.42	regulation of molecular function
44057	5.33×10^{-5}	15685	385	105	11	4.27	regulation of system process
7204	5.57×10^{-5}	15685	146	105	7	7.16	elevation of cytosolic calcium ion concentration
34702	5.96×10^{-5}	17415	220	106	8	5.97	ion channel complex
48878	6.29×10^{-5}	15685	618	105	14	3.38	chemical homeostasis
6816	6.54×10^{-5}	15685	203	105	8	5.89	calcium ion transport
44424	7.12×10^{-5}	17415	11891	106	90	1.24	intracellular part
4674	7.39×10^{-5}	17373	578	110	13	3.55	protein serine/threonine kinase activity
51480	8.47×10^{-5}	15685	156	105	7	6.70	cytosolic calcium ion homeostasis
30900	8.49×10^{-5}	15685	67	105	5	11.15	forebrain development
43406	9.18×10^{-5}	15685	158	105	7	6.62	positive regulation of MAP kinase activity
35556	9.22×10^{-5}	15685	1680	105	25	2.22	intracellular signal transduction
48856	9.36×10^{-5}	15685	2336	105	31	1.98	anatomical structure development

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
12506	9.73×10^{-5}	17415	236	106	8	5.57	vesicle membrane
6793	9.82×10^{-5}	15685	1379	105	22	2.38	phosphorus metabolic process
6796	9.82×10^{-5}	15685	1379	105	22	2.38	phosphate metabolic process
31982	9.98×10^{-5}	17415	801	106	15	3.08	vesicle
51239	1.00×10^{-4}	15685	1381	105	22	2.38	regulation of multicellular organismal process
43405	1.01×10^{-4}	15685	216	105	8	5.53	regulation of MAP kinase activity
30424	1.16×10^{-4}	17415	242	106	8	5.43	axon
10522	1.17×10^{-4}	15685	38	105	4	15.72	regulation of calcium ion transport into cytosol
6874	1.18×10^{-4}	15685	221	105	8	5.41	cellular calcium ion homeostasis
2000021	1.28×10^{-4}	15685	73	105	5	10.23	regulation of ion homeostasis
50896	1.33×10^{-4}	15685	6717	105	64	1.42	response to stimulus
55074	1.38×10^{-4}	15685	226	105	8	5.29	calcium ion homeostasis
43229	1.44×10^{-4}	17415	10129	106	80	1.30	intracellular organelle
72503	1.47×10^{-4}	15685	228	105	8	5.24	cellular divalent inorganic cation homeostasis
32502	1.51×10^{-4}	15685	3852	105	43	1.67	developmental process
55080	1.60×10^{-4}	15685	363	105	10	4.12	cation homeostasis
30425	1.62×10^{-4}	17415	190	106	7	6.05	dendrite
7015	1.67×10^{-4}	15685	174	105	7	6.01	actin filament organization
43226	1.71×10^{-4}	17415	10167	106	80	1.29	organelle
72507	1.75×10^{-4}	15685	234	105	8	5.11	divalent inorganic cation homeostasis
50790	1.88×10^{-4}	15685	1143	105	19	2.48	regulation of catalytic activity
4672	2.16×10^{-4}	17373	644	110	13	3.19	protein kinase activity
70838	2.26×10^{-4}	15685	243	105	8	4.92	divalent metal ion transport
71902	2.28×10^{-4}	15685	183	105	7	5.71	positive regulation of protein serine/threonine kinase activity
5231	2.39×10^{-4}	17373	48	110	4	13.16	excitatory extracellular ligand-gated ion channel activity
31410	2.51×10^{-4}	17415	775	106	14	2.97	cytoplasmic vesicle
72511	2.53×10^{-4}	15685	247	105	8	4.84	divalent inorganic cation transport
32309	2.61×10^{-4}	15685	19	105	3	23.59	icosanoid secretion
5216	2.71×10^{-4}	17373	409	110	10	3.86	ion channel activity
45860	2.89×10^{-4}	15685	319	105	9	4.21	positive regulation of protein kinase activity
14069	2.99×10^{-4}	17415	96	106	5	8.56	postsynaptic density
16023	3.00×10^{-4}	17415	515	106	11	3.51	cytoplasmic membrane-bounded vesicle

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
71715	3.06×10^{-4}	15685	20	105	3	22.41	icosanoid transport
22838	3.16×10^{-4}	17373	417	110	10	3.79	substrate-specific channel activity
5516	3.20×10^{-4}	17373	145	110	6	6.54	calmodulin binding
44422	3.27×10^{-4}	17415	5841	106	53	1.49	organelle part
9987	3.27×10^{-4}	15685	13399	105	101	1.13	cellular process
45859	3.36×10^{-4}	15685	475	105	11	3.46	regulation of protein kinase activity
7271	3.55×10^{-4}	15685	21	105	3	21.34	synaptic transmission cholinergic
31988	3.59×10^{-4}	17415	526	106	11	3.44	membrane-bounded vesicle
30003	3.70×10^{-4}	15685	330	105	9	4.07	cellular cation homeostasis
33674	3.78×10^{-4}	15685	331	105	9	4.06	positive regulation of kinase activity
48169	4.09×10^{-4}	15685	22	105	3	20.37	regulation of long-term neuronal synaptic plasticity
15267	4.10×10^{-4}	17373	431	110	10	3.66	channel activity
50890	4.12×10^{-4}	15685	144	105	6	6.22	cognition
50878	4.24×10^{-4}	15685	570	105	12	3.14	regulation of body fluid levels
22803	4.25×10^{-4}	17373	433	110	10	3.65	passive transmembrane transporter activity
30659	4.52×10^{-4}	17415	225	106	7	5.11	cytoplasmic vesicle membrane
34707	4.56×10^{-4}	17415	59	106	4	11.14	chloride channel complex
43549	4.60×10^{-4}	15685	493	105	11	3.33	regulation of kinase activity
51347	4.60×10^{-4}	15685	340	105	9	3.95	positive regulation of transferase activity
44446	4.60×10^{-4}	17415	5762	106	52	1.48	intracellular organelle part
51899	4.97×10^{-4}	15685	55	105	4	10.86	membrane depolarization
7596	5.00×10^{-4}	15685	498	105	11	3.30	blood coagulation

Table 4.7: GO terms significantly enriched among *AST-associated* genes from deletion CNVs - P, probability; N, total genes in the genome for each GO category (i.e. biological processes, molecular function, and cellular component); G, total genes annotated; n, number of input genes; x, number of annotated genes among the input genes; fold change, fold change over expected by random chance; GO term, term description

GOID	P	N	G	n	x	fold change	GO term
45202	1.52×10^{-11}	17415	438	85	18	8.42	synapse
7610	5.78×10^{-11}	15685	432	82	17	7.53	behavior
44456	6.92×10^{-11}	17415	320	85	15	9.60	synapse part
35176	2.07×10^{-9}	15685	24	82	6	47.82	social behavior
32947	1.41×10^{-8}	17373	33	88	6	35.89	protein complex scaffold
14069	2.15×10^{-8}	17415	96	85	8	17.07	postsynaptic density

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
65008	2.24×10^{-8}	15685	2092	82	31	2.83	regulation of biological quality
51705	5.65×10^{-8}	15685	40	82	6	28.69	behavioral interaction between organisms
7268	8.64×10^{-8}	15685	454	82	14	5.90	synaptic transmission
42391	1.69×10^{-7}	15685	117	82	8	13.08	regulation of membrane potential
48518	1.75×10^{-7}	15685	2824	82	35	2.37	positive regulation of biological process
51704	1.91×10^{-7}	15685	825	82	18	4.17	multi-organism process
30054	2.25×10^{-7}	17415	609	85	15	5.05	cell junction
51239	3.66×10^{-7}	15685	1381	82	23	3.19	regulation of multicellular organismal process
31399	5.12×10^{-7}	15685	787	82	17	4.13	regulation of protein modification process
5515	5.64×10^{-7}	17373	10841	88	76	1.38	protein binding
44057	7.10×10^{-7}	15685	385	82	12	5.96	regulation of system process
1932	7.32×10^{-7}	15685	626	82	15	4.58	regulation of protein phosphorylation
48522	1.68×10^{-6}	15685	2527	82	31	2.35	positive regulation of cellular process
42325	1.71×10^{-6}	15685	670	82	15	4.28	regulation of phosphorylation
23052	1.71×10^{-6}	15685	1999	82	27	2.58	signaling
65009	2.30×10^{-6}	15685	1417	82	22	2.97	regulation of molecular function
19220	2.46×10^{-6}	15685	690	82	15	4.16	regulation of phosphate metabolic process
51174	2.46×10^{-6}	15685	690	82	15	4.16	regulation of phosphorus metabolic process
55082	3.45×10^{-6}	15685	448	82	12	5.12	cellular chemical homeostasis
6468	3.45×10^{-6}	15685	1114	82	19	3.26	protein phosphorylation
65007	3.71×10^{-6}	15685	8625	82	65	1.44	biological regulation
19725	4.10×10^{-6}	15685	539	82	13	4.61	cellular homeostasis
32268	4.58×10^{-6}	15685	1028	82	18	3.35	regulation of cellular protein metabolic process
51607	4.64×10^{-6}	15685	83	82	6	13.83	defense response to virus
6793	5.64×10^{-6}	15685	1379	82	21	2.91	phosphorus metabolic process
6796	5.64×10^{-6}	15685	1379	82	21	2.91	phosphate metabolic process
7267	5.68×10^{-6}	15685	739	82	15	3.88	cell-cell signaling
50789	9.46×10^{-6}	15685	8161	82	62	1.45	regulation of biological process
51899	9.78×10^{-6}	15685	55	82	5	17.39	membrane depolarization
16310	1.05×10^{-5}	15685	1203	82	19	3.02	phosphorylation
42221	1.15×10^{-5}	15685	2344	82	28	2.28	response to chemical stimulus
42802	1.20×10^{-5}	17373	808	88	15	3.67	identical protein binding

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
6873	1.41×10^{-5}	15685	430	82	11	4.89	cellular ion homeostasis
45211	1.42×10^{-5}	17415	162	85	7	8.85	postsynaptic membrane
31334	1.50×10^{-5}	15685	60	82	5	15.94	positive regulation of protein complex assembly
46903	1.76×10^{-5}	15685	713	82	14	3.76	secretion
6461	1.84×10^{-5}	15685	620	82	13	4.01	protein complex assembly
5886	1.94×10^{-5}	17415	4157	85	38	1.87	plasma membrane
51246	2.20×10^{-5}	15685	1153	82	18	2.99	regulation of protein metabolic process
50804	2.27×10^{-5}	15685	163	82	7	8.21	regulation of synaptic transmission
19717	2.78×10^{-5}	17415	121	85	6	10.16	synaptosome
33138	2.86×10^{-5}	15685	34	82	4	22.50	positive regulation of peptidyl-serine phosphorylation
51049	2.96×10^{-5}	15685	747	82	14	3.58	regulation of transport
3008	3.17×10^{-5}	15685	1544	82	21	2.60	system process
23051	3.17×10^{-5}	15685	1544	82	21	2.60	regulation of signaling
1934	3.46×10^{-5}	15685	174	82	7	7.70	positive regulation of protein phosphorylation
50801	3.52×10^{-5}	15685	475	82	11	4.43	ion homeostasis
45859	3.52×10^{-5}	15685	475	82	11	4.43	regulation of protein kinase activity
31401	3.56×10^{-5}	15685	312	82	9	5.52	positive regulation of protein modification process
51969	3.59×10^{-5}	15685	175	82	7	7.65	regulation of transmission of nerve impulse
97060	3.71×10^{-5}	17415	188	85	7	7.63	synaptic membrane
48745	3.79×10^{-5}	15685	13	82	3	44.14	smooth muscle tissue development
7612	3.92×10^{-5}	15685	73	82	5	13.10	learning
6810	4.22×10^{-5}	15685	3401	82	34	1.91	transport
32273	4.47×10^{-5}	15685	38	82	4	20.13	positive regulation of protein polymerization
9653	4.78×10^{-5}	15685	1341	82	19	2.71	anatomical structure morphogenesis
43549	4.94×10^{-5}	15685	493	82	11	4.27	regulation of kinase activity
19900	5.03×10^{-5}	17373	259	88	8	6.10	kinase binding
22603	5.22×10^{-5}	15685	409	82	10	4.68	regulation of anatomical structure morphogenesis
60537	5.30×10^{-5}	15685	127	82	6	9.04	muscle tissue development
50896	5.43×10^{-5}	15685	6717	82	53	1.51	response to stimulus
51234	5.46×10^{-5}	15685	3442	82	34	1.89	establishment of localization
4707	5.47×10^{-5}	17373	15	88	3	39.48	MAP kinase activity
19226	5.66×10^{-5}	15685	188	82	7	7.12	transmission of nerve impulse
31644	5.66×10^{-5}	15685	188	82	7	7.12	regulation of neurological system process

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
35637	5.66×10^{-5}	15685	188	82	7	7.12	multicellular organismal signaling
10638	5.78×10^{-5}	15685	129	82	6	8.90	positive regulation of organelle organization
51338	6.47×10^{-5}	15685	508	82	11	4.14	regulation of transferase activity
43085	6.75×10^{-5}	15685	703	82	13	3.54	positive regulation of catalytic activity
7611	6.86×10^{-5}	15685	133	82	6	8.63	learning or memory
42327	6.90×10^{-5}	15685	194	82	7	6.90	positive regulation of phosphorylation
50790	7.15×10^{-5}	15685	1143	82	17	2.84	regulation of catalytic activity
18105	7.28×10^{-5}	15685	83	82	5	11.52	peptidyl-serine phosphorylation
33135	7.33×10^{-5}	15685	43	82	4	17.79	regulation of peptidyl-serine phosphorylation
7519	7.71×10^{-5}	15685	84	82	5	11.39	skeletal muscle tissue development
45937	7.85×10^{-5}	15685	198	82	7	6.76	positive regulation of phosphate metabolic process
10562	7.85×10^{-5}	15685	198	82	7	6.76	positive regulation of phosphorus metabolic process
50793	7.88×10^{-5}	15685	1037	82	16	2.95	regulation of developmental process
43005	8.13×10^{-5}	17415	461	85	10	4.44	neuron projection
44093	8.34×10^{-5}	15685	822	82	14	3.26	positive regulation of molecular function
48878	8.35×10^{-5}	15685	618	82	12	3.71	chemical homeostasis
43025	9.17×10^{-5}	17415	217	85	7	6.61	neuronal cell body
5217	9.70×10^{-5}	17373	18	88	3	32.90	intracellular ligand-gated ion channel activity
267	1.02×10^{-4}	17415	1257	85	17	2.77	cell fraction
42803	1.06×10^{-4}	17373	458	88	10	4.31	protein homodimerization activity
50890	1.06×10^{-4}	15685	144	82	6	7.97	cognition
30659	1.15×10^{-4}	17415	225	85	7	6.37	cytoplasmic vesicle membrane
71822	1.16×10^{-4}	15685	742	82	13	3.35	protein complex subunit organization
43434	1.19×10^{-4}	15685	365	82	9	4.72	response to peptide hormone stimulus
2682	1.20×10^{-4}	15685	851	82	14	3.15	regulation of immune system process
2252	1.24×10^{-4}	15685	367	82	9	4.69	immune effector process
9725	1.31×10^{-4}	15685	648	82	12	3.54	response to hormone stimulus

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
10033	1.33×10^{-4}	15685	1447	82	19	2.51	response to organic substance
43405	1.35×10^{-4}	15685	216	82	7	6.20	regulation of MAP kinase activity
18209	1.38×10^{-4}	15685	95	82	5	10.07	peptidyl-serine modification
44297	1.39×10^{-4}	17415	232	85	7	6.18	cell body
60341	1.46×10^{-4}	15685	463	82	10	4.13	regulation of cellular localization
17081	1.51×10^{-4}	17373	4	88	2	98.71	chloride channel regulator activity
71900	1.52×10^{-4}	15685	295	82	8	5.19	regulation of protein serine/threonine kinase activity
12506	1.54×10^{-4}	17415	236	85	7	6.08	vesicle membrane
7167	1.62×10^{-4}	15685	767	82	13	3.24	enzyme linked receptor protein signaling pathway
48812	1.69×10^{-4}	15685	224	82	7	5.98	neuron projection morphogenesis
46983	1.74×10^{-4}	17373	687	88	12	3.45	protein dimerization activity
43406	1.77×10^{-4}	15685	158	82	6	7.26	positive regulation of MAP kinase activity
50688	1.80×10^{-4}	15685	54	82	4	14.17	regulation of defense response to virus
30424	1.80×10^{-4}	17415	242	85	7	5.93	axon
19904	1.89×10^{-4}	17373	492	88	10	4.01	protein domain specific binding
32270	1.99×10^{-4}	15685	391	82	9	4.40	positive regulation of cellular protein metabolic process
48583	2.11×10^{-4}	15685	1895	82	22	2.22	regulation of response to stimulus
9887	2.19×10^{-4}	15685	487	82	10	3.93	organ morphogenesis
165	2.28×10^{-4}	15685	313	82	8	4.89	MAPKKK cascade
32502	2.34×10^{-4}	15685	3852	82	35	1.74	developmental process
48858	2.39×10^{-4}	15685	237	82	7	5.65	cell projection morphogenesis
2000027	2.42×10^{-4}	15685	107	82	5	8.94	regulation of organ morphogenesis
1763	2.47×10^{-4}	15685	168	82	6	6.83	morphogenesis of a branching structure
45860	2.59×10^{-4}	15685	319	82	8	4.80	positive regulation of protein kinase activity
60501	2.67×10^{-4}	15685	5	82	2	76.51	positive regulation of epithelial cell proliferation involved in lung morphogenesis
2000794	2.67×10^{-4}	15685	5	82	2	76.51	regulation of epithelial cell proliferation involved in lung morphogenesis
60017	2.67×10^{-4}	15685	5	82	2	76.51	parathyroid gland development

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
32990	2.71×10^{-4}	15685	242	82	7	5.53	cell part morphogenesis
32940	2.78×10^{-4}	15685	600	82	11	3.51	secretion by cell
42592	2.80×10^{-4}	15685	923	82	14	2.90	homeostatic process
48538	2.92×10^{-4}	15685	25	82	3	22.95	thymus development
55001	3.07×10^{-4}	15685	62	82	4	12.34	muscle cell development
16247	3.10×10^{-4}	17373	64	88	4	12.34	channel regulator activity
9719	3.24×10^{-4}	15685	715	82	12	3.21	response to endogenous stimulus
51649	3.26×10^{-4}	15685	1551	82	19	2.34	establishment of localization in cell
32436	3.28×10^{-4}	15685	26	82	3	22.07	positive regulation of proteasomal ubiquitin-dependent protein catabolic process
33674	3.32×10^{-4}	15685	331	82	8	4.62	positive regulation of kinase activity
33554	3.45×10^{-4}	15685	942	82	14	2.84	cellular response to stress
22607	3.47×10^{-4}	15685	1303	82	17	2.50	cellular component assembly
32870	3.67×10^{-4}	15685	336	82	8	4.55	cellular response to hormone stimulus
14706	3.81×10^{-4}	15685	118	82	5	8.11	striated muscle tissue development
71902	3.90×10^{-4}	15685	183	82	6	6.27	positive regulation of protein serine/threonine kinase activity
44422	3.91×10^{-4}	17415	5841	85	44	1.54	organelle part
51347	3.97×10^{-4}	15685	340	82	8	4.50	positive regulation of transferase activity
60502	4.00×10^{-4}	15685	6	82	2	63.76	epithelial cell proliferation involved in lung morphogenesis
60078	4.11×10^{-4}	15685	28	82	3	20.49	regulation of postsynaptic membrane potential
46982	4.12×10^{-4}	17373	267	88	7	5.18	protein heterodimerization activity
70887	4.13×10^{-4}	15685	959	82	14	2.79	cellular response to chemical stimulus
51495	4.14×10^{-4}	15685	67	82	4	11.42	positive regulation of cytoskeleton organization
16020	4.31×10^{-4}	17415	7834	85	54	1.41	membrane
7154	4.42×10^{-4}	15685	1722	82	20	2.22	cell communication
42692	4.44×10^{-4}	15685	122	82	5	7.84	muscle cell differentiation
43623	4.48×10^{-4}	15685	263	82	7	5.09	cellular protein complex assembly
35094	4.56×10^{-4}	15685	29	82	3	19.79	response to nicotine
46649	4.74×10^{-4}	15685	440	82	9	3.91	lymphocyte activation
2831	4.89×10^{-4}	15685	70	82	4	10.93	regulation of response to biotic stimulus

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
187	4.96×10^{-4}	15685	125	82	5	7.65	activation of MAPK activity
43254	5.14×10^{-4}	15685	126	82	5	7.59	regulation of protein complex assembly
32501	5.27×10^{-4}	15685	4009	82	35	1.67	multicellular organismal process
71495	5.28×10^{-4}	15685	355	82	8	4.31	cellular response to endogenous stimulus
60090	5.36×10^{-4}	17373	131	88	5	7.54	binding
32879	5.57×10^{-4}	15685	988	82	14	2.71	regulation of localization bridging

Table 4.8: GO terms significantly enriched among *AAP-associated* genes - P, probability; N, total genes in the genome for each GO category (i.e. biological processes, molecular function, and cellular component); G, total genes annotated; n, number of input genes; x, number of annotated genes among the input genes; fold change, fold change over expected by random chance; GO term, term description

GOID	P	N	G	n	x	fold change	GO term
44446	1.37×10^{-13}	17415	5762	418	204	1.48	intracellular organelle part
32502	6.03×10^{-13}	15685	3852	416	187	1.83	developmental process
5737	8.51×10^{-13}	17415	8582	418	274	1.33	cytoplasm
48523	1.19×10^{-12}	15685	2279	416	114	1.89	negative regulation of cellular process
43229	1.49×10^{-12}	17415	10129	418	316	1.30	intracellular organelle
5829	1.60×10^{-12}	17415	2092	418	99	1.97	cytosol
42995	2.22×10^{-12}	17415	976	418	70	2.99	cell projection
7268	2.33×10^{-12}	15685	454	416	50	4.15	synaptic transmission
42592	2.44×10^{-12}	15685	923	416	62	2.53	homeostatic process
5622	2.99×10^{-12}	17415	12324	418	358	1.21	intracellular
43226	3.10×10^{-12}	17415	10167	418	316	1.29	organelle
71842	3.13×10^{-12}	15685	2587	416	137	2.00	cellular component organization at cellular level
7267	3.26×10^{-12}	15685	739	416	61	3.11	cell-cell signaling
48869	3.38×10^{-12}	15685	2069	416	114	2.08	cellular developmental process
48856	4.09×10^{-12}	15685	2336	416	129	2.08	anatomical structure development
23052	4.23×10^{-12}	15685	1999	416	117	2.21	signaling
32501	5.96×10^{-12}	15685	4009	416	174	1.64	multicellular organismal process
23051	6.55×10^{-12}	15685	1544	416	86	2.10	regulation of signaling
48519	6.55×10^{-12}	15685	2489	416	124	1.88	negative regulation of biological process
32991	7.57×10^{-12}	17415	3888	418	156	1.67	macromolecular complex
44424	8.63×10^{-12}	17415	11891	418	355	1.24	intracellular part

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
7610	8.92×10^{-12}	15685	432	416	48	4.19	behavior
50793	9.05×10^{-12}	15685	1037	416	67	2.44	regulation of developmental process
50890	9.99×10^{-12}	15685	144	416	23	6.02	cognition
43025	1.02×10^{-11}	17415	217	418	26	4.99	neuronal cell body
43005	1.04×10^{-11}	17415	461	418	41	3.71	neuron projection
71841	1.05×10^{-11}	15685	2617	416	139	2.00	cellular component organization or biogenesis at cellular level
7154	1.05×10^{-11}	15685	1722	416	94	2.06	cell communication
71840	1.19×10^{-11}	15685	3403	416	168	1.86	cellular component organization or biogenesis
45202	1.75×10^{-11}	17415	438	418	48	4.57	synapse
65007	1.78×10^{-11}	15685	8625	416	294	1.29	biological regulation
51239	1.93×10^{-11}	15685	1381	416	90	2.46	regulation of multicellular organismal process
44422	2.28×10^{-11}	17415	5841	418	210	1.50	organelle part
65008	2.34×10^{-11}	15685	2092	416	122	2.20	regulation of biological quality
16043	2.40×10^{-11}	15685	3372	416	166	1.86	cellular component organization
6996	2.41×10^{-11}	15685	1688	416	100	2.23	organelle organization
44456	2.59×10^{-11}	17415	320	418	38	4.95	synapse part
5515	3.01×10^{-11}	17373	10841	433	341	1.26	protein binding
19904	3.51×10^{-11}	17373	492	433	41	3.34	protein domain specific binding
97060	3.52×10^{-11}	17415	188	418	24	5.32	synaptic membrane
50804	3.86×10^{-11}	15685	163	416	23	5.32	regulation of synaptic transmission
43228	6.37×10^{-11}	17415	2837	418	120	1.76	non-membrane-bounded organelle
43232	6.37×10^{-11}	17415	2837	418	120	1.76	intracellular non-membrane-bounded organelle
44297	6.89×10^{-11}	17415	232	418	26	4.67	cell body
7010	8.90×10^{-11}	15685	601	416	46	2.89	cytoskeleton organization
50789	1.22×10^{-10}	15685	8161	416	280	1.29	regulation of biological process
9653	1.50×10^{-10}	15685	1341	416	76	2.14	anatomical structure morphogenesis
44463	1.56×10^{-10}	17415	464	418	37	3.32	cell projection part
19226	1.98×10^{-10}	15685	188	416	24	4.81	transmission of nerve impulse
35637	1.98×10^{-10}	15685	188	416	24	4.81	multicellular organismal signaling
48731	2.27×10^{-10}	15685	945	416	60	2.39	system development
51969	2.38×10^{-10}	15685	175	416	23	4.96	regulation of transmission of nerve impulse

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
8283	3.17×10^{-10}	15685	1337	416	75	2.12	cell proliferation
45211	3.27×10^{-10}	17415	162	418	21	5.40	postsynaptic membrane
2000026	3.90×10^{-10}	15685	812	416	54	2.51	regulation of multicellular organismal development
50794	5.04×10^{-10}	15685	7745	416	267	1.30	regulation of cellular process
22607	6.13×10^{-10}	15685	1303	416	73	2.11	cellular component assembly
44444	6.78×10^{-10}	17415	6195	418	209	1.41	cytoplasmic part
65009	7.59×10^{-10}	15685	1417	416	77	2.05	regulation of molecular function
51128	7.96×10^{-10}	15685	877	416	56	2.41	regulation of cellular component organization
44057	8.36×10^{-10}	15685	385	416	34	3.33	regulation of system process
50896	8.55×10^{-10}	15685	6717	416	239	1.34	response to stimulus
19222	8.74×10^{-10}	15685	3944	416	160	1.53	regulation of metabolic process
31644	1.06×10^{-9}	15685	188	416	23	4.61	regulation of neurological system process
33043	1.33×10^{-9}	15685	333	416	31	3.51	regulation of organelle organization
32989	1.52×10^{-9}	15685	435	416	36	3.12	cellular component morphogenesis
30154	1.55×10^{-9}	15685	1717	416	87	1.91	cell differentiation
42391	1.73×10^{-9}	15685	117	416	18	5.80	regulation of membrane potential
7611	2.23×10^{-9}	15685	133	416	19	5.39	learning or memory
42221	2.25×10^{-9}	15685	2344	416	108	1.74	response to chemical stimulus
45595	2.33×10^{-9}	15685	733	416	49	2.52	regulation of cell differentiation
30036	2.44×10^{-9}	15685	322	416	30	3.51	actin cytoskeleton organization
30054	2.72×10^{-9}	17415	609	418	41	2.80	cell junction
31399	3.06×10^{-9}	15685	787	416	51	2.44	regulation of protein modification process
43234	4.34×10^{-9}	17415	3229	418	126	1.63	protein complex
48858	4.41×10^{-9}	15685	237	416	25	3.98	cell projection morphogenesis
51179	4.53×10^{-9}	15685	1177	416	66	2.11	localization
30030	5.00×10^{-9}	15685	498	416	38	2.88	cell projection organization
48812	6.63×10^{-9}	15685	224	416	24	4.04	neuron projection morphogenesis
32990	6.77×10^{-9}	15685	242	416	25	3.90	cell part morphogenesis
7015	7.17×10^{-9}	15685	174	416	21	4.55	actin filament organization
30029	8.07×10^{-9}	15685	359	416	31	3.26	actin filament-based process
14069	8.74×10^{-9}	17415	96	418	15	6.51	postsynaptic density
71844	1.05×10^{-8}	15685	967	416	57	2.22	cellular component assembly at cellular level
43227	1.11×10^{-8}	17415	9022	418	273	1.26	membrane-bounded organelle

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
12506	1.35×10^{-8}	17415	236	418	23	4.06	vesicle membrane
21987	1.40×10^{-8}	15685	44	416	11	9.43	cerebral cortex development
48468	1.47×10^{-8}	15685	609	416	42	2.60	cell development
6810	1.48×10^{-8}	15685	3401	416	139	1.54	transport
51234	1.74×10^{-8}	15685	3442	416	140	1.53	establishment of localization
43231	1.83×10^{-8}	17415	9016	418	272	1.26	intracellular membrane-bounded organelle
267	1.83×10^{-8}	17415	1257	418	63	2.09	cell fraction
5488	2.06×10^{-8}	17373	14540	433	401	1.11	binding
65003	2.09×10^{-8}	15685	884	416	53	2.26	macromolecular complex assembly
30659	2.63×10^{-8}	17415	225	418	22	4.07	cytoplasmic vesicle membrane
19717	3.26×10^{-8}	17415	121	418	16	5.51	synaptosome
122	3.55×10^{-8}	15685	425	416	33	2.93	negative regulation of transcription from RNA polymerase II promoter
51649	4.09×10^{-8}	15685	1551	416	77	1.87	establishment of localization in cell
19725	4.20×10^{-8}	15685	539	416	38	2.66	cellular homeostasis
22603	4.83×10^{-8}	15685	409	416	32	2.95	regulation of anatomical structure morphogenesis
48518	5.14×10^{-8}	15685	2824	416	119	1.59	positive regulation of biological process
7612	5.42×10^{-8}	15685	73	416	13	6.71	learning
43933	5.82×10^{-8}	15685	1042	416	58	2.10	macromolecular complex subunit organization
7399	5.99×10^{-8}	15685	640	416	42	2.47	nervous system development
42127	7.75×10^{-8}	15685	998	416	56	2.12	regulation of cell proliferation
5626	7.94×10^{-8}	17415	959	418	51	2.22	insoluble fraction
8104	8.94×10^{-8}	15685	420	416	32	2.87	protein localization
51641	9.48×10^{-8}	15685	603	416	40	2.50	cellular localization
30424	9.72×10^{-8}	17415	242	418	22	3.79	axon
31982	1.05×10^{-7}	17415	801	418	45	2.34	vesicle
48513	1.09×10^{-7}	15685	982	416	55	2.11	organ development
35176	1.16×10^{-7}	15685	24	416	8	12.57	social behavior
55082	1.23×10^{-7}	15685	448	416	33	2.78	cellular chemical homeostasis
5856	1.39×10^{-7}	17415	1574	418	71	1.88	cytoskeleton
60249	1.43×10^{-7}	15685	171	416	19	4.19	anatomical structure homeostasis
6873	1.53×10^{-7}	15685	430	416	32	2.81	cellular ion homeostasis
30425	1.66×10^{-7}	17415	190	418	19	4.17	dendrite
51899	1.68×10^{-7}	15685	55	416	11	7.54	membrane depolarization
33036	1.96×10^{-7}	15685	457	416	33	2.72	macromolecule localization
902	2.09×10^{-7}	15685	308	416	26	3.18	cell morphogenesis
7409	2.26×10^{-7}	15685	176	416	19	4.07	axonogenesis

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
51704	2.53×10^{-7}	15685	825	416	48	2.19	multi-organism process
48522	2.56×10^{-7}	15685	2527	416	107	1.60	positive regulation of cellular process
48167	2.74×10^{-7}	15685	70	416	12	6.46	regulation of synaptic plasticity
31410	2.89×10^{-7}	17415	775	418	43	2.31	cytoplasmic vesicle
51493	2.95×10^{-7}	15685	179	416	19	4.00	regulation of cytoskeleton organization
43405	3.12×10^{-7}	15685	216	416	21	3.67	regulation of MAP kinase activity
3008	3.24×10^{-7}	15685	1544	416	74	1.81	system process
44433	3.80×10^{-7}	17415	305	418	24	3.28	cytoplasmic vesicle part
9987	3.89×10^{-7}	15685	13399	416	388	1.09	cellular process
42802	3.95×10^{-7}	17373	808	433	45	2.23	identical protein binding
5624	4.07×10^{-7}	17415	925	418	48	2.16	membrane fraction
46903	4.31×10^{-7}	15685	713	416	43	2.27	secretion
50801	4.73×10^{-7}	15685	475	416	33	2.62	ion homeostasis
32268	4.80×10^{-7}	15685	1028	416	55	2.02	regulation of cellular protein metabolic process
48878	4.97×10^{-7}	15685	618	416	39	2.38	chemical homeostasis
60255	5.22×10^{-7}	15685	3335	416	131	1.48	regulation of macromolecule metabolic process
9790	5.38×10^{-7}	15685	366	416	28	2.88	embryo development
6468	5.52×10^{-7}	15685	1114	416	58	1.96	protein phosphorylation
42325	5.58×10^{-7}	15685	670	416	41	2.31	regulation of phosphorylation
50790	5.78×10^{-7}	15685	1143	416	59	1.95	regulation of catalytic activity
51049	5.98×10^{-7}	15685	747	416	44	2.22	regulation of transport
9725	6.21×10^{-7}	15685	648	416	40	2.33	response to hormone stimulus
5634	6.23×10^{-7}	17415	5650	418	183	1.35	nucleus
32940	6.43×10^{-7}	15685	600	416	38	2.39	secretion by cell
1932	6.87×10^{-7}	15685	626	416	39	2.35	regulation of protein phosphorylation
10646	7.05×10^{-7}	15685	1095	416	57	1.96	regulation of cell communication
51246	7.70×10^{-7}	15685	1153	416	59	1.93	regulation of protein metabolic process
51705	7.85×10^{-7}	15685	40	416	9	8.48	behavioral interaction between organisms
9887	8.28×10^{-7}	15685	487	416	33	2.55	organ morphogenesis
7275	8.67×10^{-7}	15685	1584	416	74	1.76	multicellular organismal development
9893	8.76×10^{-7}	15685	1241	416	62	1.88	positive regulation of metabolic process
31323	9.37×10^{-7}	15685	3568	416	137	1.45	regulation of cellular metabolic process

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
48589	9.44×10^{-7}	15685	193	416	19	3.71	developmental growth
80090	9.71×10^{-7}	15685	3537	416	136	1.45	regulation of primary metabolic process
8092	1.01×10^{-6}	17373	572	433	35	2.46	cytoskeletal protein binding
43406	1.03×10^{-6}	15685	158	416	17	4.06	positive regulation of MAP kinase activity
44464	1.06×10^{-6}	17415	16105	418	409	1.06	cell part
5623	1.08×10^{-6}	17415	16106	418	409	1.06	cell
44087	1.12×10^{-6}	15685	214	416	20	3.52	regulation of cellular component biogenesis
9719	1.18×10^{-6}	15685	715	416	42	2.21	response to endogenous stimulus
19220	1.19×10^{-6}	15685	690	416	41	2.24	regulation of phosphate metabolic process
51174	1.19×10^{-6}	15685	690	416	41	2.24	regulation of phosphorus metabolic process
34621	1.29×10^{-6}	15685	667	416	40	2.26	cellular macromolecular complex subunit organization
786	1.37×10^{-6}	17415	105	418	13	5.16	nucleosome
16310	1.43×10^{-6}	15685	1203	416	60	1.88	phosphorylation
40007	1.50×10^{-6}	15685	646	416	39	2.28	growth
5886	1.52×10^{-6}	17415	4157	418	142	1.42	plasma membrane
50877	1.63×10^{-6}	15685	1152	416	58	1.90	neurological system process
31175	1.98×10^{-6}	15685	222	416	20	3.40	neuron projection development
31325	2.14×10^{-6}	15685	1162	416	58	1.88	positive regulation of cellular metabolic process
7632	2.27×10^{-6}	15685	34	416	8	8.87	visual behavior
48011	2.28×10^{-6}	15685	224	416	20	3.37	nerve growth factor receptor signaling pathway
60341	2.31×10^{-6}	15685	463	416	31	2.52	regulation of cellular localization
1894	2.37×10^{-6}	15685	85	416	12	5.32	tissue homeostasis
9605	2.43×10^{-6}	15685	1111	416	56	1.90	response to external stimulus
44085	2.58×10^{-6}	15685	287	416	23	3.02	cellular component biogenesis
33267	2.60×10^{-6}	17415	111	418	13	4.88	axon part
19899	2.84×10^{-6}	17373	786	433	42	2.14	enzyme binding
71822	3.03×10^{-6}	15685	742	416	42	2.13	protein complex subunit organization
43549	3.05×10^{-6}	15685	493	416	32	2.45	regulation of kinase activity
34622	3.32×10^{-6}	15685	519	416	33	2.40	cellular macromolecular complex assembly
7169	3.36×10^{-6}	15685	568	416	35	2.32	transmembrane receptor protein tyrosine kinase signaling pathway
6793	3.38×10^{-6}	15685	1379	416	65	1.78	phosphorus metabolic process

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
6796	3.38×10^{-6}	15685	1379	416	65	1.78	phosphate metabolic process
8285	3.44×10^{-6}	15685	472	416	31	2.48	negative regulation of cell proliferation
10604	3.65×10^{-6}	15685	1126	416	56	1.88	positive regulation of macromolecule metabolic process
6461	3.67×10^{-6}	15685	620	416	37	2.25	protein complex assembly
32879	3.68×10^{-6}	15685	988	416	51	1.95	regulation of localization
45859	3.91×10^{-6}	15685	475	416	31	2.46	regulation of protein kinase activity
32868	4.08×10^{-6}	15685	253	416	21	3.13	response to insulin stimulus
71900	4.10×10^{-6}	15685	295	416	23	2.94	regulation of protein serine/threonine kinase activity
30182	4.28×10^{-6}	15685	384	416	27	2.65	neuron differentiation
9966	4.58×10^{-6}	15685	1305	416	62	1.79	regulation of signal transduction
51130	4.59×10^{-6}	15685	297	416	23	2.92	positive regulation of cellular component organization
44428	4.83×10^{-6}	17415	2268	418	87	1.60	nuclear part
6334	5.11×10^{-6}	15685	141	416	15	4.01	nucleosome assembly
30534	5.52×10^{-6}	15685	92	416	12	4.92	adult behavior
6325	5.62×10^{-6}	15685	532	416	33	2.34	chromatin organization
51338	5.70×10^{-6}	15685	508	416	32	2.38	regulation of transferase activity
51129	5.74×10^{-6}	15685	238	416	20	3.17	negative regulation of cellular component organization
60078	6.42×10^{-6}	15685	28	416	7	9.43	regulation of postsynaptic membrane potential
22008	6.76×10^{-6}	15685	304	416	23	2.85	neurogenesis
60537	6.95×10^{-6}	15685	127	416	14	4.16	muscle tissue development
90066	7.00×10^{-6}	15685	201	416	18	3.38	regulation of anatomical structure size
16044	7.05×10^{-6}	15685	465	416	30	2.43	cellular membrane organization
61024	7.67×10^{-6}	15685	467	416	30	2.42	membrane organization
71902	7.73×10^{-6}	15685	183	416	17	3.50	positive regulation of protein serine/threonine kinase activity
5938	7.99×10^{-6}	17415	161	418	15	3.88	cell cortex
904	8.03×10^{-6}	15685	203	416	18	3.34	cell morphogenesis involved in differentiation
6816	8.03×10^{-6}	15685	203	416	18	3.34	calcium ion transport
10638	8.34×10^{-6}	15685	129	416	14	4.09	positive regulation of organelle organization
10033	8.37×10^{-6}	15685	1447	416	66	1.72	response to organic substance
1505	8.38×10^{-6}	15685	112	416	13	4.38	regulation of neurotransmitter levels

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
51960	8.38×10^{-6}	15685	308	416	23	2.82	regulation of nervous system development
16192	8.47×10^{-6}	15685	800	416	43	2.03	vesicle-mediated transport
32993	8.95×10^{-6}	17415	124	418	13	4.37	protein-DNA complex
71495	9.48×10^{-6}	15685	355	416	25	2.66	cellular response to endogenous stimulus
30900	9.94×10^{-6}	15685	67	416	10	5.63	forebrain development
44430	1.02×10^{-5}	17415	1100	418	50	1.89	cytoskeletal part
8542	1.05×10^{-5}	15685	30	416	7	8.80	visual learning
43085	1.11×10^{-5}	15685	703	416	39	2.09	positive regulation of catalytic activity
32870	1.13×10^{-5}	15685	336	416	24	2.69	cellular response to hormone stimulus
31988	1.15×10^{-5}	17415	526	418	30	2.38	membrane-bounded vesicle
45664	1.18×10^{-5}	15685	229	416	19	3.13	regulation of neuron differentiation
34728	1.18×10^{-5}	15685	151	416	15	3.75	nucleosome organization
32956	1.23×10^{-5}	15685	116	416	13	4.23	regulation of actin cytoskeleton organization
43412	1.26×10^{-5}	15685	2357	416	95	1.52	macromolecule modification
32271	1.31×10^{-5}	15685	84	416	11	4.94	regulation of protein polymerization
32947	1.39×10^{-5}	17373	33	433	7	8.51	protein complex scaffold
8306	1.47×10^{-5}	15685	43	416	8	7.01	associative learning
10467	1.47×10^{-5}	15685	2817	416	109	1.46	gene expression
6464	1.47×10^{-5}	15685	2271	416	92	1.53	protein modification process
14706	1.49×10^{-5}	15685	118	416	13	4.15	striated muscle tissue development
60284	1.51×10^{-5}	15685	342	416	24	2.65	regulation of cell development
43434	1.52×10^{-5}	15685	365	416	25	2.58	response to peptide hormone stimulus
50767	1.56×10^{-5}	15685	276	416	21	2.87	regulation of neurogenesis
44093	1.64×10^{-5}	15685	822	416	43	1.97	positive regulation of molecular function
48169	1.75×10^{-5}	15685	22	416	6	10.28	regulation of long-term neuronal synaptic plasticity
32970	1.78×10^{-5}	15685	120	416	13	4.08	regulation of actin filament-based process
48638	1.91×10^{-5}	15685	72	416	10	5.24	regulation of developmental growth
65004	2.04×10^{-5}	15685	158	416	15	3.58	protein-DNA complex assembly
16023	2.05×10^{-5}	17415	515	418	29	2.35	cytoplasmic membrane-bounded vesicle
15276	2.05×10^{-5}	17373	129	433	13	4.04	ligand-gated ion channel activity

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
22834	2.05×10^{-5}	17373	129	433	13	4.04	ligand-gated channel activity
10975	2.14×10^{-5}	15685	140	416	14	3.77	regulation of neuron projection development
9791	2.16×10^{-5}	15685	73	416	10	5.17	post-embryonic development
7420	2.31×10^{-5}	15685	179	416	16	3.37	brain development
7411	2.32×10^{-5}	15685	351	416	24	2.58	axon guidance
51093	2.40×10^{-5}	15685	375	416	25	2.51	negative regulation of developmental process
34702	2.42×10^{-5}	17415	220	418	17	3.22	ion channel complex
33555	2.47×10^{-5}	15685	46	416	8	6.56	multicellular organismal response to stress
3006	2.54×10^{-5}	15685	330	416	23	2.63	developmental process involved in reproduction
35556	2.58×10^{-5}	15685	1680	416	72	1.62	intracellular signal transduction
48646	2.62×10^{-5}	15685	548	416	32	2.20	anatomical structure formation involved in morphogenesis
60548	2.73×10^{-5}	15685	524	416	31	2.23	negative regulation of cell death
187	2.77×10^{-5}	15685	125	416	13	3.92	activation of MAPK activity
19900	2.85×10^{-5}	17373	259	433	19	2.94	kinase binding
43254	3.01×10^{-5}	15685	126	416	13	3.89	regulation of protein complex assembly
71824	3.16×10^{-5}	15685	164	416	15	3.45	protein-DNA complex subunit organization
22604	3.20×10^{-5}	15685	204	416	17	3.14	regulation of cell morphogenesis
165	3.34×10^{-5}	15685	313	416	22	2.65	MAPKKK cascade
22414	3.39×10^{-5}	15685	1129	416	53	1.77	reproductive process
31344	3.40×10^{-5}	15685	165	416	15	3.43	regulation of cell projection organization
9952	3.40×10^{-5}	15685	165	416	15	3.43	anterior/posterior pattern formation
9628	3.41×10^{-5}	15685	505	416	30	2.24	response to abiotic stimulus
50770	3.47×10^{-5}	15685	77	416	10	4.90	regulation of axonogenesis
7167	3.52×10^{-5}	15685	767	416	40	1.97	enzyme linked receptor protein signaling pathway
51090	3.79×10^{-5}	15685	249	416	19	2.88	regulation of sequence-specific DNA binding transcription factor activity
40008	4.00×10^{-5}	15685	435	416	27	2.34	regulation of growth
43069	4.10×10^{-5}	15685	510	416	30	2.22	negative regulation of programmed cell death
10468	4.14×10^{-5}	15685	2525	416	98	1.46	regulation of gene expression
30135	4.41×10^{-5}	17415	124	418	12	4.03	coated vesicle

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
51046	4.43×10^{-5}	15685	342	416	23	2.54	regulation of secretion
45860	4.45×10^{-5}	15685	319	416	22	2.60	positive regulation of protein kinase activity
9416	4.49×10^{-5}	15685	169	416	15	3.35	response to light stimulus
9892	4.68×10^{-5}	15685	887	416	44	1.87	negative regulation of metabolic process
51716	4.77×10^{-5}	15685	5092	416	173	1.28	cellular response to stimulus
42133	4.93×10^{-5}	15685	26	416	6	8.70	neurotransmitter metabolic process
71375	4.97×10^{-5}	15685	254	416	19	2.82	cellular response to peptide hormone stimulus
23061	5.09×10^{-5}	15685	299	416	21	2.65	signal release
1666	5.10×10^{-5}	15685	191	416	16	3.16	response to hypoxia
8154	5.14×10^{-5}	15685	97	416	11	4.28	actin polymerization or depolymerization
10628	5.22×10^{-5}	15685	620	416	34	2.07	positive regulation of gene expression
50878	5.59×10^{-5}	15685	570	416	32	2.12	regulation of body fluid levels
60070	5.76×10^{-5}	15685	134	416	13	3.66	canonical Wnt receptor signaling pathway
10769	5.76×10^{-5}	15685	134	416	13	3.66	regulation of cell morphogenesis involved in differentiation
5654	6.49×10^{-5}	17415	1407	418	57	1.69	nucleoplasm
45893	6.57×10^{-5}	15685	872	416	43	1.86	positive regulation of transcription DNA-dependent
7596	6.59×10^{-5}	15685	498	416	29	2.20	blood coagulation
3682	6.79×10^{-5}	17373	230	433	17	2.97	chromatin binding
35264	6.81×10^{-5}	15685	100	416	11	4.15	multicellular organism growth
51173	7.08×10^{-5}	15685	737	416	38	1.94	positive regulation of nitrogen compound metabolic process
51480	7.11×10^{-5}	15685	156	416	14	3.38	cytosolic calcium ion homeostasis
5230	7.24×10^{-5}	17373	72	433	9	5.02	extracellular ligand-gated ion channel activity
8291	7.26×10^{-5}	15685	4	416	3	28.28	acetylcholine metabolic process
42471	7.26×10^{-5}	15685	4	416	3	28.28	ear morphogenesis
50817	7.32×10^{-5}	15685	501	416	29	2.18	coagulation
7519	7.41×10^{-5}	15685	84	416	10	4.49	skeletal muscle tissue development
7599	7.59×10^{-5}	15685	502	416	29	2.18	hemostasis
33674	7.70×10^{-5}	15685	331	416	22	2.51	positive regulation of kinase activity
48583	7.91×10^{-5}	15685	1895	416	77	1.53	regulation of response to stimulus

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
32535	8.17×10^{-5}	15685	102	416	11	4.07	regulation of cellular component size
60560	8.20×10^{-5}	15685	85	416	10	4.44	developmental growth involved in morphogenesis
43029	8.25×10^{-5}	15685	18	416	5	10.47	T cell homeostasis
5694	8.42×10^{-5}	17415	614	418	31	2.10	chromosome
70838	8.86×10^{-5}	15685	243	416	18	2.79	divalent metal ion transport
8219	9.14×10^{-5}	15685	1593	416	67	1.59	cell death
60402	9.35×10^{-5}	15685	55	416	8	5.48	calcium ion transport into cytosol
30136	9.37×10^{-5}	17415	114	418	11	4.02	clathrin-coated vesicle
44419	9.47×10^{-5}	15685	408	416	25	2.31	interspecies interaction between organisms
6913	9.47×10^{-5}	15685	408	416	25	2.31	nucleocytoplasmic transport
16265	9.49×10^{-5}	15685	1595	416	67	1.58	death
2260	9.50×10^{-5}	15685	29	416	6	7.80	lymphocyte homeostasis
44306	9.84×10^{-5}	17415	32	418	6	7.81	neuron projection terminus
51169	1.02×10^{-4}	15685	410	416	25	2.30	nuclear transport
51276	1.04×10^{-4}	15685	669	416	35	1.97	chromosome organization
44092	1.06×10^{-4}	15685	486	416	28	2.17	negative regulation of molecular function
48168	1.06×10^{-4}	15685	42	416	7	6.28	regulation of neuronal synaptic plasticity
60401	1.07×10^{-4}	15685	56	416	8	5.39	cytosolic calcium ion transport
72511	1.09×10^{-4}	15685	247	416	18	2.75	divalent inorganic cation transport
32869	1.12×10^{-4}	15685	204	416	16	2.96	cellular response to insulin stimulus
45596	1.12×10^{-4}	15685	316	416	21	2.51	negative regulation of cell differentiation
51347	1.14×10^{-4}	15685	340	416	22	2.44	positive regulation of transferase activity
48675	1.21×10^{-4}	15685	57	416	8	5.29	axon extension
2000027	1.26×10^{-4}	15685	107	416	11	3.88	regulation of organ morphogenesis
6836	1.32×10^{-4}	15685	126	416	12	3.59	neurotransmitter transport
70482	1.32×10^{-4}	15685	207	416	16	2.91	response to oxygen levels
7204	1.39×10^{-4}	15685	146	416	13	3.36	elevation of cytosolic calcium ion concentration
9889	1.39×10^{-4}	15685	2671	416	100	1.41	regulation of biosynthetic process
70887	1.40×10^{-4}	15685	959	416	45	1.77	cellular response to chemical stimulus
30001	1.41×10^{-4}	15685	599	416	32	2.01	metal ion transport
50771	1.41×10^{-4}	15685	31	416	6	7.30	negative regulation of axonogenesis

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
31056	1.41×10^{-4}	15685	31	416	6	7.30	regulation of histone modification
1776	1.41×10^{-4}	15685	31	416	6	7.30	leukocyte homeostasis
45652	1.43×10^{-4}	15685	20	416	5	9.43	regulation of megakaryocyte differentiation
31326	1.45×10^{-4}	15685	2641	416	99	1.41	regulation of cellular biosynthetic process
7165	1.46×10^{-4}	15685	4481	416	153	1.29	signal transduction
40011	1.48×10^{-4}	15685	1019	416	47	1.74	locomotion
30165	1.50×10^{-4}	17373	79	433	9	4.57	PDZ domain binding
6357	1.52×10^{-4}	15685	877	416	42	1.81	regulation of transcription from RNA polymerase II promoter
6355	1.54×10^{-4}	15685	2777	416	103	1.40	regulation of transcription DNA-dependent
16050	1.59×10^{-4}	15685	75	416	9	4.52	vesicle organization
3	1.62×10^{-4}	15685	577	416	31	2.03	reproduction
44237	1.63×10^{-4}	15685	8126	416	252	1.17	cellular metabolic process
44427	1.63×10^{-4}	17415	522	418	27	2.16	chromosomal part
22836	1.64×10^{-4}	17373	320	433	20	2.51	gated channel activity
8152	1.66×10^{-4}	15685	9213	416	280	1.15	metabolic process
6935	1.74×10^{-4}	15685	553	416	30	2.05	chemotaxis
42330	1.74×10^{-4}	15685	553	416	30	2.05	taxis
45944	1.76×10^{-4}	15685	527	416	29	2.07	positive regulation of transcription from RNA polymerase II promoter
60501	1.78×10^{-4}	15685	5	416	3	22.62	positive regulation of epithelial cell proliferation involved in lung morphogenesis
2000794	1.78×10^{-4}	15685	5	416	3	22.62	regulation of epithelial cell proliferation involved in lung morphogenesis
2000757	1.78×10^{-4}	15685	5	416	3	22.62	negative regulation of peptidyl-lysine acetylation
35067	1.78×10^{-4}	15685	5	416	3	22.62	negative regulation of histone acetylation
16247	1.82×10^{-4}	17373	64	433	8	5.02	channel regulator activity
43066	1.88×10^{-4}	15685	503	416	28	2.10	negative regulation of apoptosis
7568	1.90×10^{-4}	15685	171	416	14	3.09	aging
30041	1.95×10^{-4}	15685	77	416	9	4.41	actin filament polymerization
8022	2.02×10^{-4}	17373	161	433	13	3.24	protein C-terminus binding
55001	2.21×10^{-4}	15685	62	416	8	4.87	muscle cell development
51171	2.22×10^{-4}	15685	2704	416	100	1.39	regulation of nitrogen compound metabolic process
10941	2.36×10^{-4}	15685	1158	416	51	1.66	regulation of cell death

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
10605	2.37×10^{-4}	15685	810	416	39	1.82	negative regulation of macromolecule metabolic process
149	2.39×10^{-4}	17373	36	433	6	6.69	SNARE binding
48588	2.47×10^{-4}	15685	63	416	8	4.79	developmental cell growth
16055	2.48×10^{-4}	15685	241	416	17	2.66	Wnt receptor signaling pathway
51924	2.59×10^{-4}	15685	116	416	11	3.58	regulation of calcium ion transport
3002	2.60×10^{-4}	15685	242	416	17	2.65	regionalization
31401	2.61×10^{-4}	15685	312	416	20	2.42	positive regulation of protein modification process
9314	2.62×10^{-4}	15685	265	416	18	2.56	response to radiation
50678	2.68×10^{-4}	15685	156	416	13	3.14	regulation of epithelial cell proliferation
51258	2.71×10^{-4}	15685	136	416	12	3.33	protein polymerization
50680	2.76×10^{-4}	15685	64	416	8	4.71	negative regulation of epithelial cell proliferation
6874	2.80×10^{-4}	15685	221	416	16	2.73	cellular calcium ion homeostasis
51588	2.84×10^{-4}	15685	35	416	6	6.46	regulation of neurotransmitter transport
32880	2.87×10^{-4}	15685	244	416	17	2.63	regulation of protein localization
48771	2.87×10^{-4}	15685	81	416	9	4.19	tissue remodeling
35065	2.88×10^{-4}	15685	13	416	4	11.60	regulation of histone acetylation
2000756	2.88×10^{-4}	15685	13	416	4	11.60	regulation of peptidyl-lysine acetylation
48745	2.88×10^{-4}	15685	13	416	4	11.60	smooth muscle tissue development
10959	2.90×10^{-4}	15685	137	416	12	3.30	regulation of metal ion transport
72384	2.90×10^{-4}	15685	23	416	5	8.20	organelle transport along microtubule
30219	2.90×10^{-4}	15685	23	416	5	8.20	megakaryocyte differentiation
7173	2.94×10^{-4}	15685	99	416	10	3.81	epidermal growth factor receptor signaling pathway
10557	2.97×10^{-4}	15685	707	416	35	1.87	positive regulation of macromolecule biosynthetic process
10741	3.08×10^{-4}	15685	65	416	8	4.64	negative regulation of intracellular protein kinase cascade
226	3.12×10^{-4}	15685	201	416	15	2.81	microtubule cytoskeleton organization
44267	3.13×10^{-4}	15685	3064	416	110	1.35	cellular protein metabolic process

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
45935	3.22×10^{-4}	15685	710	416	35	1.86	positive regulation of nucleobase nucleoside nucleotide and nucleic acid metabolic process
7569	3.27×10^{-4}	15685	50	416	7	5.28	cell aging
43170	3.31×10^{-4}	15685	6690	416	212	1.19	macromolecule metabolic process
50772	3.33×10^{-4}	15685	36	416	6	6.28	positive regulation of axonogenesis
32774	3.35×10^{-4}	15685	2470	416	92	1.40	RNA biosynthetic process
30833	3.42×10^{-4}	15685	66	416	8	4.57	regulation of actin filament polymerization
5794	3.43×10^{-4}	17415	1038	418	43	1.73	Golgi apparatus
51607	3.45×10^{-4}	15685	83	416	9	4.09	defense response to virus
60502	3.49×10^{-4}	15685	6	416	3	18.85	epithelial cell proliferation involved in lung morphogenesis
9888	3.58×10^{-4}	15685	604	416	31	1.94	tissue development
60079	3.59×10^{-4}	15685	24	416	5	7.86	regulation of excitatory postsynaptic membrane potential
55074	3.60×10^{-4}	15685	226	416	16	2.67	calcium ion homeostasis
8286	3.65×10^{-4}	15685	161	416	13	3.04	insulin receptor signaling pathway
51223	3.65×10^{-4}	15685	204	416	15	2.77	regulation of protein transport
9891	3.67×10^{-4}	15685	828	416	39	1.78	positive regulation of biosynthetic process
19901	3.70×10^{-4}	17373	217	433	15	2.77	protein kinase binding
31345	3.71×10^{-4}	15685	51	416	7	5.18	negative regulation of cell projection organization
6811	3.71×10^{-4}	15685	973	416	44	1.71	ion transport
51495	3.80×10^{-4}	15685	67	416	8	4.50	positive regulation of cytoskeleton organization
6812	3.84×10^{-4}	15685	689	416	34	1.86	cation transport
7389	3.88×10^{-4}	15685	371	416	22	2.24	pattern specification process
42136	3.95×10^{-4}	15685	14	416	4	10.77	neurotransmitter biosynthetic process
45121	3.96×10^{-4}	17415	179	418	13	3.03	membrane raft
72503	3.96×10^{-4}	15685	228	416	16	2.65	cellular divalent inorganic cation homeostasis
6936	3.96×10^{-4}	15685	228	416	16	2.65	muscle contraction
46982	4.04×10^{-4}	17373	267	433	17	2.55	protein heterodimerization activity
8134	4.06×10^{-4}	17373	343	433	20	2.34	transcription factor binding
8284	4.12×10^{-4}	15685	528	416	28	2.00	positive regulation of cell proliferation
9058	4.12×10^{-4}	15685	4222	416	143	1.28	biosynthetic process

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
44249	4.16×10^{-4}	15685	4083	416	139	1.28	cellular biosynthetic process
4707	4.18×10^{-4}	17373	15	433	4	10.70	MAP kinase activity
50673	4.26×10^{-4}	15685	185	416	14	2.85	epithelial cell proliferation
71310	4.44×10^{-4}	15685	779	416	37	1.79	cellular response to organic substance
8290	4.47×10^{-4}	17415	7	418	3	17.86	F-actin capping protein complex
16564	4.50×10^{-4}	17373	245	433	16	2.62	transcription repressor activity
61387	4.51×10^{-4}	15685	38	416	6	5.95	regulation of extent of cell growth
10522	4.51×10^{-4}	15685	38	416	6	5.95	regulation of calcium ion transport into cytosol
42439	4.51×10^{-4}	15685	38	416	6	5.95	ethanolamine-containing compound metabolic process
2000112	4.57×10^{-4}	15685	2361	416	88	1.41	regulation of cellular macromolecule biosynthetic process
6887	4.58×10^{-4}	15685	231	416	16	2.61	exocytosis
31328	4.60×10^{-4}	15685	809	416	38	1.77	positive regulation of cellular biosynthetic process
51254	4.61×10^{-4}	15685	613	416	31	1.91	positive regulation of RNA metabolic process
51094	4.65×10^{-4}	15685	479	416	26	2.05	positive regulation of developmental process
50768	4.65×10^{-4}	15685	69	416	8	4.37	negative regulation of neurogenesis
46907	4.66×10^{-4}	15685	1013	416	45	1.67	intracellular transport
7613	4.72×10^{-4}	15685	53	416	7	4.98	memory
34330	4.89×10^{-4}	15685	166	416	13	2.95	cell junction organization
16568	4.93×10^{-4}	15685	403	416	23	2.15	chromatin modification
72507	5.28×10^{-4}	15685	234	416	16	2.58	divalent inorganic cation homeostasis
45653	5.28×10^{-4}	15685	15	416	4	10.05	negative regulation of megakaryocyte differentiation
7507	5.28×10^{-4}	15685	189	416	14	2.79	heart development
43523	5.29×10^{-4}	15685	126	416	11	3.29	regulation of neuron apoptosis
50688	5.30×10^{-4}	15685	54	416	7	4.89	regulation of defense response to virus
50769	5.35×10^{-4}	15685	88	416	9	3.86	positive regulation of neurogenesis
15031	5.46×10^{-4}	15685	1140	416	49	1.62	protein transport
32386	5.51×10^{-4}	15685	147	416	12	3.08	regulation of intracellular transport
43524	5.65×10^{-4}	15685	71	416	8	4.25	negative regulation of neuron apoptosis

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
1568	5.65×10^{-4}	15685	71	416	8	4.25	blood vessel development
10639	5.66×10^{-4}	15685	127	416	11	3.27	negative regulation of organelle organization
5642	5.75×10^{-4}	17415	2	418	2	41.66	annulate lamellae
45597	5.81×10^{-4}	15685	357	416	21	2.22	positive regulation of cell differentiation
45927	5.92×10^{-4}	15685	108	416	10	3.49	positive regulation of growth
3007	5.94×10^{-4}	15685	55	416	7	4.80	heart morphogenesis
71109	5.99×10^{-4}	15685	7	416	3	16.16	superior temporal gyrus development
6338	6.31×10^{-4}	15685	90	416	9	3.77	chromatin remodeling
70201	6.34×10^{-4}	15685	215	416	15	2.63	regulation of establishment of protein localization
40018	6.39×10^{-4}	15685	27	416	5	6.98	positive regulation of multicellular organism growth
48041	6.39×10^{-4}	15685	27	416	5	6.98	focal adhesion assembly
19219	6.53×10^{-4}	15685	2619	416	95	1.37	regulation of nucleobase nucleoside nucleotide and nucleic acid metabolic process
4674	6.55×10^{-4}	17373	578	433	28	1.94	protein serine/threonine kinase activity
43623	6.77×10^{-4}	15685	263	416	17	2.44	cellular protein complex assembly
35567	6.88×10^{-4}	15685	41	416	6	5.52	non-canonical Wnt receptor signaling pathway
7626	6.89×10^{-4}	15685	130	416	11	3.19	locomotory behavior
72172	7.02×10^{-4}	15685	2	416	2	37.70	mesonephric tubule formation
72181	7.02×10^{-4}	15685	2	416	2	37.70	mesonephric duct formation
70836	7.02×10^{-4}	15685	2	416	2	37.70	caveola assembly
46671	7.02×10^{-4}	15685	2	416	2	37.70	negative regulation of retinal cell programmed cell death
1765	7.02×10^{-4}	15685	2	416	2	37.70	membrane raft assembly
61038	7.02×10^{-4}	15685	2	416	2	37.70	uterus morphogenesis
7166	7.11×10^{-4}	15685	2659	416	96	1.36	cell surface receptor linked signaling pathway
31324	7.11×10^{-4}	15685	799	416	37	1.75	negative regulation of cellular metabolic process
45637	7.34×10^{-4}	15685	111	416	10	3.40	regulation of myeloid cell differentiation
51656	7.39×10^{-4}	15685	57	416	7	4.63	establishment of organelle localization
8064	7.47×10^{-4}	15685	74	416	8	4.08	regulation of actin polymerization or depolymerization
6611	7.61×10^{-4}	15685	28	416	5	6.73	protein export from nucleus
44448	7.79×10^{-4}	17415	82	418	8	4.06	cell cortex part

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
32270	7.84×10^{-4}	15685	391	416	22	2.12	positive regulation of cellular protein metabolic process
45184	7.85×10^{-4}	15685	1159	416	49	1.59	establishment of protein localization
10720	7.88×10^{-4}	15685	112	416	10	3.37	positive regulation of cell development
51259	7.93×10^{-4}	15685	243	416	16	2.48	protein oligomerization
44260	7.94×10^{-4}	15685	6044	416	192	1.20	cellular macromolecule metabolic process
12501	8.00×10^{-4}	15685	1468	416	59	1.52	programmed cell death
4672	8.00×10^{-4}	17373	644	433	30	1.87	protein kinase activity
19905	8.00×10^{-4}	17373	30	433	5	6.69	syntaxin binding
10556	8.06×10^{-4}	15685	2436	416	89	1.38	regulation of macromolecule biosynthetic process
40014	8.22×10^{-4}	15685	58	416	7	4.55	regulation of multicellular organism growth
3012	8.66×10^{-4}	15685	245	416	16	2.46	muscle system process
15075	8.69×10^{-4}	17373	767	433	34	1.78	ion transmembrane transporter activity
71514	8.82×10^{-4}	15685	17	416	4	8.87	genetic imprinting
5217	8.82×10^{-4}	17373	18	433	4	8.92	intracellular ligand-gated ion channel activity
4890	8.82×10^{-4}	17373	18	433	4	8.92	GABA-A receptor activity
1952	8.92×10^{-4}	15685	43	416	6	5.26	regulation of cell-matrix adhesion
30832	8.93×10^{-4}	15685	76	416	8	3.97	regulation of actin filament length
35094	8.99×10^{-4}	15685	29	416	5	6.50	response to nicotine
51017	9.12×10^{-4}	15685	59	416	7	4.47	actin filament bundle assembly
5215	9.34×10^{-4}	17373	1274	433	50	1.57	transporter activity
7017	9.38×10^{-4}	15685	320	416	19	2.24	microtubule-based process
31579	9.39×10^{-4}	15685	8	416	3	14.14	membrane raft organization
31063	9.39×10^{-4}	15685	8	416	3	14.14	regulation of histone deacetylation
7172	9.39×10^{-4}	15685	8	416	3	14.14	signal complex assembly
48703	9.39×10^{-4}	15685	8	416	3	14.14	embryonic viscerocranium morphogenesis
33687	9.39×10^{-4}	15685	8	416	3	14.14	osteoblast proliferation
43067	1.01×10^{-3}	15685	1142	416	48	1.58	regulation of programmed cell death
19233	1.01×10^{-3}	15685	60	416	7	4.40	sensory perception of pain
46928	1.06×10^{-3}	15685	30	416	5	6.28	regulation of neurotransmitter secretion
16070	1.06×10^{-3}	15685	3199	416	111	1.31	RNA metabolic process
48562	1.09×10^{-3}	15685	97	416	9	3.50	embryonic organ morphogenesis

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
44445	1.09×10^{-3}	17415	175	418	12	2.86	cytosolic part
48873	1.11×10^{-3}	15685	18	416	4	8.38	homeostasis of number of cells within a tissue
33198	1.11×10^{-3}	15685	18	416	4	8.38	response to ATP
48640	1.11×10^{-3}	15685	18	416	4	8.38	negative regulation of developmental growth
48872	1.11×10^{-3}	15685	61	416	7	4.33	homeostasis of number of cells
22892	1.14×10^{-3}	17373	1060	433	43	1.63	substrate-specific transporter activity
7269	1.15×10^{-3}	15685	79	416	8	3.82	neurotransmitter secretion
31346	1.15×10^{-3}	15685	79	416	8	3.82	positive regulation of cell projection organization
2682	1.18×10^{-3}	15685	851	416	38	1.68	regulation of immune system process
44431	1.19×10^{-3}	17415	594	418	27	1.89	Golgi apparatus part
32886	1.23×10^{-3}	15685	62	416	7	4.26	regulation of microtubule-based process
50885	1.23×10^{-3}	15685	31	416	5	6.08	neuromuscular process controlling balance
7405	1.23×10^{-3}	15685	31	416	5	6.08	neuroblast proliferation
33554	1.24×10^{-3}	15685	942	416	41	1.64	cellular response to stress
51402	1.27×10^{-3}	15685	140	416	11	2.96	neuron apoptosis
45165	1.27×10^{-3}	15685	140	416	11	2.96	cell fate commitment
31098	1.28×10^{-3}	15685	184	416	13	2.66	stress-activated protein kinase signaling cascade
51252	1.28×10^{-3}	15685	2207	416	81	1.38	regulation of RNA metabolic process
48747	1.28×10^{-3}	15685	46	416	6	4.92	muscle fiber development
6366	1.36×10^{-3}	15685	1067	416	45	1.59	transcription from RNA polymerase II promoter
4713	1.36×10^{-3}	17373	150	433	11	2.94	protein tyrosine kinase activity
60425	1.38×10^{-3}	15685	19	416	4	7.94	lung morphogenesis
32309	1.38×10^{-3}	15685	19	416	4	7.94	icosanoid secretion
1662	1.38×10^{-3}	15685	19	416	4	7.94	behavioral fear response
70997	1.42×10^{-3}	15685	142	416	11	2.92	neuron death
34613	1.43×10^{-3}	15685	121	416	10	3.12	cellular protein localization
60429	1.43×10^{-3}	15685	32	416	5	5.89	epithelium development
31333	1.44×10^{-3}	15685	47	416	6	4.81	negative regulation of protein complex assembly
21700	1.45×10^{-3}	15685	101	416	9	3.36	developmental maturation
3705	1.46×10^{-3}	17373	87	433	8	3.69	sequence-specific enhancer binding RNA polymerase II transcription factor activity
10721	1.47×10^{-3}	15685	82	416	8	3.68	negative regulation of cell development

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
42692	1.52×10^{-3}	15685	122	416	10	3.09	muscle cell differentiation
55002	1.61×10^{-3}	15685	48	416	6	4.71	striated muscle cell development
70727	1.62×10^{-3}	15685	123	416	10	3.07	cellular macromolecule localization
33157	1.62×10^{-3}	15685	123	416	10	3.07	regulation of intracellular protein transport
34645	1.64×10^{-3}	15685	3033	416	105	1.31	cellular macromolecule biosynthetic process
6139	1.65×10^{-3}	15685	4498	416	147	1.23	nucleobase nucleoside nucleotide and nucleic acid metabolic process
48471	1.67×10^{-3}	17415	371	418	19	2.13	perinuclear region of cytoplasm
6807	1.68×10^{-3}	15685	5036	416	162	1.21	nitrogen compound metabolic process
31058	1.69×10^{-3}	15685	20	416	4	7.54	positive regulation of histone modification
2209	1.69×10^{-3}	15685	20	416	4	7.54	behavioral defense response
51881	1.69×10^{-3}	15685	20	416	4	7.54	regulation of mitochondrial membrane potential
71715	1.69×10^{-3}	15685	20	416	4	7.54	icosanoid transport
33596	1.70×10^{-3}	17415	3	418	2	27.78	TSC1-TSC2 complex
34641	1.71×10^{-3}	15685	4930	416	159	1.22	cellular nitrogen compound metabolic process
55080	1.72×10^{-3}	15685	363	416	20	2.08	cation homeostasis
16020	1.72×10^{-3}	17415	7834	418	218	1.16	membrane
1763	1.77×10^{-3}	15685	168	416	12	2.69	morphogenesis of a branching structure
48585	1.78×10^{-3}	15685	498	416	25	1.89	negative regulation of response to stimulus
7417	1.82×10^{-3}	15685	125	416	10	3.02	central nervous system development
31327	1.85×10^{-3}	15685	555	416	27	1.83	negative regulation of cellular biosynthetic process
43269	1.86×10^{-3}	15685	169	416	12	2.68	regulation of ion transport
30072	1.86×10^{-3}	15685	169	416	12	2.68	peptide hormone secretion
31400	1.86×10^{-3}	15685	169	416	12	2.68	negative regulation of protein modification process
43009	1.87×10^{-3}	15685	192	416	13	2.55	chordate embryonic development
34329	1.88×10^{-3}	15685	147	416	11	2.82	cell junction assembly
30705	1.89×10^{-3}	15685	34	416	5	5.54	cytoskeleton-dependent intracellular transport
10970	1.89×10^{-3}	15685	34	416	5	5.54	microtubule-based transport
46822	1.93×10^{-3}	15685	126	416	10	2.99	regulation of nucleocytoplasmic transport

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
32259	1.93×10^{-3}	15685	126	416	10	2.99	methylation
46902	1.93×10^{-3}	15685	10	416	3	11.31	regulation of mitochondrial membrane permeability
90311	1.93×10^{-3}	15685	10	416	3	11.31	regulation of protein deacetylation
6582	1.93×10^{-3}	15685	10	416	3	11.31	melanin metabolic process
21756	1.93×10^{-3}	15685	10	416	3	11.31	striatum development
6605	1.94×10^{-3}	15685	315	416	18	2.15	protein targeting
51494	1.94×10^{-3}	15685	67	416	7	3.94	negative regulation of cytoskeleton organization
7243	1.98×10^{-3}	15685	614	416	29	1.78	intracellular protein kinase cascade
30168	2.00×10^{-3}	15685	241	416	15	2.35	platelet activation
43408	2.00×10^{-3}	15685	217	416	14	2.43	regulation of MAPKKK cascade
14070	2.00×10^{-3}	15685	217	416	14	2.43	response to organic cyclic compound
5891	2.02×10^{-3}	17415	23	418	4	7.25	voltage-gated calcium channel complex
46777	2.03×10^{-3}	15685	106	416	9	3.20	protein autophosphorylation
7271	2.04×10^{-3}	15685	21	416	4	7.18	synaptic transmission cholinergic

Table 4.9: GO terms significantly enriched among *AAP-associated* genes from duplication CNVs - P, probability; N, total genes in the genome for each GO category (i.e. biological processes, molecular function, and cellular component); G, total genes annotated; n, number of input genes; x, number of annotated genes among the input genes; fold change, fold change over expected by random chance; GO term, term description

GOID	P	N	G	n	x	fold change	GO term
23052	2.53×10^{-11}	15685	1999	211	66	2.45	signaling
71840	3.90×10^{-11}	15685	3403	211	88	1.92	cellular component organization or biogenesis
16043	5.58×10^{-11}	15685	3372	211	87	1.92	cellular component organization
42995	9.50×10^{-11}	17415	976	210	38	3.23	cell projection
7268	1.03×10^{-10}	15685	454	211	27	4.42	synaptic transmission
5829	5.49×10^{-10}	17415	2092	210	58	2.30	cytosol
7267	1.17×10^{-9}	15685	739	211	33	3.32	cell-cell signaling
7154	1.68×10^{-9}	15685	1722	211	54	2.33	cell communication
71841	2.97×10^{-9}	15685	2617	211	70	1.99	cellular component organization or biogenesis at cellular level
45202	3.43×10^{-9}	17415	438	210	23	4.35	synapse
5737	4.23×10^{-9}	17415	8582	210	145	1.40	cytoplasm

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
71842	4.62×10^{-9}	15685	2587	211	69	1.98	cellular component organization at cellular level
44444	8.90×10^{-9}	17415	6195	210	115	1.54	cytoplasmic part
7010	9.72×10^{-9}	15685	601	211	28	3.46	cytoskeleton organization
5515	1.11×10^{-8}	17373	10841	220	176	1.28	protein binding
44424	1.16×10^{-8}	17415	11891	210	179	1.25	intracellular part
6996	1.91×10^{-8}	15685	1688	211	51	2.25	organelle organization
32502	2.72×10^{-8}	15685	3852	211	88	1.70	developmental process
32991	3.01×10^{-8}	17415	3888	210	82	1.75	macromolecular complex
43005	4.29×10^{-8}	17415	461	210	22	3.96	neuron projection
48523	4.77×10^{-8}	15685	2279	211	61	1.99	negative regulation of cellular process
48856	4.84×10^{-8}	15685	2336	211	62	1.97	anatomical structure development
44456	6.81×10^{-8}	17415	320	210	18	4.66	synapse part
43234	8.52×10^{-8}	17415	3229	210	71	1.82	protein complex
48519	8.92×10^{-8}	15685	2489	211	64	1.91	negative regulation of biological process
23051	1.74×10^{-7}	15685	1544	211	46	2.21	regulation of signaling
5622	2.50×10^{-7}	17415	12324	210	180	1.21	intracellular
19904	3.16×10^{-7}	17373	492	220	22	3.53	protein domain specific binding
43025	3.98×10^{-7}	17415	217	210	14	5.35	neuronal cell body
97060	4.75×10^{-7}	17415	188	210	13	5.73	synaptic membrane
45211	6.36×10^{-7}	17415	162	210	12	6.14	postsynaptic membrane
43229	7.09×10^{-7}	17415	10129	210	156	1.28	intracellular organelle
44297	8.91×10^{-7}	17415	232	210	14	5.00	cell body
43226	9.70×10^{-7}	17415	10167	210	156	1.27	organelle
12506	1.09×10^{-6}	17415	236	210	14	4.92	vesicle membrane
30036	1.73×10^{-6}	15685	322	211	17	3.92	actin cytoskeleton organization
51128	2.30×10^{-6}	15685	877	211	30	2.54	regulation of cellular component organization
32501	3.47×10^{-6}	15685	4009	211	84	1.56	multicellular organismal process
5230	3.47×10^{-6}	17373	72	220	8	8.77	extracellular ligand-gated ion channel activity
30659	3.57×10^{-6}	17415	225	210	13	4.79	cytoplasmic vesicle membrane
44463	3.66×10^{-6}	17415	464	210	19	3.40	cell projection part
50794	6.32×10^{-6}	15685	7745	211	136	1.31	regulation of cellular process
30029	7.44×10^{-6}	15685	359	211	17	3.52	actin filament-based process
44422	8.03×10^{-6}	17415	5841	210	101	1.43	organelle part
33043	1.16×10^{-5}	15685	333	211	16	3.57	regulation of organelle organization
50804	1.29×10^{-5}	15685	163	211	11	5.02	regulation of synaptic transmission

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
65008	1.33×10^{-5}	15685	2092	211	51	1.81	regulation of biological quality
44446	1.48×10^{-5}	17415	5762	210	99	1.42	intracellular organelle part
30054	1.57×10^{-5}	17415	609	210	21	2.86	cell junction
19717	1.60×10^{-5}	17415	121	210	9	6.17	synaptosome
51239	1.61×10^{-5}	15685	1381	211	38	2.05	regulation of multicellular organismal process
8283	1.88×10^{-5}	15685	1337	211	37	2.06	cell proliferation
5488	1.98×10^{-5}	17373	14540	220	205	1.11	binding
48869	2.11×10^{-5}	15685	2069	211	50	1.80	cellular developmental process
51969	2.51×10^{-5}	15685	175	211	11	4.67	regulation of transmission of nerve impulse
9966	2.66×10^{-5}	15685	1305	211	36	2.05	regulation of signal transduction
50789	2.89×10^{-5}	15685	8161	211	139	1.27	regulation of biological process
30900	3.17×10^{-5}	15685	67	211	7	7.77	forebrain development
31982	3.79×10^{-5}	17415	801	210	24	2.48	vesicle
43228	4.30×10^{-5}	17415	2837	210	57	1.67	non-membrane-bounded organelle
43232	4.30×10^{-5}	17415	2837	210	57	1.67	intracellular non-membrane-bounded organelle
19226	4.86×10^{-5}	15685	188	211	11	4.35	transmission of nerve impulse
31644	4.86×10^{-5}	15685	188	211	11	4.35	regulation of neurological system process
35637	4.86×10^{-5}	15685	188	211	11	4.35	multicellular organismal signaling
7399	5.32×10^{-5}	15685	640	211	22	2.56	nervous system development
51234	5.52×10^{-5}	15685	3442	211	71	1.53	establishment of localization
43406	5.56×10^{-5}	15685	158	211	10	4.70	positive regulation of MAP kinase activity
31399	5.86×10^{-5}	15685	787	211	25	2.36	regulation of protein modification process
31410	6.42×10^{-5}	17415	775	210	23	2.46	cytoplasmic vesicle
65009	6.54×10^{-5}	15685	1417	211	37	1.94	regulation of molecular function
4890	6.66×10^{-5}	17373	18	220	4	17.55	GABA-A receptor activity
16310	6.75×10^{-5}	15685	1203	211	33	2.04	phosphorylation
6810	6.92×10^{-5}	15685	3401	211	70	1.53	transport
7610	7.81×10^{-5}	15685	432	211	17	2.93	behavior
15908	8.32×10^{-5}	15685	54	211	6	8.26	fatty acid transport
48640	8.40×10^{-5}	15685	18	211	4	16.52	negative regulation of developmental growth
44433	8.81×10^{-5}	17415	305	210	13	3.53	cytoplasmic vesicle part
6468	8.87×10^{-5}	15685	1114	211	31	2.07	protein phosphorylation

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
51129	9.35×10^{-5}	15685	238	211	12	3.75	negative regulation of cellular component organization
42802	9.49×10^{-5}	17373	808	220	24	2.35	identical protein binding
32309	1.05×10^{-4}	15685	19	211	4	15.65	icosanoid secretion
7015	1.24×10^{-4}	15685	174	211	10	4.27	actin filament organization
8092	1.28×10^{-4}	17373	572	220	19	2.62	cytoskeletal protein binding
71715	1.30×10^{-4}	15685	20	211	4	14.87	icosanoid transport
5856	1.34×10^{-4}	17415	1574	210	36	1.90	cytoskeleton
5642	1.45×10^{-4}	17415	2	210	2	82.93	annulate lamellae
19222	1.53×10^{-4}	15685	3944	211	77	1.45	regulation of metabolic process
7165	1.55×10^{-4}	15685	4481	211	85	1.41	signal transduction
51493	1.57×10^{-4}	15685	179	211	10	4.15	regulation of cytoskeleton organization
65007	1.64×10^{-4}	15685	8625	211	142	1.22	biological regulation
51246	1.65×10^{-4}	15685	1153	211	31	2.00	regulation of protein metabolic process
43405	1.68×10^{-4}	15685	216	211	11	3.79	regulation of MAP kinase activity
48731	1.72×10^{-4}	15685	945	211	27	2.12	system development
42391	1.83×10^{-4}	15685	117	211	8	5.08	regulation of membrane potential
71902	1.88×10^{-4}	15685	183	211	10	4.06	positive regulation of protein serine/threonine kinase activity
48468	2.16×10^{-4}	15685	609	211	20	2.44	cell development
5626	2.36×10^{-4}	17415	959	210	25	2.16	insoluble fraction
15276	2.39×10^{-4}	17373	129	220	8	4.90	ligand-gated ion channel activity
22834	2.39×10^{-4}	17373	129	220	8	4.90	ligand-gated channel activity
6873	2.40×10^{-4}	15685	430	211	16	2.77	cellular ion homeostasis
51716	2.55×10^{-4}	15685	5092	211	93	1.36	cellular response to stimulus
43231	2.63×10^{-4}	17415	9016	210	134	1.23	intracellular membrane-bounded organelle
32989	2.73×10^{-4}	15685	435	211	16	2.73	cellular component morphogenesis
43227	2.73×10^{-4}	17415	9022	210	134	1.23	membrane-bounded organelle
902	2.81×10^{-4}	15685	308	211	13	3.14	cell morphogenesis
42325	2.83×10^{-4}	15685	670	211	21	2.33	regulation of phosphorylation
42592	2.86×10^{-4}	15685	923	211	26	2.09	homeostatic process
16917	3.01×10^{-4}	17373	26	220	4	12.15	GABA receptor activity
5794	3.26×10^{-4}	17415	1038	210	26	2.08	Golgi apparatus
10646	3.52×10^{-4}	15685	1095	211	29	1.97	regulation of cell communication
44267	3.57×10^{-4}	15685	3064	211	62	1.50	cellular protein metabolic process

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
55082	3.79×10^{-4}	15685	448	211	16	2.65	cellular chemical homeostasis
50896	3.88×10^{-4}	15685	6717	211	115	1.27	response to stimulus
45860	3.94×10^{-4}	15685	319	211	13	3.03	positive regulation of protein kinase activity
6793	3.95×10^{-4}	15685	1379	211	34	1.83	phosphorus metabolic process
6796	3.95×10^{-4}	15685	1379	211	34	1.83	phosphate metabolic process
226	3.99×10^{-4}	15685	201	211	10	3.70	microtubule cytoskeleton organization
51649	4.02×10^{-4}	15685	1551	211	37	1.77	establishment of localization in cell

Table 4.10: GO terms significantly enriched among *AAP-associated* genes from deletion CNVs - P, probability; N, total genes in the genome for each GO category (i.e. biological processes, molecular function, and cellular component); G, total genes annotated; n, number of input genes; x, number of annotated genes among the input genes; fold change, fold change over expected by random chance; GO term, term description

GOID	P	N	G	n	x	fold change	GO term
786	2.15×10^{-11}	17415	105	177	13	12.18	nucleosome
7610	3.66×10^{-11}	15685	432	174	24	5.01	behavior
65008	4.10×10^{-11}	15685	2092	174	56	2.41	regulation of biological quality
65003	1.09×10^{-10}	15685	884	174	34	3.47	macromolecular complex assembly
32993	3.69×10^{-10}	17415	124	177	13	10.32	protein-DNA complex
22607	2.26×10^{-9}	15685	1303	174	40	2.77	cellular component assembly
34622	3.08×10^{-9}	15685	519	174	24	4.17	cellular macromolecular complex assembly
51239	3.60×10^{-9}	15685	1381	174	41	2.68	regulation of multicellular organismal process
45202	3.92×10^{-9}	17415	438	177	21	4.72	synapse
6334	5.62×10^{-9}	15685	141	174	13	8.31	nucleosome assembly
51704	7.88×10^{-9}	15685	825	174	30	3.28	multi-organism process
50793	9.28×10^{-9}	15685	1037	174	34	2.96	regulation of developmental process
43933	1.05×10^{-8}	15685	1042	174	34	2.94	macromolecular complex subunit organization
34728	1.30×10^{-8}	15685	151	174	13	7.76	nucleosome organization
34621	2.23×10^{-8}	15685	667	174	26	3.51	cellular macromolecular complex subunit organization
65004	2.25×10^{-8}	15685	158	174	13	7.42	protein-DNA complex assembly
32502	2.31×10^{-8}	15685	3852	174	76	1.78	developmental process
44456	2.99×10^{-8}	17415	320	177	17	5.23	synapse part

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
71824	3.52×10^{-8}	15685	164	174	13	7.15	protein-DNA complex subunit organization
48518	5.30×10^{-8}	15685	2824	174	61	1.95	positive regulation of biological process
71844	7.83×10^{-8}	15685	967	174	31	2.89	cellular component assembly at cellular level
48869	1.21×10^{-7}	15685	2069	174	49	2.13	cellular developmental process
9653	1.67×10^{-7}	15685	1341	174	37	2.49	anatomical structure morphogenesis
35176	1.95×10^{-7}	15685	24	174	6	22.54	social behavior
51705	2.51×10^{-7}	15685	40	174	7	15.78	behavioral interaction between organisms
48522	2.62×10^{-7}	15685	2527	174	55	1.96	positive regulation of cellular process
2000026	2.98×10^{-7}	15685	812	174	27	3.00	regulation of multicellular organismal development
45595	5.42×10^{-7}	15685	733	174	25	3.07	regulation of cell differentiation
14069	5.73×10^{-7}	17415	96	177	9	9.22	postsynaptic density
44422	6.44×10^{-7}	17415	5841	177	91	1.53	organelle part
50896	1.05×10^{-6}	15685	6717	174	106	1.42	response to stimulus
32947	1.07×10^{-6}	17373	33	182	6	17.36	protein complex scaffold
43229	1.60×10^{-6}	17415	10129	177	133	1.29	intracellular organelle
48856	1.92×10^{-6}	15685	2336	174	50	1.93	anatomical structure development
43226	2.11×10^{-6}	17415	10167	177	133	1.29	organelle
48858	2.44×10^{-6}	15685	237	174	13	4.94	cell projection morphogenesis
44424	2.45×10^{-6}	17415	11891	177	148	1.22	intracellular part
7268	3.06×10^{-6}	15685	454	174	18	3.57	synaptic transmission
32990	3.07×10^{-6}	15685	242	174	13	4.84	cell part morphogenesis
22603	3.12×10^{-6}	15685	409	174	17	3.75	regulation of anatomical structure morphogenesis
44446	3.14×10^{-6}	17415	5762	177	88	1.50	intracellular organelle part
60249	3.25×10^{-6}	15685	171	174	11	5.80	anatomical structure homeostasis
32501	3.36×10^{-6}	15685	4009	174	72	1.62	multicellular organismal process
42592	3.50×10^{-6}	15685	923	174	27	2.64	homeostatic process
6461	5.31×10^{-6}	15685	620	174	21	3.05	protein complex assembly
65007	6.96×10^{-6}	15685	8625	174	124	1.30	biological regulation
32989	7.07×10^{-6}	15685	435	174	17	3.52	cellular component morphogenesis
7267	7.40×10^{-6}	15685	739	174	23	2.81	cell-cell signaling
48812	7.66×10^{-6}	15685	224	174	12	4.83	neuron projection morphogenesis

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
71840	7.76×10^{-6}	15685	3403	174	63	1.67	cellular component organization or biogenesis
5622	1.01×10^{-5}	17415	12324	177	150	1.20	intracellular
71841	1.01×10^{-5}	15685	2617	174	52	1.79	cellular component organization or biogenesis at cellular level
42221	1.14×10^{-5}	15685	2344	174	48	1.85	response to chemical stimulus
16043	1.20×10^{-5}	15685	3372	174	62	1.66	cellular component organization
60501	1.32×10^{-5}	15685	5	174	3	54.09	positive regulation of epithelial cell proliferation involved in lung morphogenesis
2000794	1.32×10^{-5}	15685	5	174	3	54.09	regulation of epithelial cell proliferation involved in lung morphogenesis
6996	1.55×10^{-5}	15685	1688	174	38	2.03	organelle organization
71842	1.59×10^{-5}	15685	2587	174	51	1.78	cellular component organization at cellular level
7612	1.61×10^{-5}	15685	73	174	7	8.64	learning
45653	1.82×10^{-5}	15685	15	174	4	24.04	negative regulation of megakaryocyte differentiation
30154	2.27×10^{-5}	15685	1717	174	38	2.00	cell differentiation
60502	2.62×10^{-5}	15685	6	174	3	45.07	epithelial cell proliferation involved in lung morphogenesis
9887	3.03×10^{-5}	15685	487	174	17	3.15	organ morphogenesis
31325	3.05×10^{-5}	15685	1162	174	29	2.25	positive regulation of cellular metabolic process
5217	3.18×10^{-5}	17373	18	182	4	21.21	intracellular ligand-gated ion channel activity
50890	3.26×10^{-5}	15685	144	174	9	5.63	cognition
9893	3.97×10^{-5}	15685	1241	174	30	2.18	positive regulation of metabolic process
32271	4.05×10^{-5}	15685	84	174	7	7.51	regulation of protein polymerization
43228	4.06×10^{-5}	17415	2837	177	50	1.73	non-membrane-bounded organelle
43232	4.06×10^{-5}	17415	2837	177	50	1.73	intracellular non-membrane-bounded organelle
46903	4.29×10^{-5}	15685	713	174	21	2.66	secretion
48468	4.73×10^{-5}	15685	609	174	19	2.81	cell development
42391	4.77×10^{-5}	15685	117	174	8	6.16	regulation of membrane potential
30054	5.01×10^{-5}	17415	609	177	18	2.91	cell junction
31334	5.21×10^{-5}	15685	60	174	6	9.01	positive regulation of protein complex assembly

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
7275	5.33×10^{-5}	15685	1584	174	35	1.99	multicellular organismal development
32273	5.92×10^{-5}	15685	38	174	5	11.86	positive regulation of protein polymerization
50789	5.93×10^{-5}	15685	8161	174	116	1.28	regulation of biological process
45652	6.17×10^{-5}	15685	20	174	4	18.03	regulation of megakaryocyte differentiation
55001	6.28×10^{-5}	15685	62	174	6	8.72	muscle cell development
31579	7.21×10^{-5}	15685	8	174	3	33.80	membrane raft organization
33687	7.21×10^{-5}	15685	8	174	3	33.80	osteoblast proliferation
71822	7.59×10^{-5}	15685	742	174	21	2.55	protein complex subunit organization
43025	7.71×10^{-5}	17415	217	177	10	4.53	neuronal cell body
65009	7.87×10^{-5}	15685	1417	174	32	2.04	regulation of molecular function
122	7.99×10^{-5}	15685	425	174	15	3.18	negative regulation of transcription from RNA polymerase II promoter
43254	8.07×10^{-5}	15685	126	174	8	5.72	regulation of protein complex assembly
60537	8.53×10^{-5}	15685	127	174	8	5.68	muscle tissue development
50804	8.58×10^{-5}	15685	163	174	9	4.98	regulation of synaptic transmission
6325	9.04×10^{-5}	15685	532	174	17	2.88	chromatin organization
9605	9.33×10^{-5}	15685	1111	174	27	2.19	response to external stimulus
9952	9.42×10^{-5}	15685	165	174	9	4.92	anterior/posterior pattern formation
44057	1.02×10^{-4}	15685	385	174	14	3.28	regulation of system process
8154	1.02×10^{-4}	15685	97	174	7	6.51	actin polymerization or depolymerization
19725	1.06×10^{-4}	15685	539	174	17	2.84	cellular homeostasis
30219	1.10×10^{-4}	15685	23	174	4	15.68	megakaryocyte differentiation
48519	1.15×10^{-4}	15685	2489	174	47	1.70	negative regulation of biological process
7611	1.18×10^{-4}	15685	133	174	8	5.42	learning or memory
70836	1.22×10^{-4}	15685	2	174	2	90.14	caveola assembly
1765	1.22×10^{-4}	15685	2	174	2	90.14	membrane raft assembly
32940	1.24×10^{-4}	15685	600	174	18	2.70	secretion by cell
23052	1.28×10^{-4}	15685	1999	174	40	1.80	signaling
6950	1.31×10^{-4}	15685	2285	174	44	1.74	response to stress
97060	1.33×10^{-4}	17415	188	177	9	4.71	synaptic membrane
44297	1.34×10^{-4}	17415	232	177	10	4.24	cell body
32991	1.40×10^{-4}	17415	3888	177	61	1.54	macromolecular complex
7015	1.41×10^{-4}	15685	174	174	9	4.66	actin filament organization

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
42995	1.45×10^{-4}	17415	976	177	23	2.32	cell projection
51969	1.47×10^{-4}	15685	175	174	9	4.64	regulation of transmission of nerve impulse
33555	1.50×10^{-4}	15685	46	174	5	9.80	multicellular organismal response to stress
7409	1.54×10^{-4}	15685	176	174	9	4.61	axonogenesis
45165	1.68×10^{-4}	15685	140	174	8	5.15	cell fate commitment
902	1.70×10^{-4}	15685	308	174	12	3.51	cell morphogenesis
267	1.72×10^{-4}	17415	1257	177	27	2.11	cell fraction
43623	1.73×10^{-4}	15685	263	174	11	3.77	cellular protein complex assembly
51707	1.78×10^{-4}	15685	357	174	13	3.28	response to other organism
42133	1.81×10^{-4}	15685	26	174	4	13.87	neurotransmitter metabolic process
2682	1.88×10^{-4}	15685	851	174	22	2.33	regulation of immune system process
30424	1.89×10^{-4}	17415	242	177	10	4.07	axon
31982	1.91×10^{-4}	17415	801	177	20	2.46	vesicle
5737	2.06×10^{-4}	17415	8582	177	111	1.27	cytoplasm
15671	2.07×10^{-4}	15685	11	174	3	24.58	oxygen transport
7569	2.24×10^{-4}	15685	50	174	5	9.01	cell aging
16986	2.26×10^{-4}	17373	29	182	4	13.17	transcription initiation factor activity
9607	2.35×10^{-4}	15685	522	174	16	2.76	response to biotic stimulus
51049	2.38×10^{-4}	15685	747	174	20	2.41	regulation of transport
2053	2.43×10^{-4}	15685	28	174	4	12.88	positive regulation of mesenchymal cell proliferation
1505	2.50×10^{-4}	15685	112	174	7	5.63	regulation of neurotransmitter levels
19226	2.52×10^{-4}	15685	188	174	9	4.32	transmission of nerve impulse
31644	2.52×10^{-4}	15685	188	174	9	4.32	regulation of neurological system process
35637	2.52×10^{-4}	15685	188	174	9	4.32	multicellular organismal signaling
45211	2.54×10^{-4}	17415	162	177	8	4.86	postsynaptic membrane
43227	2.55×10^{-4}	17415	9022	177	115	1.25	membrane-bounded organelle
7389	2.59×10^{-4}	15685	371	174	13	3.16	pattern specification process
10033	2.67×10^{-4}	15685	1447	174	31	1.93	response to organic substance
43005	2.69×10^{-4}	17415	461	177	14	2.99	neuron projection
5515	2.69×10^{-4}	17373	10841	182	136	1.20	protein binding
5833	2.74×10^{-4}	17415	13	177	3	22.71	hemoglobin complex
10604	2.87×10^{-4}	15685	1126	174	26	2.08	positive regulation of macromolecule metabolic process
44427	2.93×10^{-4}	17415	522	177	15	2.83	chromosomal part
51607	3.18×10^{-4}	15685	83	174	6	6.52	defense response to virus
42802	3.21×10^{-4}	17373	808	182	20	2.36	identical protein binding

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
7519	3.40×10^{-4}	15685	84	174	6	6.44	skeletal muscle tissue development
14706	3.44×10^{-4}	15685	118	174	7	5.35	striated muscle tissue development
31410	3.50×10^{-4}	17415	775	177	19	2.41	cytoplasmic vesicle
51899	3.52×10^{-4}	15685	55	174	5	8.19	membrane depolarization
48745	3.54×10^{-4}	15685	13	174	3	20.80	smooth muscle tissue development
1894	3.62×10^{-4}	15685	85	174	6	6.36	tissue homeostasis
10464	3.64×10^{-4}	15685	31	174	4	11.63	regulation of mesenchymal cell proliferation
3002	3.76×10^{-4}	15685	242	174	10	3.72	regionalization
9719	3.76×10^{-4}	15685	715	174	19	2.40	response to endogenous stimulus
5344	3.78×10^{-4}	17373	14	182	3	20.45	oxygen transporter activity
15980	3.83×10^{-4}	15685	199	174	9	4.08	energy derivation by oxidation of organic compounds
48646	4.03×10^{-4}	15685	548	174	16	2.63	anatomical structure formation involved in morphogenesis
42692	4.21×10^{-4}	15685	122	174	7	5.17	muscle cell differentiation
43209	4.29×10^{-4}	17415	15	177	3	19.68	myelin sheath
43231	4.34×10^{-4}	17415	9016	177	114	1.24	intracellular membrane-bounded organelle
6112	4.34×10^{-4}	15685	161	174	8	4.48	energy reserve metabolic process
904	4.43×10^{-4}	15685	203	174	9	4.00	cell morphogenesis involved in differentiation
42136	4.46×10^{-4}	15685	14	174	3	19.32	neurotransmitter biosynthetic process
4707	4.69×10^{-4}	17373	15	182	3	19.09	MAP kinase activity
55082	4.80×10^{-4}	15685	448	174	14	2.82	cellular chemical homeostasis
51130	4.88×10^{-4}	15685	297	174	11	3.34	positive regulation of cellular component organization
30659	5.01×10^{-4}	17415	225	177	9	3.94	cytoplasmic vesicle membrane
5886	5.15×10^{-4}	17415	4157	177	62	1.47	plasma membrane
33138	5.23×10^{-4}	15685	34	174	4	10.61	positive regulation of peptidyl-serine phosphorylation
44464	5.66×10^{-4}	17415	16105	177	174	1.06	cell part
5623	5.71×10^{-4}	17415	16106	177	174	1.06	cell
34724	5.86×10^{-4}	15685	35	174	4	10.30	DNA replication-independent nucleosome organization
34080	5.86×10^{-4}	15685	35	174	4	10.30	CenH3-containing nucleosome assembly at centromere

Continued on next page

GOID	P	N	G	n	x	fold change	GO term
6336	5.86×10^{-4}	15685	35	174	4	10.30	DNA replication-independent nucleosome assembly
10638	5.88×10^{-4}	15685	129	174	7	4.89	positive regulation of organelle organization

4.6.3 REVIGO-summarised GO terms

The following figures in this section (Appendix Figures 4.5-4.8) show summarised GO terms that were significantly enriched among ASD candidate genes identified here. REVIGO was used for this analysis with the following parameters: allowed similarity=0.7, scores=p-values associated with a GO term, database with GO term sizes=whole UniProt, semantic similarity measure=SimRel.

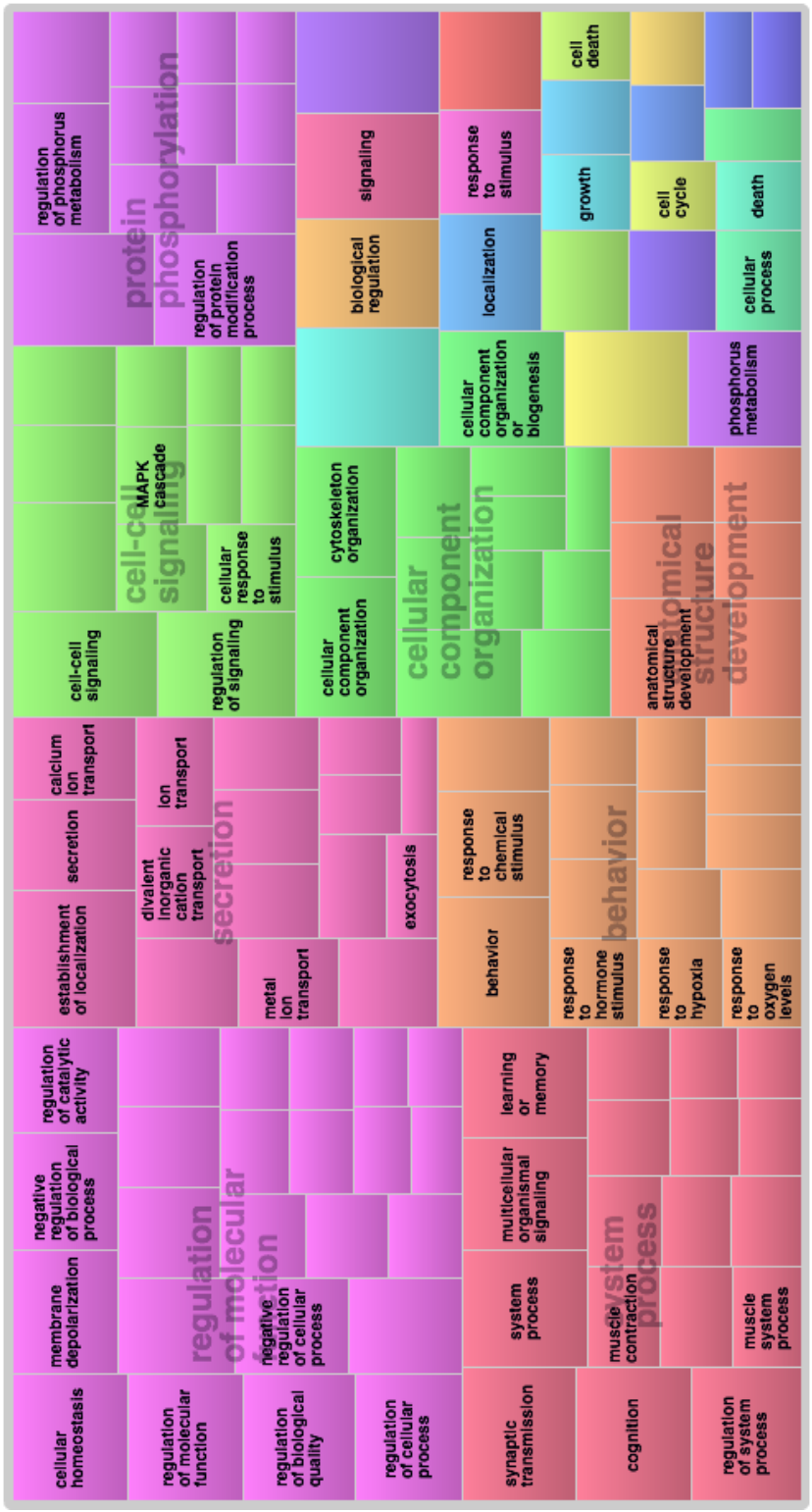


Figure 4.8: REVIGO-summarised Biological Process GO terms enriched within *AST-associated* network genes - i.e. *AST+ASD-implicated+dirPPI* genes. The summarised GO terms are given in full in Table 4.5.

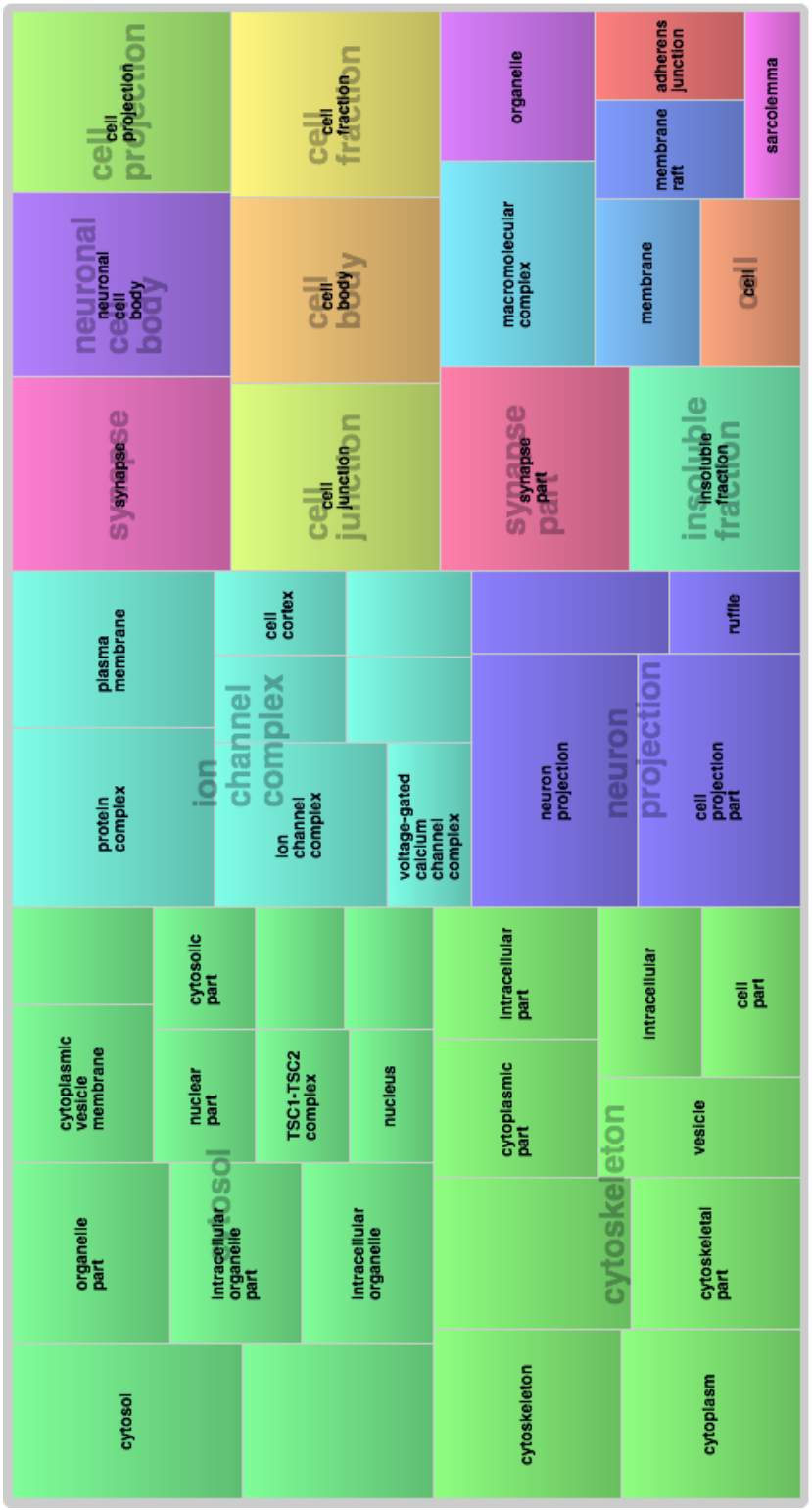


Figure 4.9: REVIGO-summarised Cellular Component GO terms enriched within *AST-associated* network genes - i.e. *AST+ASD-implicated+dirPPI* genes. The summarised GO terms are given in full in Table 4.5.

Figure 4.10: REVIGO-summarised Molecular Function GO terms enriched within *AST-associated* network genes - i.e. *AST+ASD-implicated+dirPPI* genes. The summarised GO terms are given in full in Table 4.5.

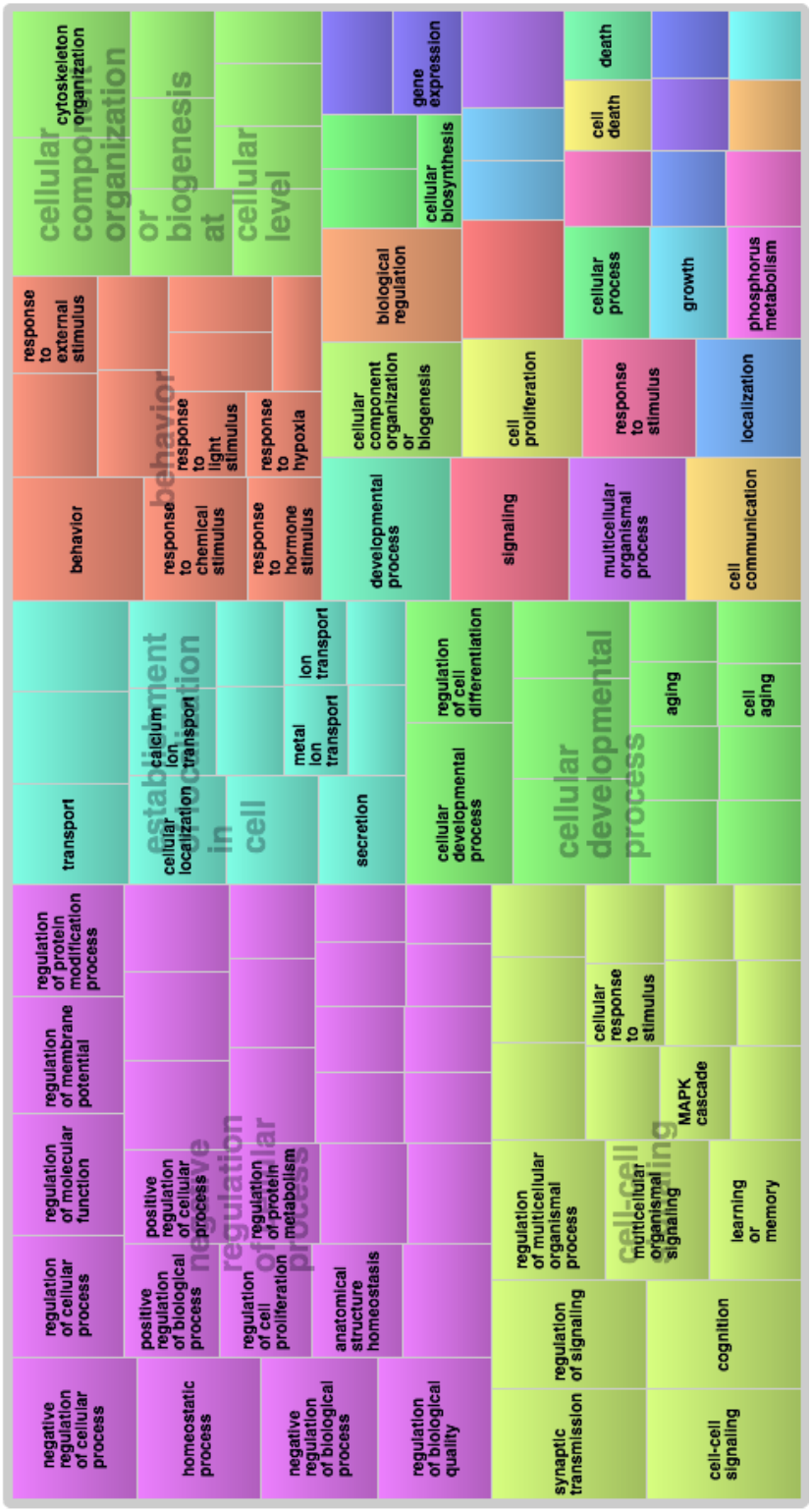


Figure 4.11: REVIGO-summarised Biological Process GO terms enriched within *AAP-associated* network genes - i.e. *AAP+ASD-implicated+dirPPI* genes. The summarised GO terms are given in full in Table 4.8.

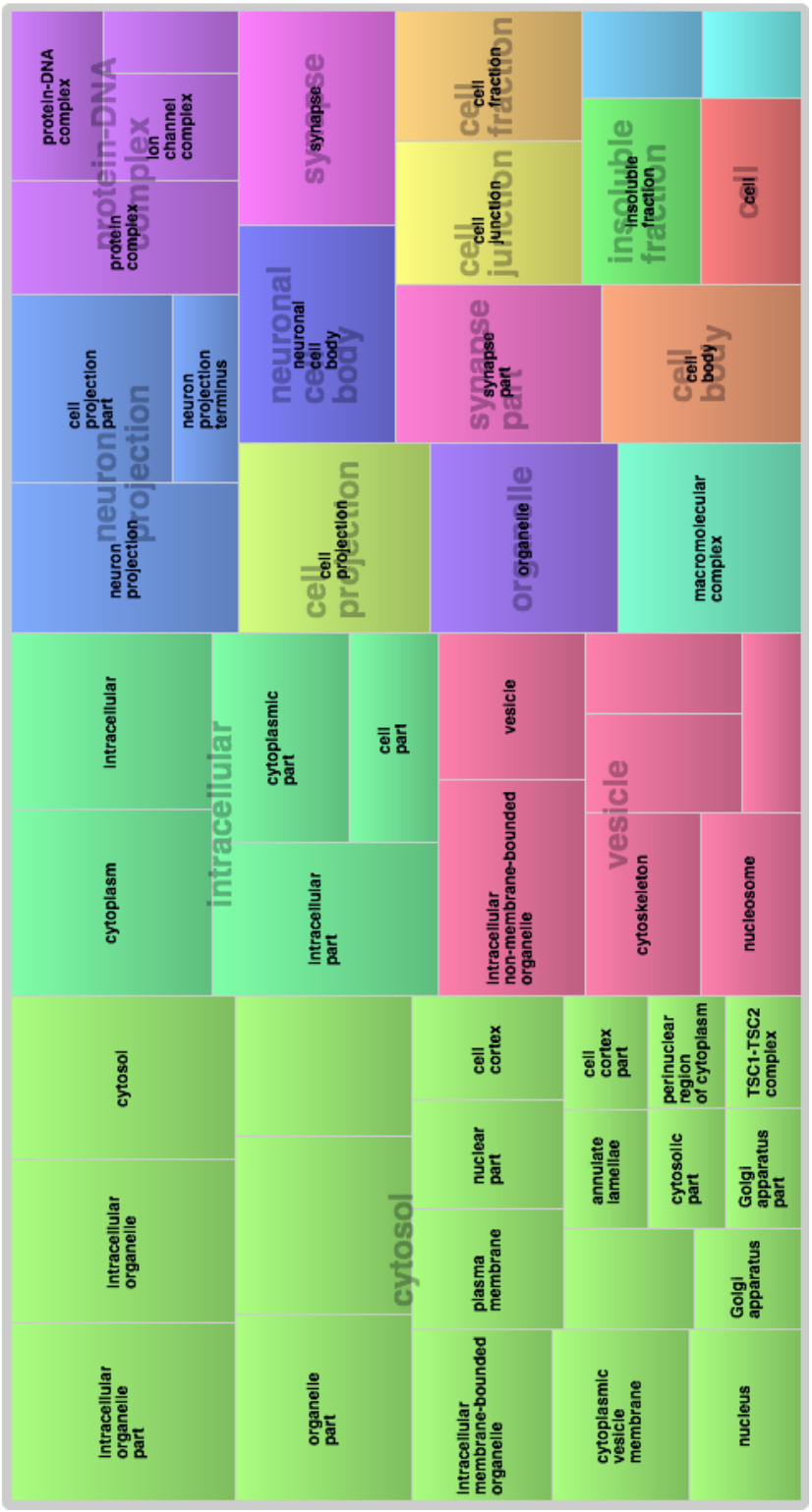


Figure 4.12: REVIGO-summarised Cellular Component GO terms enriched within *AAP-associated* network genes - i.e. *AAP+ASD-implicated+dirPPI* genes. The summarised GO terms are given in full in Table 4.8.

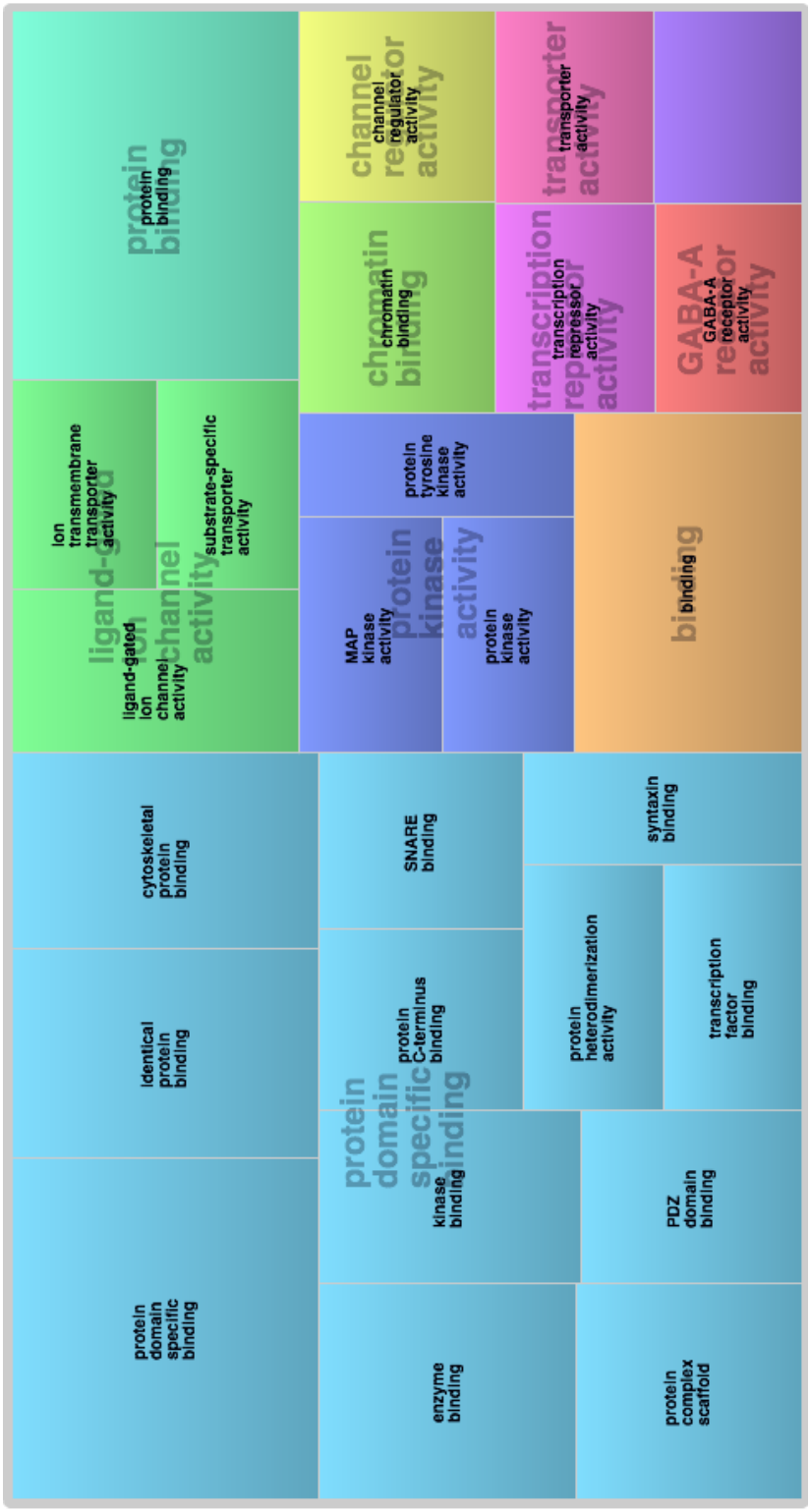


Figure 4.13: REVIGO-summarised Molecular Function GO terms enriched within *AAP-associated* network genes - i.e. *AAP+ASD-implicated+dirPPI* genes. The summarised GO terms are given in full in Table 4.8.

4.6.4 PPI network connectivity to recently published exome studies

Table 4.11: PPI network connectivity between this study and two recently published exome-sequencing studies for ASD - Gene sets prefixed with 'eich' are reported in reference (274), and sets prefixed with 'neal' in (275); Filter, genes that have HGNC symbols and DAPPLE entry are considered; Background, the expected number of protein interaction is calculated from the whole genome (WG); # Genes, number of genes in a set shown in the first column (numbers of genes in a combined set of a gene set from an exome-sequencing study and *AST-associated* network genes); # Connections, number of protein interactions between an exome-sequencing set and *AST-associated* network genes (number of total protein protein interaction identified in a combined set); P-value, probabilities obtained from 500 randomisations; Fold Change, number of protein interactions between two given sets over that expected by chance.

Gene sets	Filter	Background	# Genes	# Connections	P-value	Fold Change
eichsup12	HGNC+DAPPLE	WG	126 (326)	138 (496)	<0.002	12.23
eichsup12con	HGNC+DAPPLE	WG	49 (251)	86 (440)	<0.002	11.75
eichtab1	HGNC+DAPPLE	WG	55 (257)	98 (451)	<0.002	30.49
eichtab1cnv	HGNC+DAPPLE	WG	60 (261)	103 (459)	<0.002	30.62
nealfig2	HGNC+DAPPLE	WG	45 (247)	84 (437)	<0.002	37.4
nealsupfig3	HGNC+DAPPLE	WG	137 (323)	847 (1537)	<0.002	5.41

4. FUNCTIONAL GENOMICS IDENTIFIES CONVERGENT AETIOLOGIES FOR BOTH DELETIONS AND DUPLICATIONS FOUND IN ASD

5

Targeted resequencing of human genomic regions for obsessive-compulsive disorder

5.1 Abstract

Obsessive-compulsive disorder (OCD), a critical mental condition exhibited as extreme repetition of behaviours, affects 1-3% of the human population. Twin studies have suggested a strong heritability for OCD, with concordance rates of 80-84% and 47-50% in monozygotic and dizygotic twins, respectively. Despite >60 candidate gene studies, no genome-wide significant, replicable genetic associations have been found thus far. However, neuroimaging studies have provided compelling evidence for an involvement of the cortico-striatal-thalamo-cortical (CSTC) circuit in OCD. Furthermore, animal models of OCD have suggested defects in glutamatergic neurotransmission in the CSTC circuit in OCD. Thus, I resequenced >600 genes that occur in the CSTC circuit and those suggested from animal genetic studies, in 400 human patients and controls, in order to identify rare variants contributing to OCD. Targeted resequencing yielded >90,000 variants of which approximately 40% were novel. Case-control comparison of allele frequencies of the detected variants led to an identification of several nominally significantly associated non-coding variants that might act within transcription factor binding sites or functional RNA genes in OCD. Examining groupwise effects of allele frequency differences between cases and controls per functional unit (i.e. gene) yielded significantly ($p_{perm} < 0.05$) associated genes including *ROBO1, CDH13, ABCF3*, and *TMEM41A*. Pathway analysis of the associated genes suggested several OCD-relevant functional themes including neuron projection organisation, synaptic transmission, cell adhesion, and telencephalon development. Finally, extrapolation of the simulated results from sub-sampled data suggested that doubling the sample size may yield more strongly associated genes.

5.2 Introduction

5.2.1 Obsessive-Compulsive Disorder (OCD)

5.2.1.1 Classification and prevalence

Obsessive-Compulsive Disorder (OCD) is a debilitating psychiatric condition that is characterised by obsessions (i.e. recurrent and persistent thoughts, impulses, or images) or compulsions that are defined by repetitive behaviours (e.g. hand washing, ordering, checking) or mental acts (e.g. praying, counting, repeating words silently). These behaviours are aimed at preventing or reducing distress, but they are not connected in a realistic way with what they are designed to neutralise. Individuals with OCD recognise that their obsessions/compulsions are excessive and unreasonable, hence try to ignore or suppress them. Such obsessions/compulsions are time consuming (taking >1 hour a day) and as a consequence of this condition the affected person often takes longer to complete tasks and may have been significantly affected in terms of the their normal routine, occupational/academic functioning, or usual social activities and relationships (140). The APA (American Psychiatric Association) provides a diagnostic index of OCD subtypes. The categories include: cleaners, repeaters, completers, checkers, overly meticulous, compulsive avoiders, hoarders, and slowness (320).

The prevalence of OCD is estimated as 1-3% of the world-wide population (104)(105). Although the disorder may occur at any age, individuals from childhood to early adulthood may be at the greatest risk of developing the disorder (108)(109). When including other neuropsychiatric disorders that may share aetiology with OCD, such as body dysmorphic disorder (BDD), Tourette syndrome (TS), trichotillomania (TTM; compulsive hair-pulling), and dermatillomania (compulsive skin-picking [CSP]) (321)(322), OC-spectrum disorders may affect 10% of the general population (323). According to the World Health Organization, OCD was estimated as being the 11th leading cause of non-fatal burden in the world, accounting for 2.2% of total Years Lived with Disability (YLD), which is about the same level as schizophrenia (324). Medical treatment including pharmacological and behavioural therapies may partially alleviate the symptoms, but the cause of the disorder is largely unknown, and cure is rare.

5.2.1.2 Neurobiology

The neurobiological basis of OCD has been proposed mainly based on pharmacological studies and neuroimaging: (*i*) although partially effective, current drug treatments such as selective serotonin reuptake inhibitors (SSRIs) have demonstrated efficacy in reducing OC-symptoms among 40-60% affecteds (325)(326), and (*ii*) brain imaging studies have identified specific brain regions to be highly active (e.g. metabolic activity and blood flow) among OCD patients (327).

Several imaging studies have repeatedly reported an elevated glucose metabolism (328)(329)(330) and perfusion (331)(332)(333) in the orbitofrontal cortex, anterior cingulate cortex, caudate, and/or thalamus of OCD patients. In addition, other imaging studies that examined the effect of medical treatments for OCD in brain region activity have shown decreased glucose metabolism and perfusion in the same regions. The treatments that showed such effects include drug therapies (e.g. SSRIs or serotonin reuptake inhibitor) (330)(334)(335)(336)(337), behavioural therapy (335)(338), and neurosurgical lesion (339). Alterations in the level of (most likely glutamate) neurotransmitters (340) of the same brain regions were also observed among OCD patients using an independent technology (341)(342)(343). Taken together, these brain regions that are commonly affected among OCD patients implicate the cortico-striatal-thalamo-cortical (CSTC) circuit in the manifestation of OCD (Figure 5.1).

The CSTC circuit is well described in literature for its role and mechanism in controlling selection and exclusion of behaviours (345). Cortical neurons are densely projected to the striatum, forming glutamatergic cortico-striatal synapses onto medium spiny neurons (MSNs). MSNs then connect to the main output architecture of the basal ganglia, the globus pallidus pars internalis (GPi) and the substantia nigra pars reticulata (SNr). This connection between the input and the output structure of the basal ganglia may be achieved by two distinct, complementary pathways: the "direct pathway" formed by dopamine receptor type 1 (D1R)-expressing MSNs (blue pathway in Figure 5.1), and the "indirect pathway" by dopamine receptor type 2 (D2R)-expressing MSNs, whose signal reaches the GPi/SNr via the globus pallidus pars externalis (GPe), and the subthalamic nucleus (STN) (red pathway in Figure 5.1) (346). The output structure (GPi/SNr) of the basal ganglia then projects inhibitory signal to the thalamus that is regulated by the balance between the direct and indirect pathways (347).

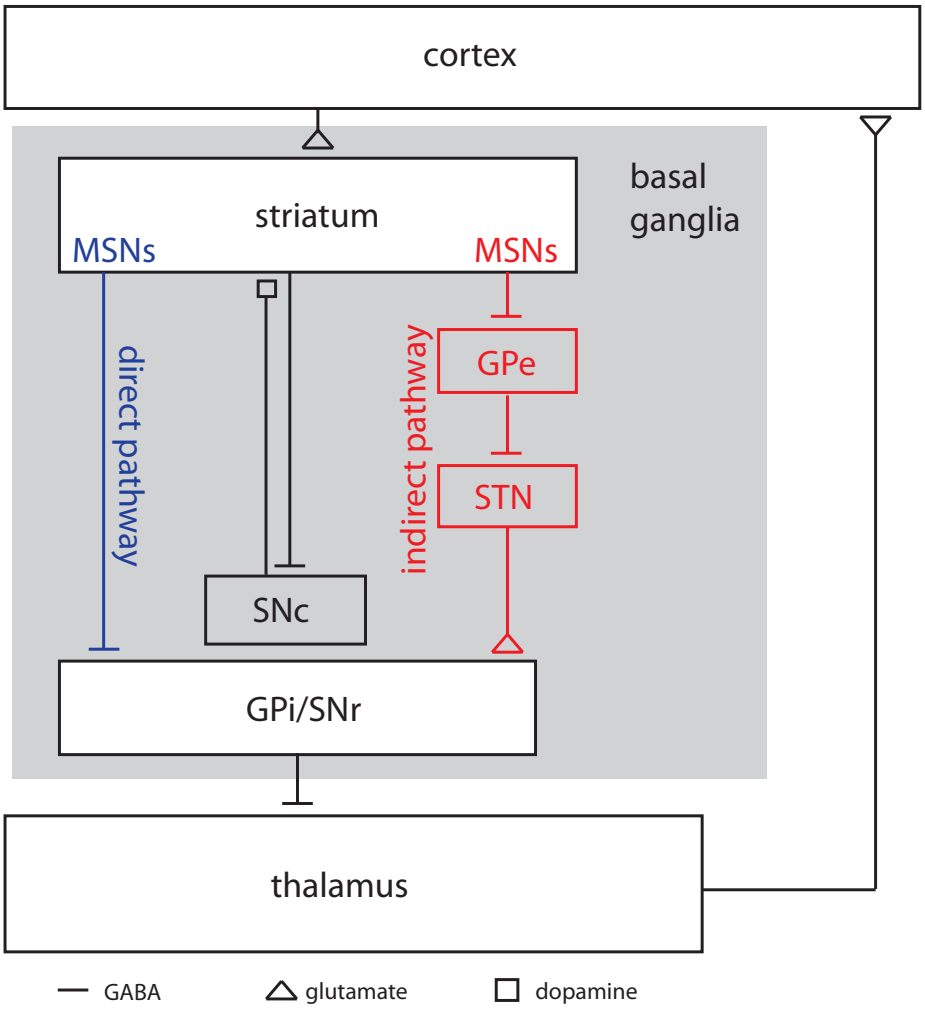


Figure 5.1: Diagram of the cortico-striatal-thalamo-cortical (CSTC) circuit - See in-figure legend for abbreviations and keys for neurotransmission systems. Detailed mechanism is described in the main text. Adapted from (344).

Depending on the inhibitory [via the direct pathway] or "dis"inhibitory signals [via the indirect pathway] from the basal ganglia, the thalamus inhibits or promotes the selection of behavioural sequences. The thalamus then sends excitatory signals back to the cortex, completing the CSTC loop (348) (Figure 5.1).

Current neuropharmacological evidence, mainly based on the efficacy of SSRIs, suggests an implication of the serotonergic systems in OCD (325)(326). Notably, the serotonin receptor subtypes and transporters implicated in OCD are most highly expressed in the ventral striatum (349), in which the serotonergic receptors and transporters may influence the CSTC circuit function. However, given the limited efficacy of the agents modulating the serotonergic system in OCD patients, other neurotransmitter systems such as dopaminergic and glutamatergic systems in the CSTC circuit are also likely to play roles in OCD (350) (Figure 5.1).

5.2.1.3 Genetics

Multiple lines of evidence suggest that OCD has a strong genetic basis. The concordance rate for OCD is 80-87% in monozygotic twins and 47-50% in dizygotic twins (112)(351). First-degree relatives are at 3-12 times greater risk than the general population (352). Despite >60 candidate gene publications, no genome-wide significant, replicated genetic associations with OCD have been achieved thus far (353)(354), which is likely to reflect the heterogeneous, complex nature of the genetic basis of OCD (see section 1.2).

While human genetic studies have not yielded much success, studying genetic animal models of OCD has been a productive approach. Mice with deletion of *Sapap3*, which encodes a postsynaptic scaffolding protein at the excitatory synapse, manifested compulsive grooming that was alleviated by SSRIs (203). Electrophysiological and biochemical studies of *Sapap3* knock-out mice revealed defects in cortico-striatal synapses, supporting the CSTC hypothesis of OCD. Sequencing exons and splice-sites of the human *SAPAP3* gene in 165 individuals with OCD and trichotillomania (TTM) revealed six nonsynonymous changes (355). In addition, a study on human patients with an OCD subtype, pathological grooming, identified an associated SNP of the *SAPAP3* gene (356). Finally, a genome-wide association study on a spontaneous animal model, canine compulsive disorder (CCD), revealed a strong association ($p = 7 \times 10^{-7}$, $p_{genome} = 0.01$)

of *CDH2*, the neural cadherin involved in synaptic plasticity, and other genes that are potentially linked to the CSTC circuit (357).

Given the compelling evidence for the CSTC neurocircuitry model of OCD and animal model studies, here I sought to reseqeuence >600 genes related to the CSTC circuit and those identified from dog and mouse studies among ∼200 OCD patients and ∼200 matched healthy controls, and to identify rare genetic variants contributing to OCD. For a polygenic disorder such as OCD, it is likely that not every patient shares the same mutation, but that a disruption at any point in the common neural pathway leads to a similar symptom (see section 1.2), hence the mutations are uncommon even in the affected population. Accordingly, I employed statistical tests that consider aggregated effects of the multiple variants and genes per functional unit, in order to identify the common biological functions underlying OCD.

5.3 Materials and Methods

The sample collection (Appendix Tables 5.10 and 5.11) and resequencing experiment were conducted by a collaboration between the Lindblad-Toh group (Broad Institute) and the Feng group (MIT). Here, I will describe the relevant experimental procedures as well as the analytical methods, in order to help to understand how the raw data were generated. Note that the work described in sections 5.3.1, 5.3.2, and 5.3.3 were conducted by Dr Ruqi Tang (Feng group), Dr Elinor Karlsson (Lindblad-Toh group), and the Broad Sequencing Platform.

5.3.1 DNA preparation and pooling

In this study, a pooled sequencing approach was employed to minimise the sequencing cost, while maximising the utility of the billions of DNA sequence data that are generated by each sequencing run. As I wanted to identify rare variants in pooled DNA samples, it was important that the study design provided sufficient power to detect rare variants. Currently, the standard error rate of most massively-parallel sequencing runs is 1-2%. Given this error rate, the number of chromosomes in each pool should not exceed 50, because otherwise the true single copy alleles would not be discerned from sequencing error. Hence, in this experiment 16 individuals were grouped, making 32 chromosomes per pool, in order to permit the identification of rare variants.

The pooled sequencing experiment requires a special preparation of the samples before the sequencing step, as described below: The samples of 201 affected individuals defined by rigorous clinical phenotyping (140) and 201 matched controls of European ancestry were obtained (Appendix Tables 5.10 and 5.11). Once the DNA was quantified and normalised, 16 samples of the same phenotype were pooled together using the Multiprobe or Packard Robotic, making a total of 24 pools that consist of 384 samples. Each pool was then barcoded (358), so that the case and the control pools are identifiable for the downstream analysis (Figure 5.3).

5.3.2 Target selection

Coding and non-coding regions of 611 genes (Appendix Table 5.9) were sequenced, 11.7% of which was coding sequence. The targeted genes were selected based on the following criteria: (*i*) the genes that encode postsynaptic proteins based on published

biochemical and mass spectrometry data (359)(360), (*ii*) genes that are highly expressed in the striatum based on published expression data and the Allen Brain Atlas (361), (*iii*) genes that are differentially expressed in the medium spiny neurons (MSNs) of the direct and indirect pathways within the basal ganglia (362), (*iv*) candidate genes from human association or linkage studies (363), (*v*) genes from canine and mouse genetic studies (364)(203)(357), and (*vi*) genes that are related to other neuropsychiatric disorders whose aetiology may be shared with OCD (260)(365). These regions of interest were targeted using the customised Nimblegen Hybrid Selection array (Figure 5.3).

5.3.3 Sequencing and mapping

Libraries were constructed by a modified Illumina paired-end library protocol and PCR enrichments. The first step of the library preparation was DNA fragmentation using Covaris^{TM}. The fragmented DNA samples were then enriched for the target region using the Nimblegen hybrid capture array. Subsequently, the prepared DNA was sequenced in 16 lanes, each of which contains 12 pools, of the Illumina's HiSeq 2000 machines at the Broad Sequencing Platform (Figure 5.3). The resulting 76bp-long reads were aligned to the reference genome (hg19) using the MAQ algorithm within the Picard analysis pipeline (http://picard.sourceforge.net/), and further processed using the GATK pipeline (Figure 5.2 and Phase 1 of Appendix Figure 5.20) (366)(367). The quality score, Q, was calculated for quality control. Q is logarithmically linked to error probabilities: when the probability of an inaccurate base call is 1 in 10 (i.e. accuracy of base call is 90%), Q is denoted as Q10, and when the probability is 1 in 100 (i.e. 99% base calling accuracy), it is denoted as Q20, and so forth. Typically, alignments with Q20 are considered to be of high quality, hence Q20 alignments were kept for downstream analysis.

5.3.4 Variant discovery and allele frequency estimation

Once the raw sequencing data were processed via the Picard pipeline (Phase 1 of Appendix Figure 5.20), I employed SYZYGY to call variants and estimate their allele frequencies (368)(369). SYZYGY is an analysis tool that addresses various issues arising from pooled sequencing, where each read does not retain the identity of its origin. SYZYGY computes a likelihood that each genomic position contains a non-reference allele, applies filters based on *(i)* the error rates considering the strand consistency

(whether the numbers of + and - strands for an allele are comparable), *(ii)* sequence context (di- or tri-nucleotide alleles tend to have different error estimates), and *(iii)* relative depth of coverage (a drop in coverage, arising from bases in otherwise high quality read, predicts a reduced accuracy at those bases), and provides estimated allele frequencies of the detected variants that are used for downstream analysis (369) (Figure 5.2).

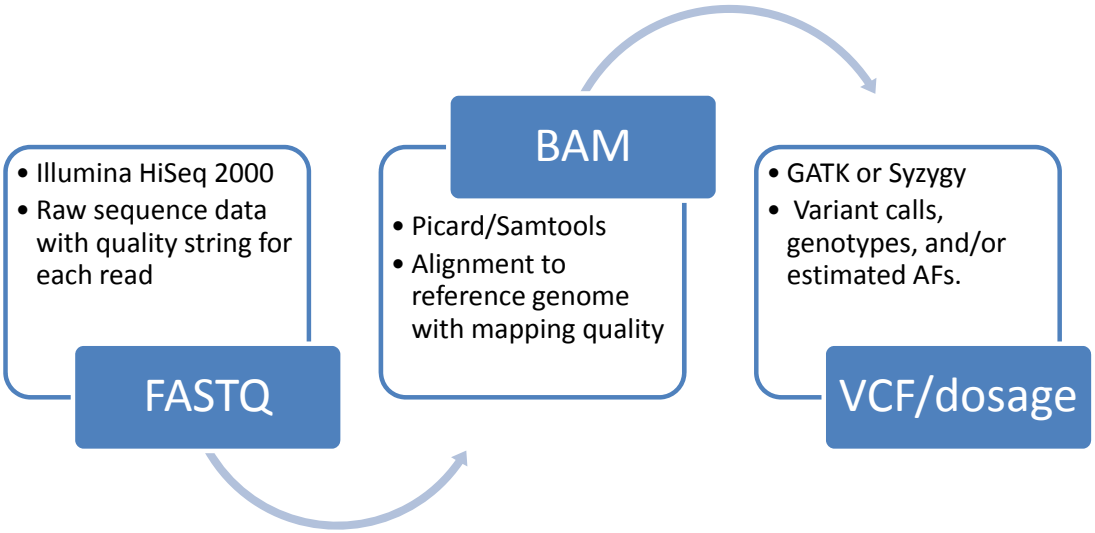


Figure 5.2: Analysis steps and relevant file formats for next generation sequencing (NGS) data - Illumina HiSeq 2000 produces raw sequence data with quality strings in FASTQ file format. An aligner program (e.g. Picard or Samtools) then takes the FASTQ file and generates alignments to a reference genome with mapping quality scores in BAM file format. BAM files are then taken either into GATK for the sequencing data where the individual sample's identity is available for each read (see Chapter 6), or into SYZYGY for the pooled sequencing data. GATK produces variant calls and genotypes in VCF format, while SYZYGY outputs variant calls and estimated allele frequencies (AFs) in a dosage file.

5.3.5 Pathway analysis

I used Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 (370) to identify significantly enriched Gene Ontology (GO) terms among the genes of interest. In order to increase the power of the test, a subset of the full GO terms (GO FAT) was used (370). Unlike GO Slim, which consists of a subset of the broadest GO terms based on the terms' level-specificity (http://www.geneontology.org/GO.slims.shtml),

GO FAT determines the term-specificity based on the number of child terms, and keeps more specific terms rather than the broadest terms. This way, one can keep the specificity of the examined GO terms, while increasing the power by reducing the number of tested GO terms. In addition, the term-specificity based on the number of child terms can universally describe the specificity of a given GO term, while the level information cannot, due to the variable levels of the hierarchical structure for each GO term. The examined genes include those that were significantly associated ($p_{perm} < 0.05$) with the case cohort resulting from the nine association test categories (see below).

5.3.6 Association tests

For the single variant association test, the t-test was employed to examine whether allele frequencies for each variant in case pools differ from those in control pools. For the groupwise (a.k.a. gene-based) association test, the differences in allele frequencies between the case and the control population were summed for each gene, and the sum of the differences were compared with the expectation by 1,000 times permutation of case and control labels, to compute the p_{perm}-values. Nine categories of variants, [namely, *(i) all*, *(ii) non-synonymous (ns)*, *(iii) coding*, *(iv) conserved*, *(v) conserved & ns* (i.e. nonsynonymous variants in conserved region), *(vi) MAF*$<$ 10%, *(vii) MAF*$<$ 1%, *(viii) MAF*$<$ 10% *& conserved* (i.e. variants with MAF$<$ 10% in conserved elements), and *(ix) MAF*$<$ 1% *& conserved* (i.e. variants with MAF$<$ 1% in conserved elements], were grouped for each gene and tested separately. In total, the nine test categories resulted in 123 genes that are significantly associated with any of the test. These 123 genes are herein termed "investigative" genes. Despite the potential decrease in the significance of association when combining the results from multiple tests, I considered all of the 123 genes for pathway analysis, in order to reduce the number of false negative results which may be particularly high when examining a sub-population of individuals with a complex disorder where the disease-contributing mutations may be distinct from one another. For the power calculation, the same number of pools were sub-sampled from each case and control cohort, followed by the groupwise association test.

5.4 Results

Targeted resequencing of ∼13Mb candidate genomic regions for obsessive-compulsive disorder (OCD) in 192 affected individuals and 192 matched healthy controls yielded: *(i)* high-quality sequencing alignments of >2.4 billion reads, *(ii)* >90,000 genetic polymorphisms and variants enriched for rare variants, and *(iii)* highly accurate allele frequency (AF) estimates [that were confirmed by the comparison with 355 genotypes from an independent experiment; correlation coefficient $rho = 0.99$, $p < 2.2 \times 10^{-16}$].

The comparison between allele frequencies of detected variants in cases and controls revealed that the most strongly associated variants reside in non-coding sequence. Of the top ten most associated variants, two variants, whose genomic positions are evolutionarily conserved, reside in the 3'-untranslated region (3'-UTR) of either *THBD* (Chr20:23028063 T→C, $p = 0.00002$) or *CLDN1* (Chr3:190025964 C→T, $p = 0.00062$). Examining the collective difference in all of the genic variants' allele frequencies between cases and controls identified 36 significantly associated genes ($p_{perm} < 0.05$). 8 additional tests that considered the types of mutation and allele frequency of variants were also performed, resulting in 87 additional genes. Altogether, the groupwise association tests yielded 123 "investigative" genes for OCD. Pathway analysis on these investigative genes yielded interesting functional categories including telencephalon development, regulation of cellular component size, and regulation of neurogenesis.

Lastly, in order to gain insights whether sequencing of 200 cases and 200 additional controls would be worthwhile or not, I performed the groupwise association tests on the varying size of sub-sampled data, an extrapolation of which suggested that a doubled sample size may improve the associations.

5.4.1 Pooled targeted sequencing yielded large amounts of high-quality sequence alignments

This study employed state-of-the-art technologies including hybrid selection (HS) and next generation sequencing (NGS) to resequence ∼13Mb candidate regions for OCD in 192 cases and 192 matched controls (Appendix Tables 5.10 and 5.11). The DNAs from the cohort were combined into 24 pools that were then individually barcoded, each of which containing 16 individuals (Figure 5.3). The pooled DNAs were then captured for the target regions, which encompass 611 genes related to the CSTC circuit and those

suggested from the dog and mouse studies (Appendix Table 5.9), using Nimblegen's customized hybrid capture array. The captured DNAs were then sequenced using the Illumina HiSeq 2000 in 16 lanes of 12 pools (see Materials and Methods).

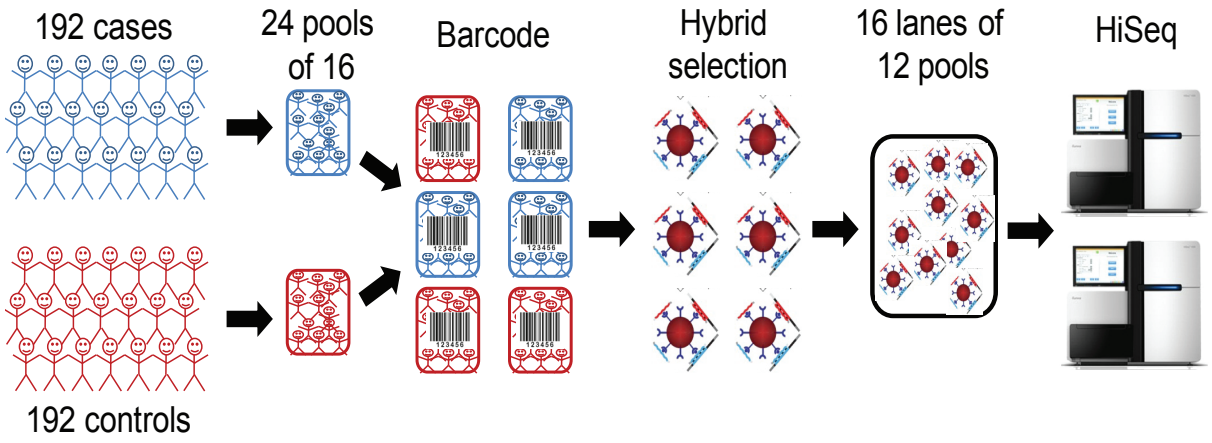


Figure 5.3: Experimental design for targeted pooled resequencing - The samples of 192 cases and 192 controls were pooled into 24 pools of 16 individuals, resulting in 12 case pools and 12 control pools. Each pool was then barcoded, and hybrid selection was performed to target approximately 13Mb genomic regions of interest. Those captured genomic regions were then sequenced using the Illumina HiSeq 2000 machine.

As a result, a total of 60-170 million reads per pool were generated (Appendix Figure 5.28). Of those, 53-156 million reads for each pool had a mapping quality of Q20 (i.e. 1/100 chance of wrong alignment) or higher (Table 5.1; see Appendix text 5.6.1 for how duplicated reads were handled.). Also, 4-11 billion bases were confidently called at >99% accuracy (*PF HQ Aligned Q20 Bases* in Table 5.1). 21 of 24 pools achieved a coverage of >90% of target bases at 20x (Table 5.1), and the mean coverage excluding three low-quality pools was 89x per pool and 5.5x per individual (Table 5.2 and Figure 5.4). 93% of all targets had a coverage of at least 2x per individual, and only 0.05% of the total target bases had no coverage (Table 5.2).

Sample	Mean Read Length (bp)	PF HQ Aligned Reads	PF HQ Aligned Q20 Bases
Case1	76	124,154,961	8,985,945,434
Case2	76	110,282,143	7,956,278,297
Case3	76	121,218,361	8,796,043,814
Case4	76	121,643,316	8,815,806,624
Case5	76	95,420,508	6,918,095,998
Case6	76	107,387,912	7,811,734,896
Case7	76	52,787,870	3,891,794,707
Case8	76	86,622,361	6,345,670,171
Case9	76	155,846,389	11,429,784,858
Case10	76	100,263,856	7,342,435,735
Case11	76	92,661,278	6,797,899,807
Case12	76	82,830,923	6,073,388,370
Control1	76	109,909,194	7,960,889,020
Control2	76	85,821,056	6,221,128,777
Control3	76	101,972,105	7,431,952,519
Control4	76	128,232,811	9,337,744,511
Control5	76	94,025,184	6,765,450,800
Control6	76	85,009,815	6,169,560,062
Control7	76	103,836,929	7,614,317,692
Control8	76	98,880,620	7,265,600,981
Control9	76	118,016,608	8,661,705,310
Control10	76	57,602,355	4,219,177,484
Control11	76	100,941,431	7,411,078,780
Control12	76	79,443,170	5,818,763,773

Table 5.1: Sequencing statistics showing read length and number of high quality (HQ) aligned reads and bases for each pool - *Sample* column indicates each of the 12 case and 12 control pools. *Read Length* is defined as the number of bases identified contiguously generated from the sequencing method. *PF HQ Aligned Reads* indicates the number of PF reads (i.e. passing Illumina's filter) that were aligned to the reference sequence (hg19) with a mapping quality of Q20 or higher (i.e. reads with 1 in 100 [or less] chance that the alignment was incorrect). *PF HQ Aligned Q20 Bases* displays the number of bases aligned to the reference sequence in reads that were mapped at high quality (>Q20). This is close to (*PF HQ ALIGNED READS*)×(*READ LENGTH*), but can differ when either mixed read lengths are present or many reads are aligned with gaps.

Sample	Mean Target Coverage (read depth)	Mean Target Coverage per individual (read depth)	PCT Target Bases 2x	PCT Target Bases 20x
Case1	39.9	2.5	98.6	79.1
Case2	36.2	2.3	98.7	75.2
Case3	73.2	4.6	99.1	94.5
Case4	65.2	4.1	99.0	94.0
Case5	93.3	5.8	99.1	96.1
Case6	69.2	4.3	99.0	93.9
Case7	101.4	6.3	98.7	96.6
Case8	77.6	4.8	98.9	96.5
Case9	128.5	8.0	99.1	97.6
Case10	84.2	5.3	98.9	96.3
Case11	110.6	6.9	98.9	96.9
Case12	90.2	5.6	98.9	96.8
Control1	104.0	6.5	99.1	96.8
Control2	84.5	5.3	99.0	95.8
Control3	66.4	4.1	98.9	94.2
Control4	81.6	5.1	99.1	95.8
Control5	29.5	1.8	98.6	63.9
Control6	52.8	3.3	98.8	90.3
Control7	123.3	7.7	99.0	97.1
Control8	114.4	7.1	98.9	96.9
Control9	127.7	8.0	99.0	97.1
Control10	60.4	3.8	98.7	94.8
Control11	100.1	6.3	98.9	96.8
Control12	62.3	3.9	98.7	94.9

Table 5.2: Target coverage information per pool - The table shows the mean target coverage per pool (*Mean Target Coverage*), the mean target coverage per individuals in each pool (*Mean Target Coverage per individual*), and % targeted bases with read depth of at least 2x (*PCT Target Bases 2x*) and % targeted bases with read depth of at least 20x (*PCT Target Bases 20x*).

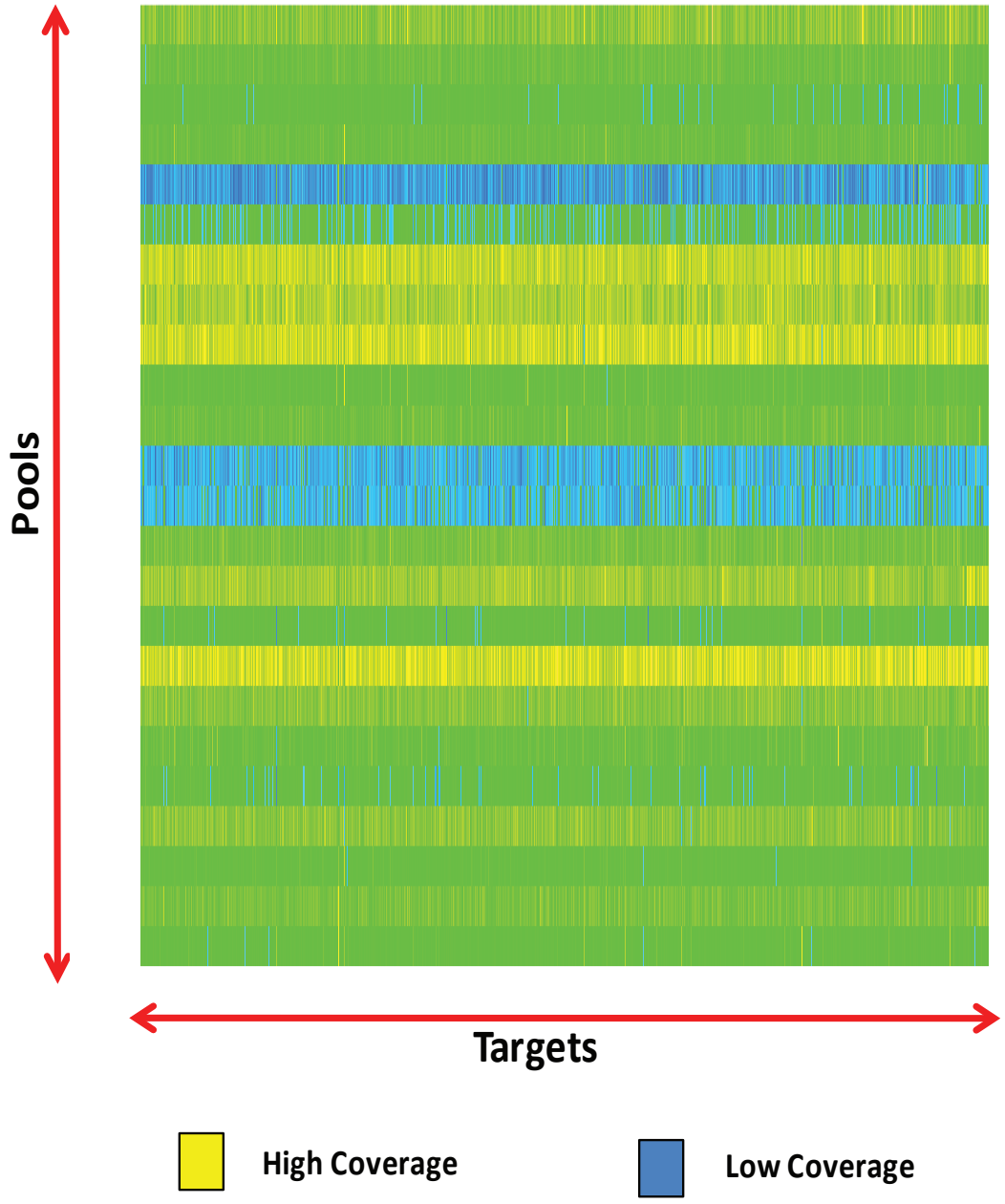


Figure 5.4: Heatmap for read depth of target region per pool - The heatmap shows the read depth for each target across all 24 pools. The x-axis indicates 31,471 segments of the targeted sequence while the y-axis represents 24 pools. The read depth is shown as a colour spectrum, starting from low coverage in blue to deep coverage in yellow.

5.4.2 SYZYGY identified a number of functional genetic variants and polymorphisms

Using these high-quality alignments, I then aimed to identify genetic variants including single nucleotide variants (SNVs) and short insertions/deletions (indels) from the sample cohort. Using SYZYGY, I identified 96,487 variants in the case and control cohort, and of these, 88,021 variants passed the quality control (see Materials and Methods). This quality control improved the Transition/Transversion (Ti/Tv) ratio from 1.83 (50,112/27,420) to 2.03 (47,974/23,649), as well as the proportion of detected variants found among previously reported SNPs, from 62% to 66% in dbSNP 132 and 56% to 62% in the 1000 genomes project. Of 88,021 QC+ variants, there were 38 nonsense mutations, 2,635 missense mutations, and 2,421 synonymous mutations, in which 37% (1,892/5,094) were novel. Almost 7,000 additional variants were identified in the splice sites, 3'-UTR, or 5'-UTR, 35% (2,405/6,924) of which were novel (Table 5.3).

Variant type	# QC+ variants	# Novel* variants
Nonsense	38	27
Missense	2,635	1,189
Splice	28	21
Synonymous	2,421	676
3'-UTR	6,198	2,109
5'-UTR	698	275
Total	88,021	29,797

*Not present in dbSNP 132

Table 5.3: Functional annotations of detected variants - Among 88,021 QC+ variants detected from the cohort, 34% (29,797/88,021) were novel. 14% (12,018/88,021) of the total QC+ variants had functional types of mutations, of which 36% (4,297/12,018) were novel.

Notably, 94% (82,965/88,021) of the QC+ variants were identified in non-coding sequence, including 16,482/82,965 (19.9%) variants that reside within the evolutionarily conserved elements as defined in (3). The identification of a large number of non-coding variants was perhaps because a majority (88.3%) of the target region was non-coding sequence, and non-coding bases tend to show more polymorphisms than coding bases because of a lower degree of selection on average.

5.4.3 Comparison with genotyping array confirmed the accuracy of allele frequency (AF) estimation

The pooled targeted sequencing employed here is different from the conventional sequencing strategy in that the reads do not retain the identity of the individual from whom the reads originated. Thus, I was only able to distinguish the reads by each pool, but not by each individual. In other words, the allele frequencies of the detected variants could be only estimated, rather than measured. For the accurate estimation of allele frequency, I employed an analysis tool, SYZYGY, which was especially designed for pooled, targeted resequencing studies (368) (see Materials and Methods).

In order to assess SYZYGY's performance, I compared SYZYGY's allele frequency estimation with the actual allele frequencies measured from an independent SNP genotyping experiment for the same individuals. The allele frequencies of the 355 SNPs that were both estimated from the pooled sequencing data and measured from the SNP array were strongly correlated (Spearman's correlation coefficient=0.99, $p < 2.2 \times 10^{-16}$; Figure 5.5), confirming that the estimated allele frequencies are reliable.

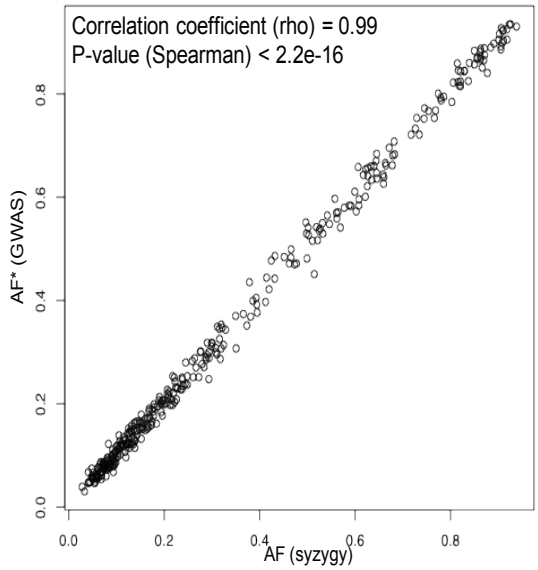


Figure 5.5: Comparison of AFs between pooled sequencing data and SNP array data - Allele frequencies that were estimated by SYZYGY (x-axis) were highly concordant with the actual allele frequencies obtained from SNP array data (y-axis).

5.4.4 Many of the detected variants were rare

Among the detected variants, I observed an excess of rare (minor allele frequency [MAF]<1%) and low-frequency (1%<MAF<5%) variants (panel A in Figures 5.6 and 5.7). More specifically, 32.7% (745,788/2,315,105) of the total allele frequency estimates for all 24 pools showed MAF<1%, and 17.3% (399,982/2,315,105) were 1%<MAF<5%. Taken together, 50% of the estimated allele frequencies for detected variants were found to be uncommon, which would have been undetectable by the popular genome-wide association study (GWAS) method that uses common SNPs.

Subsequently, I surveyed the types of variants across the allele frequency spectrum. Classifying coding and non-coding variants for each allele frequency bin revealed that the ratio of the number of coding:non-coding variants stayed relatively constant across different allele frequencies, at 4-10% of the variants being coding (panel B in Figure 5.6).

When considering the evolutionary conservation of the genomic positions (3) which may be an indication of the functional proportion of the genome (see section 1.1), the percentage of non-coding variants within conserved elements relative to the total detected variants for each allele frequency bin was ∼20% (panel B in Figure 5.7). Despite some variation (12-24%), this percentage stayed relatively constant across allele frequencies >1%. Among rare variants (MAF<1%), however, there was an increase in the proportion of non-coding variants within the constrained regions relative to other variants (20→30%). This surfeit of rare, non-coding variants within conserved regions might reflect the presence of many disease-contributing mutations that are rare, regulatory non-coding variants (see section 1.1). Indeed, all of the most strongly associated variants were found to be non-coding variants in later analyses presented below (see section 5.4.5).

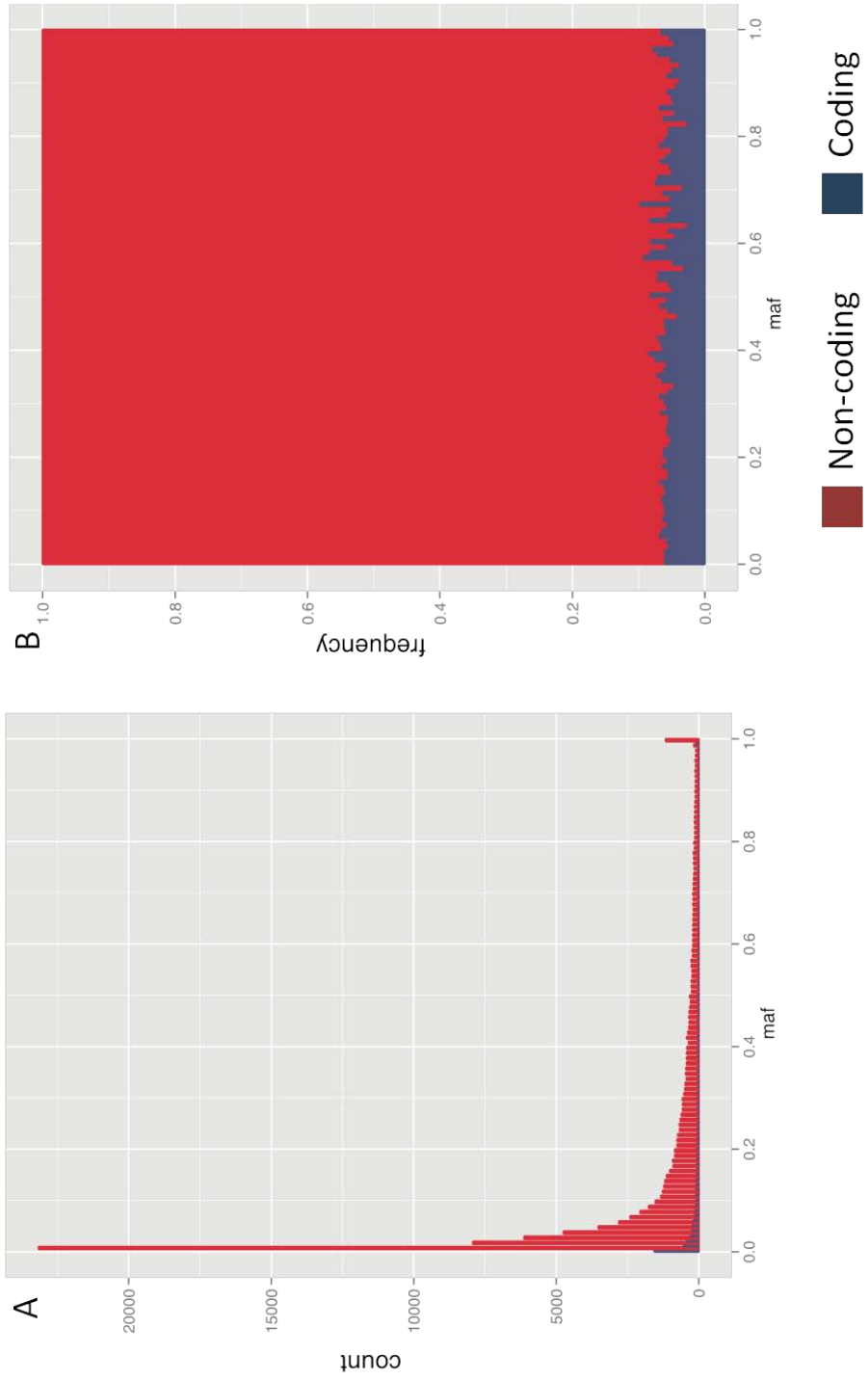


Figure 5.6: Distribution of coding and non-coding variants across different AFs - A. The x-axis shows allele frequencies within the cohort population, while the y-axis shows the number of detected variants that fall into each allele frequency bin. B. The x-axis shows allele frequencies in the cohort population, while the y-axis indicates the fraction of the coding and non-coding variants that constitute each allele frequency bin.

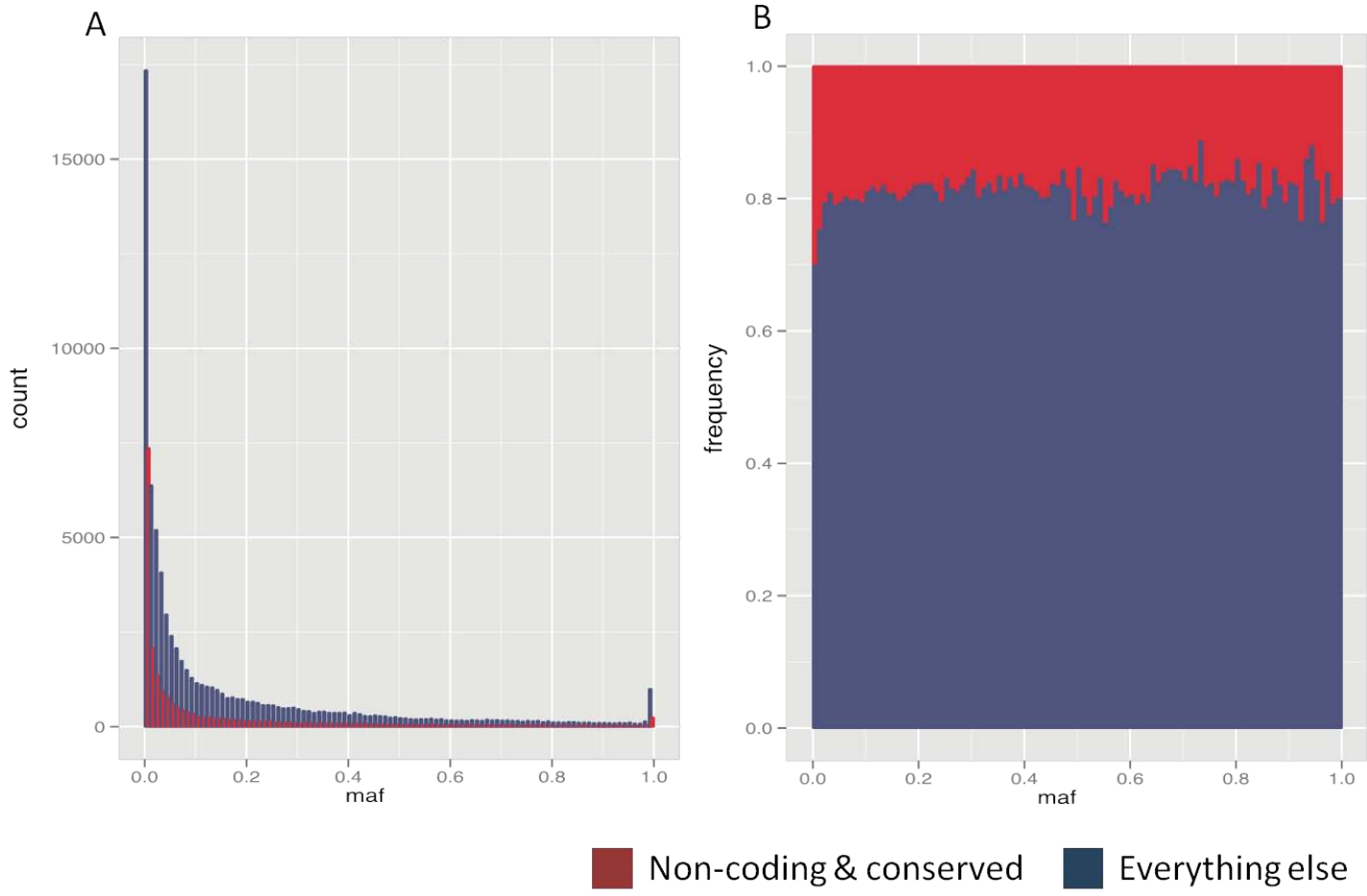


Figure 5.7: Distribution of non-coding variants within evolutionarily conserved elements across different AFs - A. The x-axis shows allele frequencies in the cohort population, while the y-axis shows the number of detected variants that fall into each allele frequency bin. B. The x-axis shows allele frequencies in the cohort population, while the y-axis indicates the fraction of coding and non-coding variants that constitute each allele frequency bin.

5.4.5 Case-control comparison of AFs yielded strongly associated variants, genes, and functional themes for OCD

Using the allele frequencies of the variants estimated from the pooled sequencing experiment, I then sought to compare allele frequencies between cases and controls, in order to identify variants that are significantly more common in the case cohort than in the controls.

In order to permit such comparison, I first checked whether there is any population stratification between the case and the control samples that may lead to an artefactual association that reflects the genetic difference arising from different population ancestries rather from that related to disease state. As expected from the sample collection, where the cases and the controls were matched (Appendix Tables 5.10 and 5.11), the genetic markers (genome-wide data) of these individuals showed no evidence of population stratification between case and control cohorts (Appendix text 5.6.2).

Next, I sought to check whether there was any bias in allele frequencies between cases and controls that might be either a true difference, or an artefact that occurred during the sequencing experiment and/or variant calling. I compared the distribution of allele frequencies between cases and controls and found that the two distributions do not differ (Appendix text 5.6.3). This confirmed that there was no bias in the allele frequencies of detected variants between cases and controls (e.g. an abundance of rare variants in cases), and eliminated the possibility of an artefactual skew in the allele frequencies.

Given that the case-control allele frequency data were confirmed as being usable for association tests, I performed the genetic association tests for OCD based on either:

- allele frequency difference for each variant between cases and controls, or
- aggregated allele frequency differences of multiple variants per gene between cases and controls.

When testing the allele frequency difference for a single variant, more low-frequency variants were observed among highly associated variants (Figure 5.8). However, the outnumbered associations among the low-frequency variants were not more than those expected by random chance, given the excess of the low-frequency variants among the total detected variants (panel A of Figures 5.6 and 5.7).

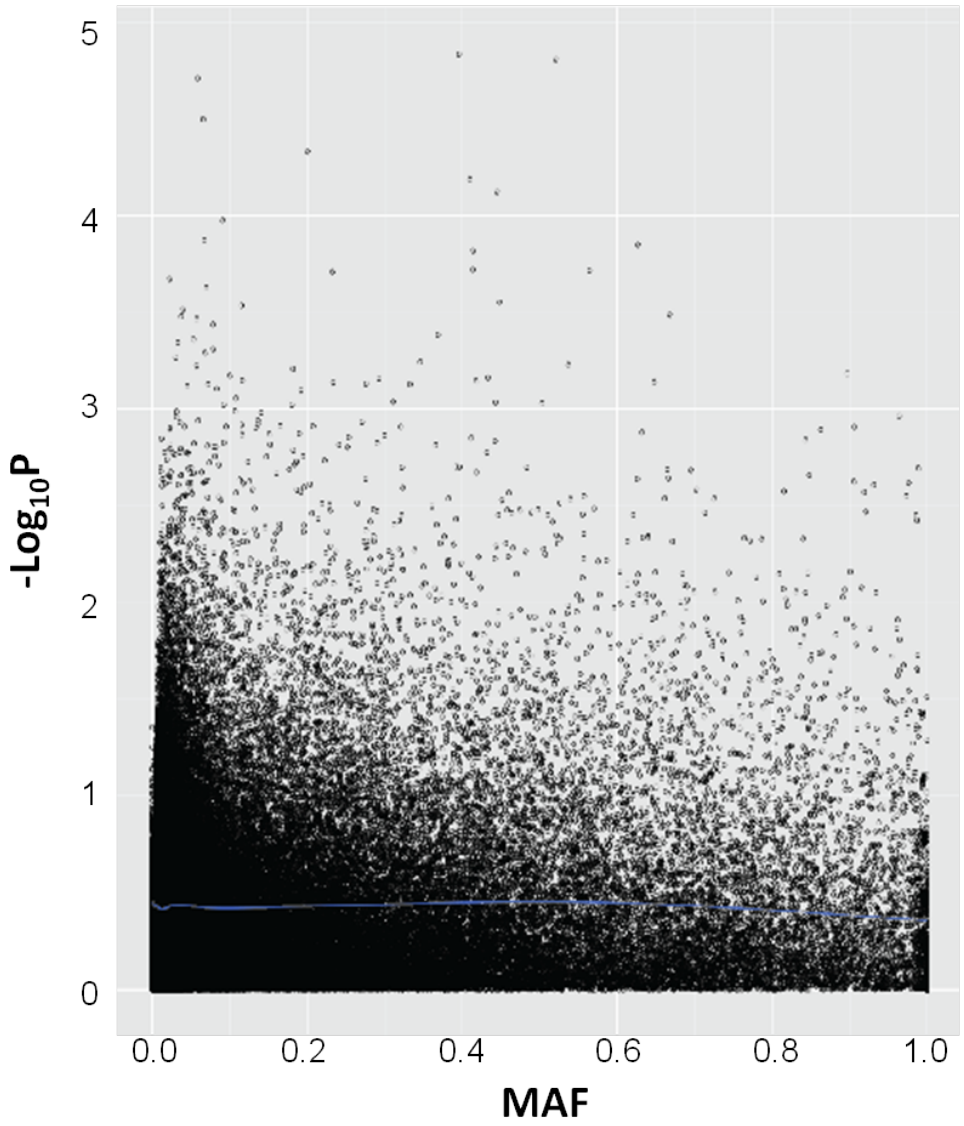


Figure 5.8: Distribution of the association levels across different AFs - Each dot represents a detected variant, for which its position is determined by its allele frequency (x-axis) and $-log_{10}P$ of case-control association test (y-axis). The almost horizontal regression line (blue line) suggests that there is no obvious bias in the association level across different allele frequencies.

On the other hand, when considering the functional annotations of detected variants, I observed an interesting pattern in the distribution of the variant types across different association levels. The ratio of [coding]:[non-coding]:[non-coding variants in conserved elements] stayed relatively constant in the lower level associations (i.e. [coding]:[non-coding]:[non-coding variants in conserved elements] = 3-9%:91-97%:17-26% in the association level $0 < -log_{10}P < 4$; Figure 5.9), whereas the most strongly associated variants ($-log_{10}P > 4$) were all non-coding variants (Figure 5.9). The dominance of the non-coding variants among the most associated variants is again consistent with the hypothesis that regulatory non-coding variants may underlie OCD (see section 1.1.3).

association

-log p	coding		non-coding		non-coding & conserved	
0.5	4418	6%	64282	94%	13480	20%
1.0	1227	6%	18828	94%	3801	19%
1.5	316	6%	5232	94%	1013	18%
2.0	84	5%	1447	95%	276	18%
2.5	20	4%	445	96%	80	17%
3.0	4	3%	137	97%	27	19%
3.5	3	9%	31	91%	9	26%
4.0	0	0%	12	100%	0	0%
4.5	0	0%	4	100%	0	0%
5.0	0	0%	3	100%	1	33%

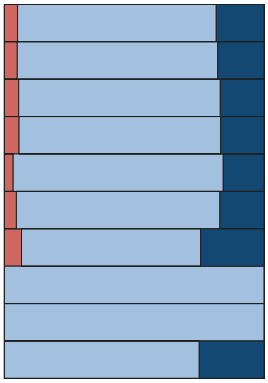


Figure 5.9: The most strongly associated variants are non-coding - The first column indicates the level of association with the OCD cohort. The larger the $-log_{10}P$ value, the stronger the association. The second column indicates the number of coding variants and its percentage to the total number of variants for each p-value indicated in the first column. The third and fourth columns show the same information for non-coding variants and for non-coding variants in conserved elements, respectively. The bar graph on the right represents the proportion of coding variants (red), non-coding variants that are not in conserved region (light blue), and non-coding variants within conserved region (dark blue) relative to the total number of detected variants for the corresponding p-values.

5.4.5.1 Strongly associated variants may have functional effects

Although there was no apparent evidence that the association levels achieved from the single variant tests were unexpectedly higher than the expectation from uniformly distributed p-values (Appendix text 5.6.4 and Appendix Figure 5.23), I still went on

to investigate the top associated variants as a pilot analysis for a future experiment with an increased sample size (see section 5.4.6).

For the functional investigation, I selected the seven most strongly associated coding variants ($-log_{10}P > 3$) and the top ten non-coding variants in conserved elements ($-log_{10}P > 3.5$). Of the top seven associated coding variants, two were nonsynonymous mutations: Chr8:41566438 C→T variant ($p = 0.0005$; allele count in case [AC_{case}]=4, allele count in control [$AC_{control}$]=19) that changes *ANK1* (ankyrin 1) gene's residue 1954 from arginine (R) to histidine (H) (herein termed R1954H) and Chr3:196888560 A→G variant ($p = 0.0017$; AC_{case}=10, $AC_{control}$=21) that replaces *DLG1* (discs, large homolog 1) gene's residue 532 valine (V) with alanine(A) (herein termed V532A) (Table 5.4).

Chr	BP	REF/ALT	Residue	AC_{case}	$AC_{control}$	GENE	P-value
chr8	41566438	CT	R1954H	4	19	*ANK1*	5.00×10^{-4}
chr17	56396587	TC	E1561E	32	21	*BZRAP1*	5.00×10^{-4}
chr9	4576680	TC	T1108T	76	120	*SLC1A1*	7.00×10^{-4}
chr3	196888560	AG	V532A	10	21	*DLG1*	1.70×10^{-3}
chr9	124083614	CT	G1411G	23	7	*GSN*	2.40×10^{-3}
chr22	18900868	GA	A1621A	8	22	*PRODH*	2.80×10^{-3}
chr12	63544209	AG	F406F	191	184	*AVPR1A*	2.80×10^{-3}

Table 5.4: Top associated coding variants - This table shows the top seven coding variants with $-log_{10}P > 3$. The *Chr* and *BP* columns represent the genomic position of each variant. In the *REF/ALT* column, the first and the second characters represent the reference and each variant's (alternative) alleles, respectively. The *Residue* column shows the original residue, the location of the residue, and the new residue changed by the corresponding variant. The AC_{case} and the $AC_{control}$ columns show the allele counts (AC) in each case and control population. The *GENE* column indicates the genes in which the variants reside. The *P-value* column shows the probability obtained from a t-test.

Among the top ten non-coding variants in conserved elements, three were residing within 3'-UTRs of protein-coding genes. Of these three variants, two were present at higher allele frequencies in cases than in controls. Although the variant with higher frequency in the control population might be of interest as a protective allele, here I focused on the variants with higher frequency in cases in order to identify causative mutations. Chr20:23028063 T→C variant ($p = 0.00002$; AC_{case}=28, $AC_{control}$=11) was located within the 3'-UTR of *THBD* (thrombomodulin) gene, where the reference allele T was evolutionarily highly conserved (Table 5.5, see section 5.5.2). Chr3:190025964 C→T variant ($p = 0.00062$; AC_{case}=72, $AC_{control}$=51) overlapped with the 3'-UTR of

CLDN1 (claudin 1) gene, where the variant's genomic position was also well conserved across multiple species (Table 5.5, see section 5.5.3).

Chr	BP	R/A	AC_{case}	$AC_{control}$	GENE	P-value	Cons.
chr20	23028063	TC	28	11	*THBD* (3'-UTR)	2.00×10^{-5}	+++
chr9	34502037	GA	17	28	*DNAI1*	5.10×10^{-4}	+
chr7	69837848	GA	26	13	*AUTS2*	5.90×10^{-4}	-
chr3	190025964	CT	72	51	*CLDN1* (3'-UTR)	6.20×10^{-4}	+++
chr2	31240899	GC	120	160	*GALNT14*	7.20×10^{-4}	+
chr3	143283337	CT	190	245	*SLC9A9*	7.20×10^{-4}	-
chr3	173306684	CD	21	10	*NLGN1*	7.60×10^{-4}	+
chr5	78742980	GA	157	140	*HOMER1*	9.20×10^{-4}	-
chr11	115045237	TC	148	190	*CADM1* (3'-UTR)	9.30×10^{-4}	+
chr6	154568823	GD	53	68	*IPCEF1*	9.50×10^{-4}	++

Table 5.5: Top ten non-coding variants in conserved region - This table shows the top ten non-coding variants in conserved region with $-\log_{10}P$ >3.5. The *Chr* and the *BP* columns represent the genomic position of each variant. In the R/A column, the first and second characters represent the reference and each variant's (alternative) alleles, respectively. The AC_{case} and the $AC_{control}$ columns show the allele counts (AC) in each case and control population. The *GENE* column indicates the closest genes to the variant's genomic position. The *P-value* column shows the level of association with the case population. The *Cons.* column shows the degree of evolutionary conservation at the genomic position of the corresponding variant, across multiple species (+++, >80% species with available alignments share the same allele at the position; ++, >70%; +, >60%; -, <60%; Appendix Figure 5.62).

5.4.5.2 Examining burden of multiple variants per gene among cases identified significantly associated genes

Given that a group of multiple rare mutations, each with a mild effect, together can explain a large proportion of the genetic susceptibility for complex disorders (371), it may be beneficial to group the detected variants into a functional unit, such as a gene, and to measure the groupwise difference between cases and controls. This approach not only permits to identify disease-associated genes [rather than discrete variants], but also alleviates the multiple testing problem arising from single variant-based tests (372).

I assessed several groupwise association tests for their statistical power (Appendix text 5.6.5 and Appendix Figure 5.24), and chose the "burden" test that examines whether the sum of the allele frequencies of variants within a gene is significantly higher in the case population than in the control (see Materials and Methods). I con-

sidered nine categories of variants that were classified based on their types of mutation, allele frequency, and evolutionary conservation status. The nine categories include: *(i) all*, *(ii) non-synonymous (ns)*, *(iii) coding*, *(iv) conserved* (i.e. variants in conserved region), *(v) conserved & ns* (i.e. nonsynonymous mutation in conserved region), *(vi) MAF<10%*, *(vii) MAF<1%*, *(viii) MAF<10% & conserved* (i.e. variants with MAF<10% in conserved elements), and *(ix) MAF<1% & conserved* (i.e. variants with MAF<1% in conserved elements).

As a result, I identified 7-36 significantly associated (p_{perm} <0.05) genes from each of the nine categories (*all*, 36 genes; *ns*, 13 genes; *coding*, 17 genes; *conserved*, 26 genes; *conserved & ns*, 7 genes; *MAF*< 10%, 32 genes; *MAF*< 1%, 29 genes; *MAF*< 10% *& conserved*, 24 genes; *MAF*< 1% *& conserved*, 18 genes; Appendix Table 5.7).

These significantly associated genes from each category might contribute to OCD in different genetic mechanisms. For example, the protein products of the genes identified from the *non-synonymous* and the *coding* categories might be affected by coding changes that have larger marginal effects, while the genes identified from the *conserved* category may have disruptions in their non-coding regulatory regions where the mutations may exert smaller more marginal effects. The genes identified from the *conserved & ns* are perhaps those affected by coding mutations with the strongest marginal effects, as may be indicated by the sequence constraint. The *MAF<10%* and *MAF<1%* categories were tested to increase the sensitivity to rare/uncommon mutations that are also likely to underlie OCD. Finally, applying the evolutionary conservation on top of these low allele frequency criteria may increase the power to detect functional mutations among these rare variants, possibly identifying genes with excessive rare variants affecting functional non-coding regulatory regions.

In any case, these significantly associated genes are of high interest for the genetics of OCD. Thus, I took all of these associated genes, forming a set of 123 genes (herein termed "investigative" genes), and examined their biological functions (see below).

5.4.5.3 Associated genes were enriched for functions related to synaptic formation and development of the cortex and the basal ganglia

Knowing which biological processes or pathways are involved in a disease can provide extremely useful insight into the common disease mechanism underlying a complex disorder (see section 4.2.1). Here, I considered all of the 123 investigative genes that

have variants with significantly higher frequencies in cases than in controls, in order to take into account various possible scenarios of how causative mutations play their roles in OCD (see section 5.4.5.2 and Materials and Methods), and sought to identify significantly over-represented biological functions among these investigative genes.

Employing the Database for Annotation, Visualization and Integrated Discovery (DAVID) functional annotation clustering (see Materials and Methods), I tested whether certain functional themes (i.e. Gene Ontology (GO) FAT terms; see Materials and Methods) are significantly enriched (FDR<5%) among these 123 investigative genes. For comparison between investigative genes and all of the resequenced genes, I repeated the same analysis on all 611 targeted genes. Subsequently, I identified 43 GO FAT terms including *telencephalon development, postsynaptic density*, and *axon* that were significantly enriched (FDR<5%) among the 123 investigative genes ($p_{corr} = 5.2 \times 10^{-14}$ - 2.8×10^{-5}, 1.7-15 fold-change; Figure 5.10 and Appendix Table 5.12). Of these 43, 40 (93%) GO FAT terms were also found in 236 GO FAT terms that were significantly enriched (FDR< 5%) among the 611 resequenced genes ($p_{corr} = 5 \times 10^{-39}$ - 5.3×10^{-3}, 1.6-10.5 fold-change; Figure 5.10 and Appendix Table 5.13).

In order to take any functional biases arising from the initial target genes into account, I compared the fold-enrichment of the 40 GO FAT terms that were significantly enriched in both the investigative genes and all the resequenced genes. Of these, 36 (90%) GO FAT terms, including *telencephalon development, regulation of cellular component size*, and *regulation of neurogenesis*, were identified with higher fold-enrichment in the investigative genes than in all the resequenced genes (1.1 (6.6/6.3) - 2.1 (15/7) fold-change ratio; Figure 5.10). This is consistent with the rationale of the study design, where a subset of the (functionally related) pre-selected genes are expected to contribute to the disruption of disease-relevant biological processes, thus a proper identification of disease contributing genes would amplify the functional enrichment signals that are perhaps also present in the pre-selected genes. Lastly, the three GO FAT terms that were only identified in the investigative genes were related to cell death, namely, *regulation of apoptosis, regulation of programmed cell death*, and *regulation of cell death* (Figures 5.10 and 5.11).

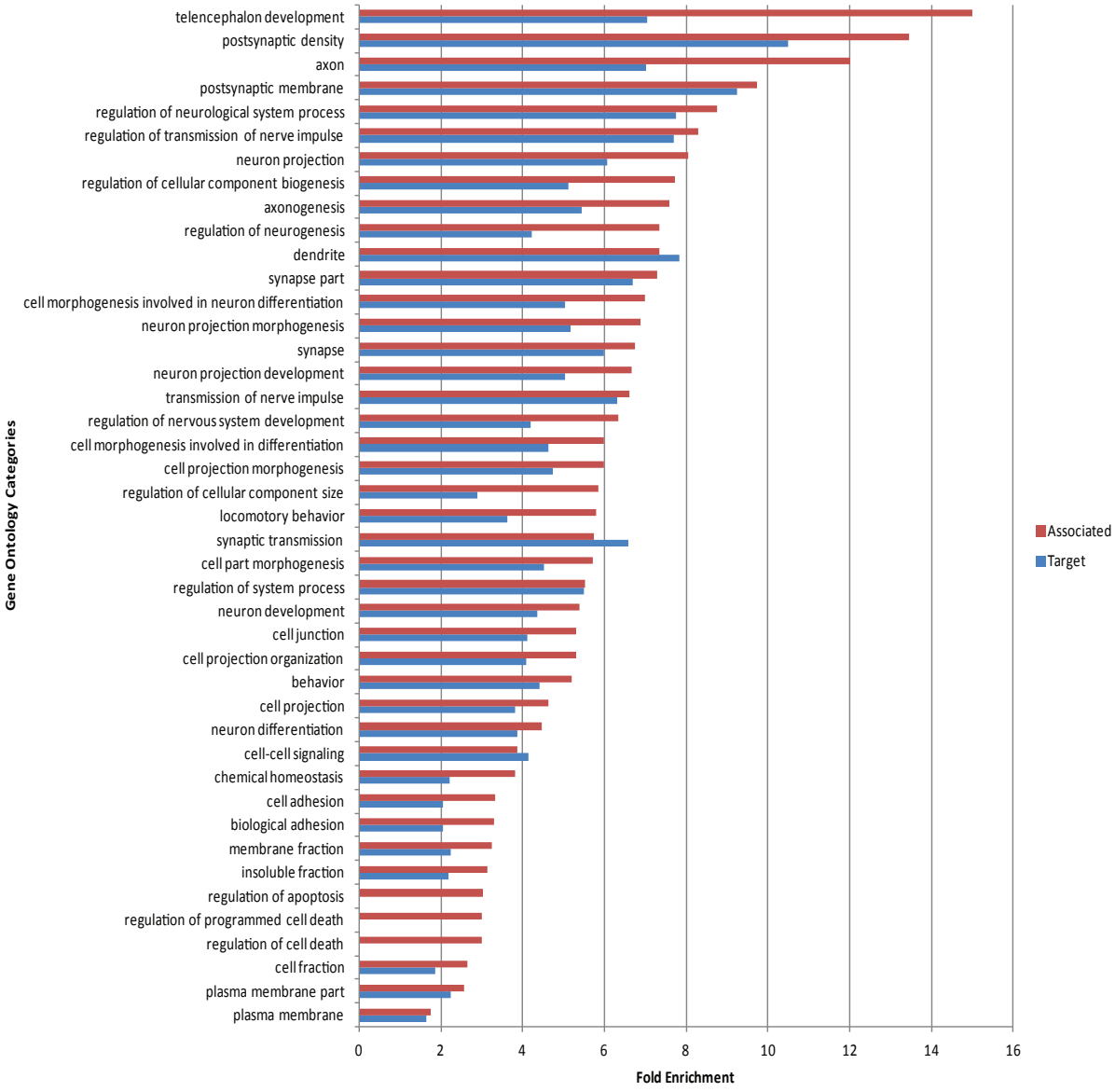


Figure 5.10: GO FAT terms that are significantly associated (FDR <5%) with 123 investigative genes and their fold-enrichments - The blue bars show the fold-enrichment of each GO FAT term among the 611 resequenced genes, while the red bars show the fold-changes among the 123 investigative genes.

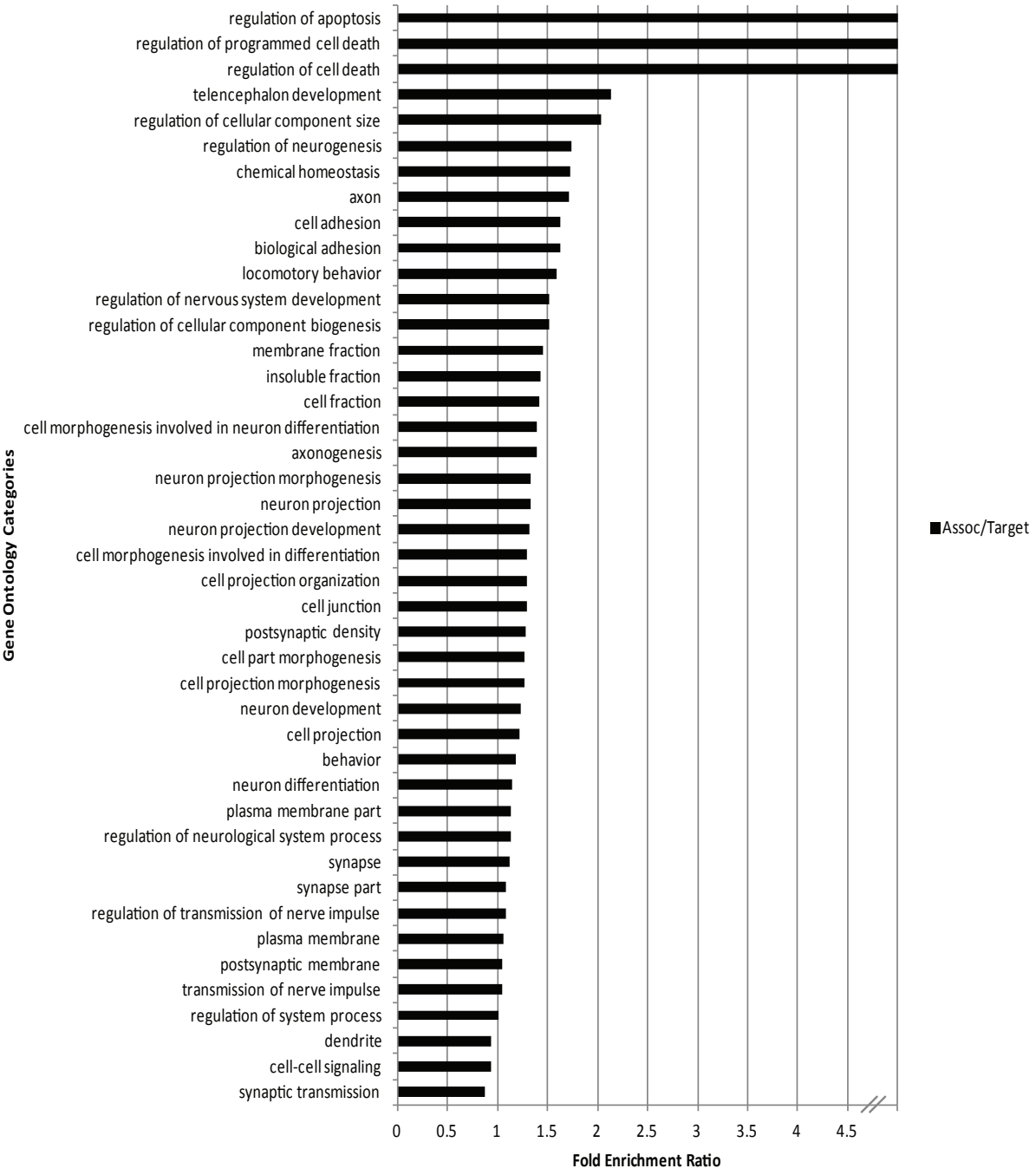


Figure 5.11: Significantly enriched GO FAT terms and the ratio of the fold-change in 123 investigative genes to the fold-change in 611 resequenced genes - The bar represents the fold enrichment ratio, i.e. [fold-enrichment in investigative genes]/[fold-enrichment in all resequenced genes]. The top three GO terms are significantly enriched (FDR $< 5\%$) only among investigative genes, but not in all resequenced genes.

5.4.6 Power calculation

Given that the case-control comparison of the identified variants and genes failed to achieve genome-wide significance (Appendix Figures 5.23 and 5.24), I sought to investigate whether the resequencing of additional samples would be worthwhile. As 200 additional cases and 200 matched controls were to be available for further resequencing, I sought to find out whether adding these available samples which doubles the current sample size, would provide sufficient power.

Initially, I made up a simulated set of 400 cases and 400 controls (i.e. the equivalent size to the final dataset) by sampling the current dataset, and examined whether the associations detected from the simulated data exceed the current level of significance. The results, however, indicated that the associations from the simulated data could not be used for such an analysis (Appendix Figure 5.56). This is probably due to a severe population stratification introduced by sampling the same objects multiple times within the limited number of case and control pools, which is unlikely to occur in the real world. Other options, such as introducing "normal" variations in the simulated data may mimic the real situation better, but it is difficult to decide what "normal" variations are, and its implementation lies beyond the scope of this study.

Accordingly, I took an approach to extrapolate the association results of several sub-sampled data from the original dataset. Thereby, I could avoid an artefactual population stratification introduced by sampling the same objects more than once. First, I generated four sets of sub-sampled data with different sizes, namely, 48 (case individuals) + 48 (control individuals), 64 + 64, 96 + 96, and 160 + 176 (i.e. the largest available QC+ sample size) cases and controls, and employed the same groupwise association test as before (see section 5.4.5.2). As a result, the largest data set (160 + 176) yielded the strongest association at $-log_{10}P$=2.8, the second largest set (96 + 96) at $-log_{10}P$=2.3, the third largest set (64 + 64) at $-log_{10}P$=1.9, and the smallest set (48 + 48) at $-log_{10}P$=1.4 (Figure 5.12). Extrapolation of these association levels for each size of the four sets suggests that doubling the original sample size may yield stronger associations, exceeding the 95% confidence interval from the expected distribution. Finally, this analysis was repeated multiple times, confirming that the sub-sampled data with the same size yield consistent association levels (Appendix Figures 5.57, 5.58, 5.59, and 5.60).

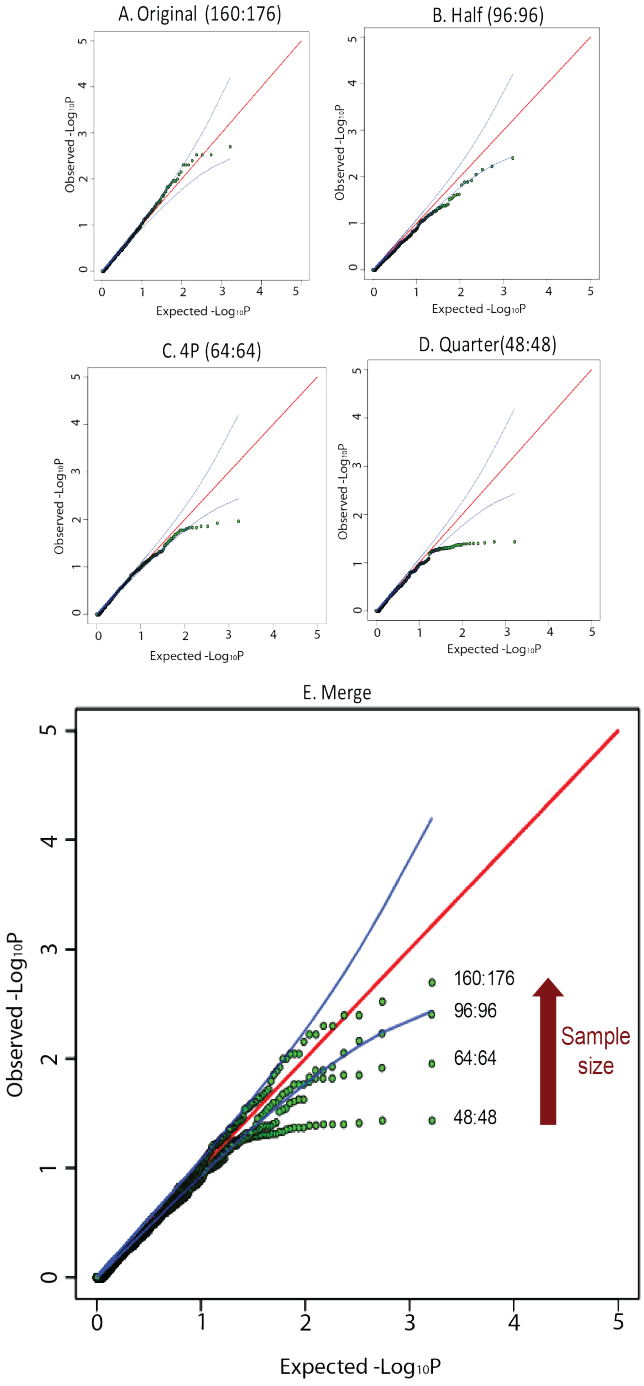


Figure 5.12: Quantile-quantile (Q-Q) plots for groupwise association test in sub-sampled data - The plots show the observed distribution of $-log_{10}P$ values obtained from groupwise association test (y-axis) of the original dataset and three sub-sampled datasets against the expected distribution (x-axis). A. Q-Q plot for the largest data set (160 cases: 176 controls), B. Q-Q plot for the ∼half-sized sub-samples (96 cases: 96 controls), C. Q-Q plot for the ∼1/3-sized sub-samples (64 cases: 64 controls), D. Q-Q plot for ∼quarter-sized sub-samples (48 cases: 48 controls). E. A-D overlaid

5.5 Discussion

5.5.1 Associated coding variants may be less likely to be causative to OCD

Although single variant association tests had limited power to detect genome-wide significant associations with the current sample size, here I investigated the top variants as pilot analyses. Among the most strongly associated coding variants ($-log_{10}P > 3$, Table 5.4), two variants would cause amino acid changes: Chr8:41566438 C→T variant ($p = 5 \times 10^{-4}$) changes R (arginine) 1954 of *ANK1* gene to H (histidine), and Chr3:196888560 A→G variant ($p = 0.002$) changes V (valine) 532 of *DLG1* gene to A (alanine). Nevertheless, these amino acid changing coding variants may be less likely to be causal to OCD. Both arginine and histidine are amino acids with positively charged side chains, and both valine and alanine have hydrophobic side chains. Thus, these amino acid changes may have little effect on the electric charges of their side chains that play an important role in the protein folding process, hence possibly leaving their final protein products functionally essentially intact. Furthermore, if those variants yield any effects, they might work as protective alleles rather than causative alleles, as both variants are more commonly present in the controls than they are in the cases (Table 5.4).

5.5.2 A variant in the 3'-UTR of *THBD* and its flanking sequence might serve as a new transcription factor binding site

Among the most strongly associated non-coding variants ($-log_{10}P > 3.5$), two variants showed substantial evidence suggesting that they may play functional roles (Table 5.5). The two variants are: *Chr20:23028063 T→C* variant ($p = 2 \times 10^{-5}$) in the 3'-UTR of the thrombomodulin gene (*THBD*, Figure 5.13) which encodes an endothelial-specific type I membrane receptor that binds thrombin, and *Chr3:190025964 C→T* variant ($p = 6.2 \times 10^{-4}$) in the 3'-UTR of the claudin 1 gene (*CLDN1*, Figure 5.17) which encodes an integral membrane protein that is required for cell-to-cell adhesion. Here, I will first discuss a potential role for the variant in *THBD* gene based on multiple lines of evidence inferred from publicly available bioinformatics resources, followed by a similar analysis for the other variant within *CLDN1* gene.

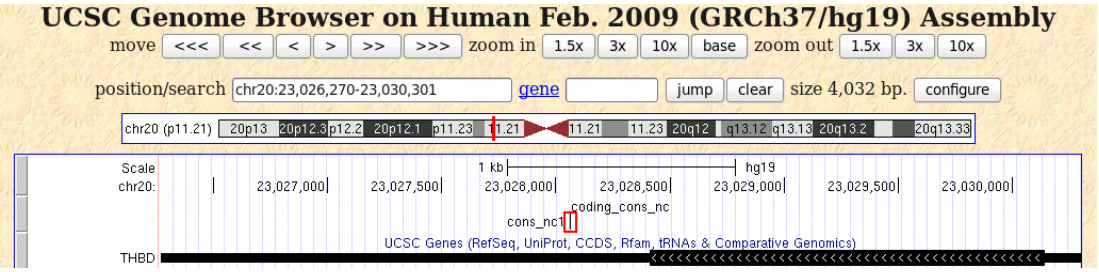


Figure 5.13: UCSC tracks showing a variant *Chr20:23028063 T→C* and its overlapping gene (hg19) - A strongly associated non-coding variant, *Chr20:23028063 T→C* (shown in red box), resides within the 3'-UTR of the *THBD* gene.

5.5.2.1 *Chr20:23028063 T→C* is uncommon in the general population

The dbSNP build 135 reports that the alternative C allele at position Chr20:23028063 (SNP ID: rs11696919) is uncommon in the general population (allele frequency, 2.69% [59/2,190]). In the OCD cohort employed here, the alternative C allele was present in 17.5% (28/160) of individuals with OCD, considerably higher also than in the control cohort (6.25% [11/176]). The strong association of this uncommon variant is consistent with an emerging view that a collection of rare or uncommon variants may contribute to a complex disorder (see section 1.2).

5.5.2.2 Reference allele T at Chr20:23028063 is evolutionarily conserved

In order to assess whether the variant is likely to be functional, I checked whether the variant position is conserved across multiple species. This is based on the assumption that if a specific nucleotide position of the genome plays an essential role, the allele is likely to be left intact because of purifying selection during evolution. Indeed, the reference allele T at Chr20:23028063 is evolutionarily well conserved across primates and most mammals, as it is conserved in 27 of 32 (84%) species in which the alignment information was available (Appendix Figures 5.26 and 5.62).

5.5.2.3 Its neighbouring region overlaps regulatory marks

Supported by the evolutionary conservation status, I hypothesised that the non-coding *Chr20:23028063 T→C* variant may reside within functional sequence, such as a genomic region that regulates the expression of a gene involved in OCD. Accordingly, I

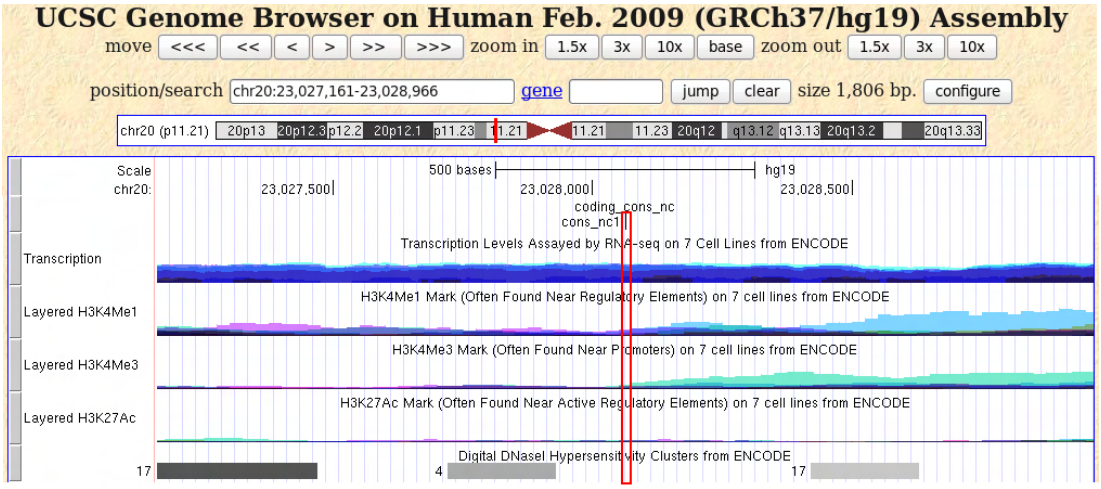


Figure 5.14: Chr20:23028063 and its flanking region with regulatory marks - The variant's position, Chr20:23028063, is shown in the red box. The *Transcription* track shows the transcription levels assayed by sequencing poly-adenylated RNA from 7 cell types (Appendix Table 5.14). The *H3K4Me1, H3K4Me3* and *H3K27Ac* tracks show the genomic regions where mono-methylation, tri-methylation, and acetylation of histone H3 proteins occur, respectively. The *H3K4Me3* track shows a histone mark associated with promoters, while the other two are suggestive of enhancers and other regulatory activities. The histone marks are, however, only informative at a high level, due to its low resolution (∼200 bases). The *DNase1 Clusters* track shows the regions where the chromatin is hypersensitive to cutting by the DNase1 enzyme. The DNase1 hypersensitive assay is higher in resolution at the DNA level than the histone marks and can be done on a large number of cell types. Regulatory regions, in general, tend to be DNase1 sensitive, and promoters are particularly DNase1 sensitive. A gray box indicates the extent of the hypersensitive region. The darkness is proportional to the maximum signal strength observed in any cell line. The number at the left of the grey box shows how many cell lines are hypersensitive in the region.

compared the variant's position with the genomic regions annotated with various regulatory markers that were available in the UCSC Genome Browser. Indeed, multiple regulatory markers showed compelling evidence that suggests regulatory activities in this region: (*i*) the transcription level is very high in this region, (*ii*) the histone marks that are indicative of an enhancer (H3K4Me1) and a promotor (H3K4Me3) become stronger from the area where the variant resides towards the coding portion of *THBD*, and (*iii*) the DNase1 hypersensitivity site marks are present in the variant's flanking region, suggesting that the region is in general within open chromatin (Figure 5.14). Considering that in metazoans, microRNAs (miRNAs) often regulate the level of gene expression by targeting 3'-UTR of messenger RNA transcripts (mRNA) (373), I also examined this region for conserved mammalian microRNA regulatory target sites (374). However, I could not find any evidence suggesting this region as being miRNA target site at the time.

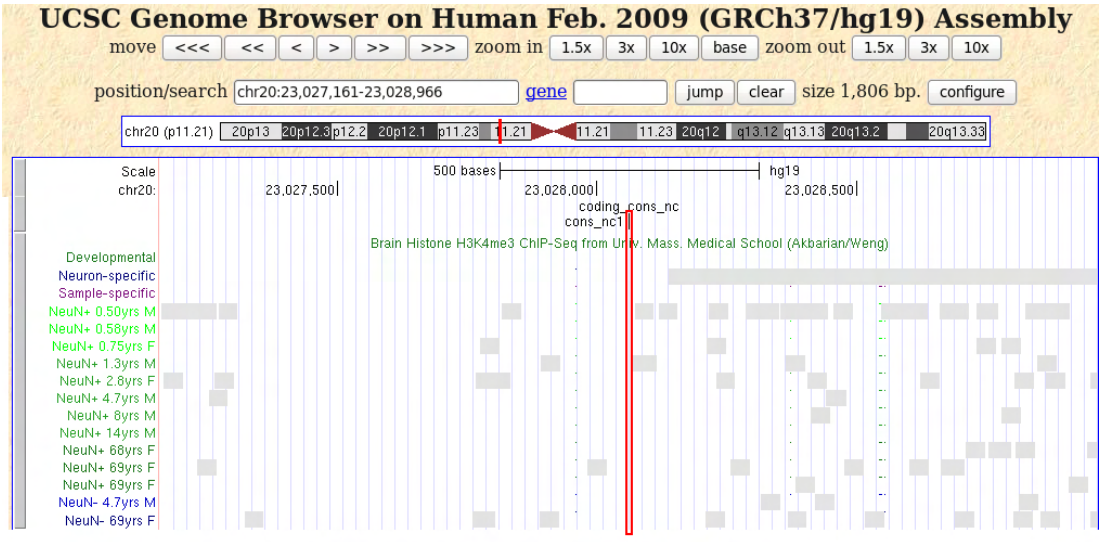


Figure 5.15: Brain histone H3K4Me3 ChIP-seq marks in Chr20:23028063 and its flanking region - The track displays the H3K4me3 marks in the human brain, identified by chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq). The measurements were made in neuronal and non-neuronal nuclei collected from the prefrontal cortex of 11 individuals ranging in age from 0.5 to 69 years (375).

Considering that the 7 cell lines that were assayed for the histone marks shown in ENCODE do not include brain tissue (ENCODE project, Appendix Table 5.14), I also compared the variant's position with H3K4Me3 histone marks in the prefrontal cortex

of the brain (375). The region is generally enriched with H3K4Me3 (promoter) marks, with some variability in their specific position depending on the age of tissue donors. Chr20:23028063 is particularly in close proximity to the H3K4Me3 marks shown in infants (0.5 and 1.3 yr, Figure 5.15), suggesting that this region may play a role in early infantile development of the prefrontal cortex.

5.5.2.4 *Chr20:23028063 T→C* may affect transcription factor binding

In order to assess whether the variant Chr20:23028063 T→C affects transcription factor binding, I investigated whether any transcription factor (TF) binding DNA motifs can be confidently mapped (matrix similarity>0.86 and core similarity>0.96; see Appendix text 6.6.7) to this region using TRANSFAC® (129). The reference T allele and its flanking sequence matched the consensus sequence of the DNA-binding site for the transcription factor OCT7 and the reference T allele appeared to serve as an important site in this TF-binding DNA motif (Figure 5.16). The actual binding of OCT7 to this region could have been confirmed bioinformatically if there was chromatin immunoprecipitation-sequencing (ChIP-seq) data for OCT7, but such data were not available at the time.

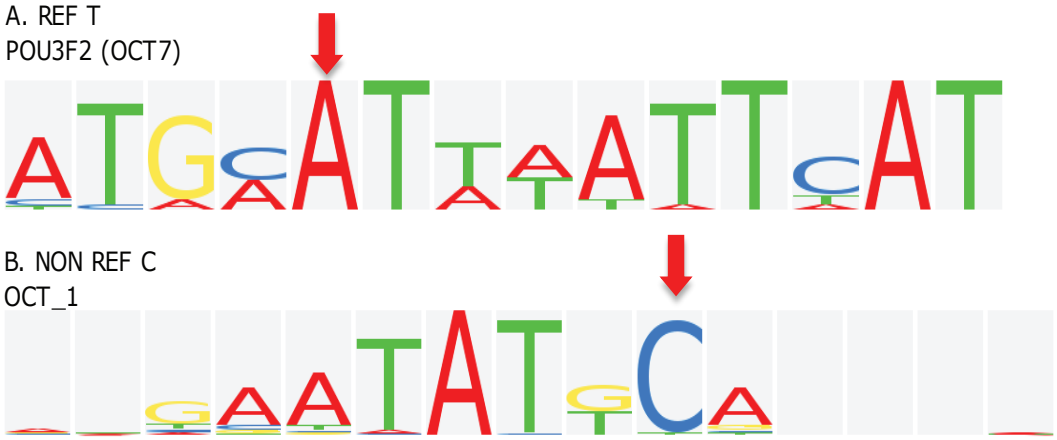


Figure 5.16: Transcription factor binding DNA motifs detected for Chr20:23028063 and its neighbouring region - The red arrow indicates the Chr20:23028063 position. The larger the letters are, the more important bases they are, in order to be recognised as a TF-binding DNA site for the suggested transcription factor. A. Consensus sequence for the transcription factor OCT7 binding (shown as the reversed strand). B. Consensus sequence for the transcription factor OCT1 binding.

The alternative C allele and its flanking sequence, on the other hand, matched the consensus sequence of the DNA binding site for another transcription factor, OCT1. The C allele is also strongly constrained in this sequence motif for OCT1 binding (Figure 5.16). This concordance between the variant's flanking sequences and the transcription binding sites suggest that the *Chr20:23028063 T→C* change may affect the original transcription factor binding site for OCT7, creating a different binding site for OCT1, and this change may contribute to OCD.

In fact, OCT7 is expressed predominantly in the central nervous system (CNS) during development and in the adult brain, and such CNS-specific transcription factors are suggested to play an important role in mammalian neurogenesis by regulating diverse patterns of gene expression (376)(377). OCT1, on the other hand, is ubiquitous (378) and activates the promoters of diverse genes from small nuclear RNAs (snRNA) to histone H2B to immunoglobulins (379)(380). This change in TF-binding DNA site from the CNS-specific OCT7 to the ubiquitous OCT1 perhaps affects the regulation of brain genes that are involved in OCD. *THBD*, in which this TF-binding DNA motif occurs within its 3'-UTR, may be a gene whose expression is regulated by OCT7, although the gene's function in the brain remains unknown.

This hypothesis immediately suggests a series of experiments to test. First, whether the Chr20:23028063 region serves as a protein binding site or not can be tested using an Electrophoretic Mobility Shift Assay (EMSA), where the interaction between a given DNA sequence and a protein(s) slows down the traveling of the DNA fragment during electrophoresis, hence resulting in a shift of its location in the gel. Once the DNA-protein interaction is confirmed, the identity of the protein(s) may be identified by *supershift*, where an antibody that recognizes the protein (in this case, OCT7 or OCT1 specific antibodies) is added, creating an even larger complex with a greater shift in the gel. In order to identify the genes that are affected by this variant, standard mRNA quantification methods (e.g. RT-qPCR) may be used on transfected cells, specifically for the *THBD* gene or CNS-specific genes, although interrogating the more relevant tissue from a subject would be problematic.

5.5.3 A variant in the 3'-UTR of *CLDN1* might affect functional non-coding RNA

The dbSNP build 135 reports that the population frequency of the complementary A allele (rs3172404) of the variant *Chr3:190025964 C→T* is 12.85% (282/2,195), while the complementary G allele to the reference C allele is 87.15% (1,913/2,195). In the OCD sample cohort employed here, the alternative T allele is present in 45% (72/160) of the cases, also considerably higher than 29% (51/176) of the controls. This variant overlaps with the 3'-UTR of *CLDN1* gene, which encodes a protein that is important in tight junction formation and function (Figure 5.17). The reference C allele appears to be well conserved across multiple species, unchanged across 29 of 36 (81%) vertebrates. Moreover, of the six vertebrates that have an alternative allele, the T allele only appears in non-primates (Appendix Figure 5.27).

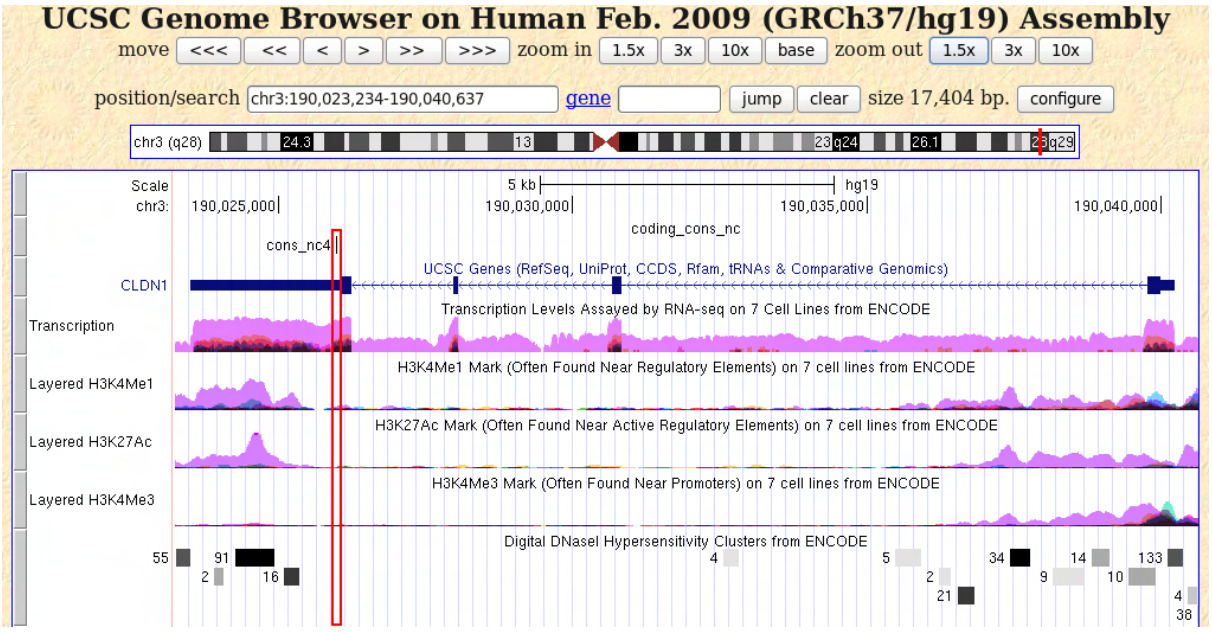


Figure 5.17: UCSC tracks showing a non-coding variant *Chr3:190025964 C→T* and its overlapping gene in hg19 - See Figure 5.14 for the track description.

Given the strong evolutionary conservation of the region, I again sought to investigate the transcription level and the regulatory marks in this region. Indeed, the variant and its flanking region showed a very high transcription level in general, while there was no apparent evidence for promoter or enhancer regions, according to histone marks

and the DNase1 hypersensitivity sites (Figure 5.17). Concordantly with the weak regulatory marks, the variant's position and its flanking region did not appear to be a promising transcription factor binding site, as the variant position was not constrained in the consensus sequence (Appendix Figure 5.61).

Excluding the possibility of the region's role for a transcription factor binding site, a non-coding RNA (ncRNA) gene, an emerging molecular mechanism, may be another possibility. Non-coding RNAs are active on the RNA level as opposed to producing protein product. Functional roles for some ncRNAs are well characterised, such as transfer RNA (tRNA) and ribosomal RNA (rRNA) during the protein synthesis. Other ncRNAs have been described more lately, such as small nucleolar RNA (snoRNA) and microRNA (miRNA). In addition, thousands of mRNA-like transcripts with little protein-coding potential have been identified via high-throughput transcriptome studies, but their roles are largely unknown (381)(382).

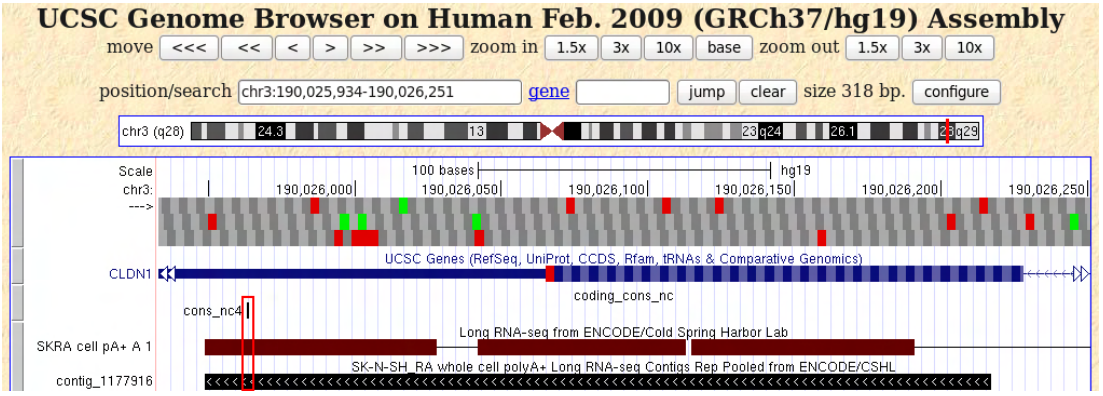


Figure 5.18: A non-coding RNA transcript overlaps Chr3:190025964 position - The RNA-seq data show a non-coding RNA transcript expressed in the neuroblastoma cell line from brain tissue (383), SK-N-SH_RA (SKRA) that is differentiated with retinoic acid (384). The contig (i.e. continuous regions covered by uniquely aligned reads) 1177916 is 268 bases long (chr3:190025950-190026217). The bases per kilobase per million mapped bases (BPKM, the sum of base coverage normalised by total aligned bases and the length of a given feature) was 0.055, and the irreproducible discovery score (IRD, the probability of NOT detecting that element in a third experiment equivalent in depth to the sum of the bio-replicates) was 0.17 for this contig.

Indeed, an RNA-seq experiment on the neuroblastoma cell line from brain tissue (383) identified a lowly (Bases per Kilobase per Million mapped bases [BPKM] = 0.055) but confidently (irreproducible discovery score [IRD] = 0.17) expressed 268 nucleotide-long RNA transcript (chr3:190025950-190026217) that overlaps with the

Chr3:190025964 C→T variant (Figure 5.18). While the transcript might simply represent 3'-UTR of *CLDN1*, given that the start and end positions of the transcript contig (i.e. continuous regions covered by uniquely aligned reads) do not coincide with 3'-ends of *CLDN1*, the transcript may have a potential to play a role as functional ncRNA.

The hypothesis that the *Chr3:190025964 C→T* variant may affect the functional non-coding RNA transcript (contig 1177916 in Figure 5.18) again immediately suggests a testable experiment. The control transcript may be silenced using RNA interference to examine its function in the normal cell, and the transcript with the variant may be expressed to investigate the mutation's effect.

However, in order to design such an experiment, a prediction of the ncRNA's role should be investigated before the development of an appropriate functional assay. Current prediction methods include (*i*) homology search among known RNA families and (*ii*) *de novo* ncRNA prediction based on secondary structure. Homology search is considered to be a reliable method, but it is restricted to thousands of known ncRNA sequences (385)(386), while the total number of ncRNAs in the human genome is estimated to be tens of thousands (387)(388). Indeed, I could not identify any known ncRNAs that are related to the ncRNA transcript identified here using a homology search, but this may be due to the incomplete annotation of all functional ncRNAs. *De novo* prediction, on the other hand, relies on the potential of a sequence to adopt a secondary structure, hence is not restricted to the small, already known subset of ncRNAs. However, it has been shown that the predicted structures of ncRNAs are not significantly more stable than predictions from random RNA sequences (389), thus the predicted RNA structure does not necessarily indicate a functional significance (382). In addition, accuracy of sequence-structure prediction is affected by imperfect models and scoring systems, even when a RNA sequence of known structure was tested (390). Therefore, although I could predict a RNA secondary structure for the ncRNA transcript using RNA fold (391) (data not shown), the secondary structure may not provide much useful information about the ncRNA's function. A more recent, combined approach that measures the probability of a given alignment of sequences to adopt a consensus secondary structure predicted from the alignments of multiple species (387)(388) perhaps provides a better indication for a novel ncRNA's functional significance, but this approach would still not provide sufficient information for the prediction of the ncRNA's role.

5.5.4 Evolutionary conservation status is a good indication to detect functional variants associated with OCD

To find out which types of mutations may commonly underlie OCD, I calculated the proportion of the strongly associated genes ($p < 0.01$) to all the associated genes ($p < 0.05$) for each of the test variant categories. Among the nine variant type categories, [i.e. *all, nonsynonymous (ns), low-frequency (MAF<10%)*, and *rare* variants (*MAF<1%*), their four corresponding variant types that overlap with evolutionarily conserved elements (termed '*in cons*'), and *coding* variants] three of four categories that consider conserved elements were ranked higher in terms of the proportion of the strong association than each of their corresponding variant type categories that include the variants regardless of their conservation status (*ns* → *ns in cons*, 9th (7.7% [1/13]) → 1st (57% [4/7]); *all* → *all in cons*, 5th (27.8% [10/36]) → 3rd (38.5% [10/26]); *rare* → *rare in cons*, 7th (24% [7/29]) → 4th (33.3% [6/18]); Table 5.6).

Test variant type	% strong association to all	Ratio
Nonsyn in conserved	57.1%	4/7
Low-frequency	50.0%	16/32
In conserved	38.5%	10/26
Rare in conserved	33.3%	6/18
All	27.8%	10/36
Low-frequency in cons	25.0%	6/24
Rare	24.1%	7/29
Coding	23.5%	4/17
Nonsynonymous	7.7%	1/13

Table 5.6: Proportion of strongly associated ($p < 0.01$) genes to all associated genes for each test variant type - The low-frequency variants are with MAF<10% and the rare variants are those with MAF<1%. The different variant types are shown in different colours (i.e. *nonsynonymous* (Nonsyn, in green), *low-frequency* (in blue), *rare* (in magenta), and *all* (in red)). The ratio displays the number of strongly associated genes over the number of all associated genes for each test variant type.

This may indicate that the consideration of base conservation status improved the power to detect associations with the OCD cohort, by decreasing the background noise produced by non-disease contributing variants. This observation supports the rationale that was used when classifying variant type categories, in which the application of evolutionary conservation perhaps filters out variants that do not contribute to OCD, while keeping the variants that are more likely to be functional (see section 1.1.3).

5.5.5 Groupwise association test allowed a prioritisation of candidate genes for OCD

Although all of the strongly associated genes are of interest for OCD, it may be unrealistic to follow-up all of the 123 investigative genes in parallel. One way to prioritise the associated genes is to consider the genes' p-values (i.e. association level) and how disrupted their protein products would be (e.g. number of variants found in that gene). Accordingly, I plotted the associated genes for each of the nine test variant type categories (Table 5.6), in terms of their $-log_{10}P$ and the number of detected variants per gene (Appendix Figures 5.36-5.44). Of the 123 genes plotted in the 9 separate graphs, two genes were ranked as the top in both the p-values and the number of variants per gene: *ROBO1* (roundabout, axon guidance receptor, homolog 1 [Drosophila]) gene in the *nonsynonymous variants in conserved element* category ($p = 0.004$, the number of variants=6; Appendix Figure 5.37), and *CDH13* (cadherin 13, H-cadherin [heart]) gene in the *MAF<10%* category ($p = 0.001$, the number of variants=1,258; Appendix Figure 5.41). Notably, both genes are involved in axon growth and cell adhesion: *ROBO1* encodes an integral membrane protein that is both an axon guidance receptor and a cell adhesion receptor, involved in the axons that cross the midline of the central nervous system (392). *CDH13* is a member of the cadherin superfamily that mediates cell adhesion, and acts as a negative regulator of axon growth during neural differentiation (393). Furthermore, a GWAS has identified another cadherin, *CDH2*, to be strongly associated with compulsive disorder in dogs (357).

In addition to the approach of counting the number of variants per functional unit (i.e. gene), I computed the mutation frequency per 1kb per gene to take into account gene size (Appendix Figures 5.35 and 5.45-5.53). Under this metric, *ABCF3* (ATP-binding cassette, sub-family F, member 3) gene was ranked as the top in both association level and mutation frequency in the *all* variants category ($p = 0.003$, the number of variants per 1kb per gene=4.1; Appendix Figure 5.45), and *TMEM41A* (transmembrane protein 41A) gene in the *rare (MAF< 1%)* variant category ($p = 0.002$, the number of variants per 1kb per gene=1.7; Appendix Figure 5.51) The functions of both *ABCF3* and *TMEM41A* in the brain are currently not well understood, but a genome-wide linkage scan for OCD identified these regions as susceptibility loci (394). These four genes (*ROBO1, CDH13, ABCF3*, and *TMEM41A*) may be a good starting point for functional studies of candidate genes for OCD.

5.5.6 Pathway analysis identified functional themes for OCD and prioritised further candidate genes

Pathway analysis of the 123 investigative genes resulted in 43 significantly associated (FDR< 5%) GO FAT terms (Figures 5.10 and 5.11). Using REVIGO (271), I summarised these terms into higher-level functional themes. The summarised functional themes include biological processes related to cell projection organisation, chemical homeostasis, behaviour, cell-cell signalling, and adhesion (Appendix Figure 5.54), and cellular components such as plasma membrane part, neuron projection, synapse, and cell junction (Appendix Figure 5.55). These functional themes may implicate processes involved in neuron projection organisation, synaptic transmission, and adhesion in OCD.

Also, the strongest association of the *telencephalon development* GO term with 123 investigative genes may suggest an implication of the defects during the developmental process of the embryonic structure (telencephalon), from which the mature cerebral cortex and the basal ganglia derive. Developmental defects in the telencephalon are likely to perturb the adult CSTC circuit that may be responsible for OCD (see section 5.2.1.2), hence that may lead to OCD symptoms. This hypothesis may be experimentally tested by disrupting each of the 8 genes (*TSC1, DRD2, NF1, GRIN1, RAC1, AVPR1A, BCAN*, and *LMX1A*) that contribute to the *telencephalon development* term enrichment to examine whether these genes affect the development of telencephalon and the neuro-circuit in the mature brain that was developed from the telencephalon, and whether such defects exert OCD-like behaviours in model organisms.

5.5.7 Inclusion of non-coding regulatory region increased statistical power

Lastly, I was curious whether the up- and down-stream regions of a gene that may potentially regulate the flanked gene are also involved in OCD (see sections 1.1.3 and 1.2). Thus, I created genomic intervals that are inclusive of genes and their potential regulatory regions (i.e. 50kb flanking region) for each targeted gene, and performed the same groupwise association test as before (see section 5.6.5). Indeed, the association levels resulted from the groupwise test including the up- and down-stream regions were enhanced compared to the original test performed on gene only regions. The observed

association level with OCD followed the expected distribution at lower association levels (i.e. $-log_{10}P < 1.6$) but then exceeded the upper 95% confidence interval of the expected distribution up to the range of $-log_{10}P = 2$ (Appendix Figure 5.25). In contrast, the original groupwise association test on gene only regions yielded association levels around the upper 95% confidence interval, but failed to exceed this confidence interval (Appendix Figure 5.24). This increase in the level of associations with OCD by including the potential regulatory region is consistent with the hypothesis that a disruption of regulatory elements may also underlie OCD (see sections 1.1.3 and 1.2). Given the increased statistical power based on the assumption that intergenic variants affect the regulation of the closest gene, these additional intergenic variants that contributed to a genic associations with OCD may be worth testing for their *cis*-regulatory activity on their neighbouring genes.

5.5.8 Conclusions

In summary, resequencing ∼13Mb genomic regions of 192 individuals with OCD and 192 matched healthy controls yielded promising association signals with OCD, despite the modest sample size and the complexity of the disorder. The close examination of the top associated variants, genes, and biological processes using publicly available bioinformatics database produced several experimentally tractable hypotheses for OCD. A power calculation showed that the borderline significance of the genome-wide association may be overcome by doubling the sample size. Additional resequencing is thus planned with the same study design, but given the required resource and time for the new sequencing experiment, the analysis of the new sequence dataset lies beyond the scope of my doctoral thesis. Finally, the results from human OCD may be compared with a parallel resequencing study of compulsive disorder in dogs (see Chapter 6) in order to investigate whether there is a common mechanism underlying compulsive disorders across different species.

5.6 Appendices

5.6.1 Handling duplicated reads

Hybrid capture sequencing in this study yielded a total of 60-170 million reads per pool. The number of reads that were uniquely mapped was 17-63 million per pool, indicating a large proportion of duplicated reads (45-85% per pool).

Duplicated reads can be problematic because when a "sequencing" error is introduced by PCR and is amplified by subsequent PCR cycles, the error can be repeatedly duplicated, resulting in multiple reads containing this PCR error. This can eventually lead to a bias in variant calling and allele frequency estimation. Thus, PCR duplicates are recommended to be removed from the raw sequence data. In addition, the removal of PCR duplicates may provide a computational benefit by reducing the number of reads to be processed in the downstream steps.

Although it is difficult to distinguish these artefactual duplicates from naturally occurring alignment duplicates arising from deep sequencing data, one can inspect whether there are more duplicates than expected by chance. Generally, reads are considered to be duplicates if the read pairs have identical 5'-end coordinates. The chance of two read pairs having the identical coordinates (i.e. alignment duplicates) depends on the coverage, but generally speaking, is very small: According to Heng Li's calculation (http://www.broadinstitute.org/gsa/wiki/images/1/10/Samtools.pdf), theoretical alignment duplicates can be estimated as: $d' = \frac{m}{2\sqrt{\pi}\sigma L} \sim \frac{0.282m}{\sigma L}$, where m is the number of read pairs, σ is the standard deviation of the insert size distribution, and L is the length of the reference. As the empirical duplication rate in the human OCD sequencing data, 0.66 ([2,651,508,194 total reads – 910,362,456 unique reads]/2,651,508,194 total reads) was substantially larger than the alignment duplication rate, 0.39 ($m = 1,336,461,572$ $\sigma = 73.6$, $L = 13,238,099$), a significant amount of true PCR duplicates probably existed in the sequence data. Thus, in order to avoid the potential bias caused by PCR sequencing error, duplicated reads were marked and excluded for downstream analysis.

The duplication rate in this study was probably exacerbated by a known technical artefact of the hybrid capture sequencing, where copies of the same DNA molecule tend to be sequenced repeatedly (395). These duplicates usually arise when too few total DNA molecules are used at any stage of DNA molecule manipulations, particularly immediately prior to the PCR step. The rate of PCR duplicates is: $d \sim \frac{m}{2N}$ (http://www.broadinstitute.org/gsa/wiki/images/1/10/Samtools.pdf), where m is the number of sequenced reads and N is the number of DNA molecules before PCR amplification. This equation implies that in order to reduce PCR duplicates, the original pool of distinct DNA molecules (N) should be increased before PCR amplification, or to reduce the number of reads to be sequenced (m). Reducing the number of PCR cycles, however, has little effect. This also implies that it is inevitable to produce more PCR duplicates as one sequences more. Indeed, I observed a correlation between duplication rate and sequencing depth from the OCD sequence data (Appendix Figure 5.19).

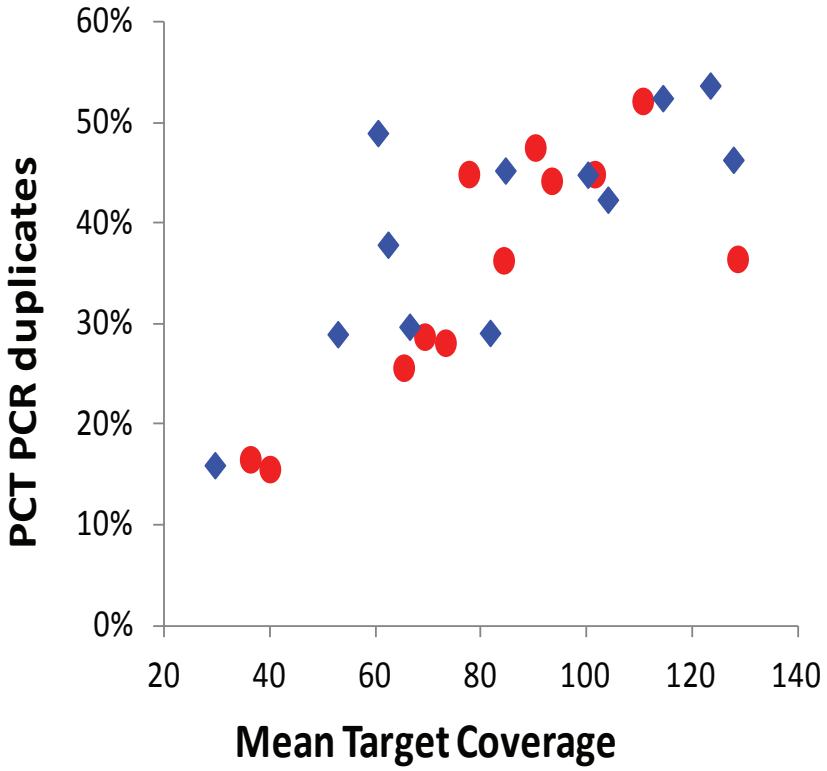


Figure 5.19: Correlation between mean target coverage and duplicate rates - Red circles represent case pools and blue squares indicate control pools.

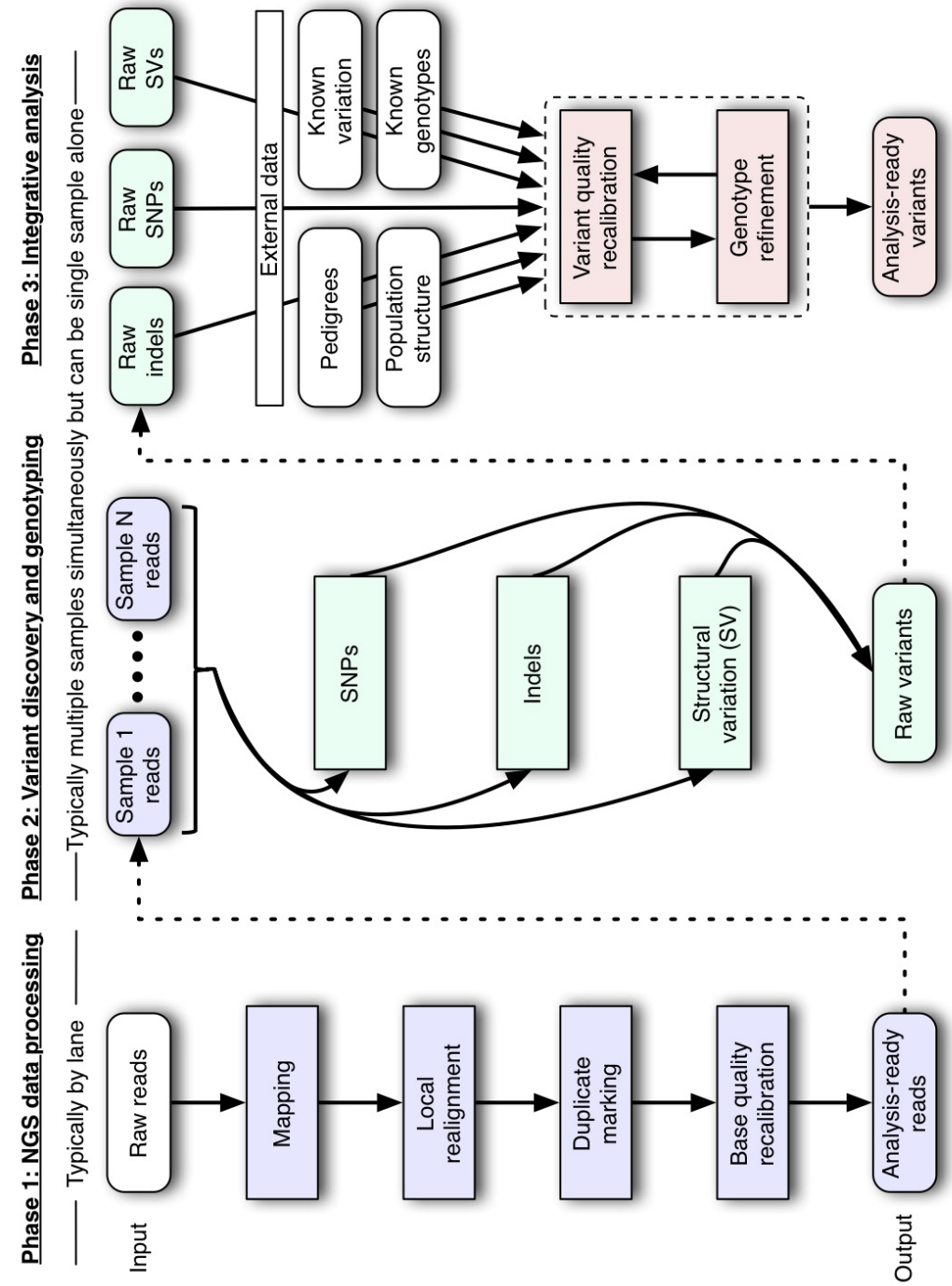


Figure 5.20: Overview of GATK data processing pipeline - The pipeline consists of three phases: NGS data processing, variant discovery and genotyping, and integrative analysis. (367).

5.6.2 Population stratification

As all of the discovered variants were present in both the case and the control cohort, it was not possible to prioritise variants that were found only in the case pools. Instead, I compared the allele frequencies of the detected variants between case and control cohorts to test whether any of the detected variants were associated with OCD. To be able to use such a genetic association approach, I first examined whether there is a population stratification between the case and the control cohorts that may yield misleading associations that reflect the genetic differences arising from ancestry rather than the actual association with OCD.

There are two principles in SNP usage for association studies: identity by state (IBS) and identity by descent (IBD). IBS is the mere sharing of alleles between individuals while IBD is the sharing of alleles between individuals with an identified, common ancestral source of the alleles. For instance, consider a family with a father having genotype CC, a mother with genotype AC, and two children with AC genotypes. Each sibling's A allele must have been inherited from the mother on the same physical chromosome, hence shared IBD. The C alleles, however, are not possible to be distinguish whether the C alleles were transmitted from the same physical chromosome or not, as it could have been transmitted from any of the father's two C alleles. Thus, the C alleles are shared IBS between the siblings.

Although IBD may provide more insights into the inheritance of alleles, the use of IBD requires knowledge of the relationship between individuals, or the population allele frequencies to calculate IBD probabilities. In contrast, IBS can be calculated without the knowledge of pedigree structure and does not require allele frequency information. Considering that pedigree information was not available from the OCD cohort, IBS was the only possible metric that could be used for an association study.

Using the genome-widely distributed 51,462 genetic markers from 139 cases and 136 controls that passed quality control for both sequencing and SNP array experiments, I calculated the pairwise IBS distance that indicates genomic similarity/difference between individuals using PLINK (135). To do this, I examined every pair of SNPs between two individuals and classified them into one of three following categories:

- IBS2 (identical): Both individuals have the same genotype. (e.g. AA and AA, CC and CC, AC and AC)
- IBS1 (one allele in common) : One allele is shared between the two compared individuals. (e.g. AA and AC, AC and CC)
- IBS0 (no allele in common): No alleles are shared between the two individuals. (e.g. AA and CC)

The mean IBS between the case and the control groups, within the cases, and within the controls were all 0.76, with *standard deviation* values of 0.004, 0.005, and 0.003, respectively, hence the mean genetic distance between case and control groups did not differ from the genetic distance within each group. For more detailed examination of the OCD cohort, I asked whether or not there are group differences or similarities in the average IBS distances, with regard to various within- and between- comparison of the case and the control groups, using the pairwise IBS distances calculated as above. In order to obtain the expected values for each comparison, the case/control labels were permuted 10,000 times, and group metrics were recalculated. To provide a full description of the variability between and within the groups, 12 different tests were carried out as below:

- **T1: Case/control less similar** $p = 0.003$
- T2: Case/control more similar $p \sim 1$

- **T3: Case/case less similar than control/control** $p = 0.03$
- T4: Case/case more similar than control/control $p = 0.97$
- **T5: Case/case less similar** $p = 0.04$
- T6: Case/case more similar $p = 0.96$
- T7: Control/control less similar $p = 0.97$
- **T8: Control/control more similar** $p = 0.03$
- T9: Case/case less similar than case/control $p = 0.05$
- T10: Case/case more similar than case/control $p = 0.95$
- T11: Control/control less similar than case/control $p = 0.98$
- **T12: Control/control more similar than case/control** $p = 0.02$

Of the 12 tests carried out, 5 tests yielded significant differences between the tested groups. The case and the control groups were less similar than expected (T1, $p = 0.003$), the cases were more variable within the group than expected (T3, $p = 0.03$; T5, $p = 0.04$), and the controls were more similar within the group than expected by random chance (T8, $p = 0.03$; T12, $p = 0.02$). Nevertheless, the multi-dimensional scaling (MDS) analysis (that condenses the information for a large number of loci and alleles into one or a few synthetic variables) (396) showed no evident substructure in the sample population (Appendix Figure 5.21). Thus, the less similarity between the case and the control groups (T1) probably represents a few outliers rather than the true population substratification.

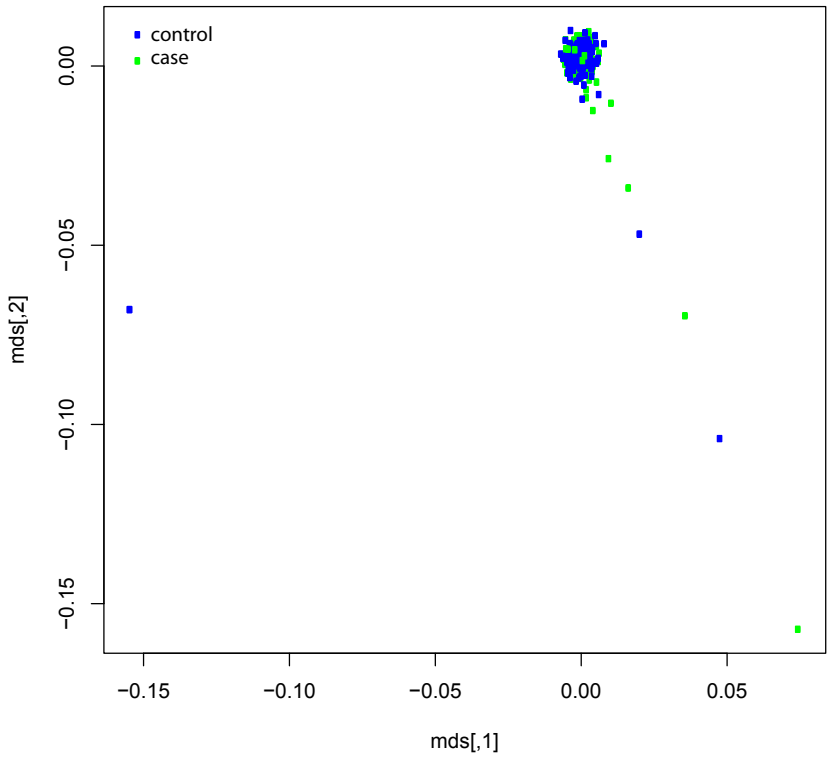


Figure 5.21: A multi-dimensional scaling (MDS) plot of samples in the OCD cohort - The x-axis explains the greatest amount of variance, while the y-axis explains the next greatest amount of variance. Blue dots represent control samples and green dots represent case samples.

5.6.3 Comparison of AF distribution between cases and controls

Besides controlling the population stratification based on the genome-wide markers, I also compared the distribution of estimated allele frequencies of the variants in target sequence, in order to examine whether there is any bias in the allele frequency distributions of the detected variants between cases and controls. Plotting quantiles of allele frequencies in cases against the quantiles in controls revealed that there is no evidence of deviation in allele frequency distributions between cases and controls (Appendix Figure 5.22).

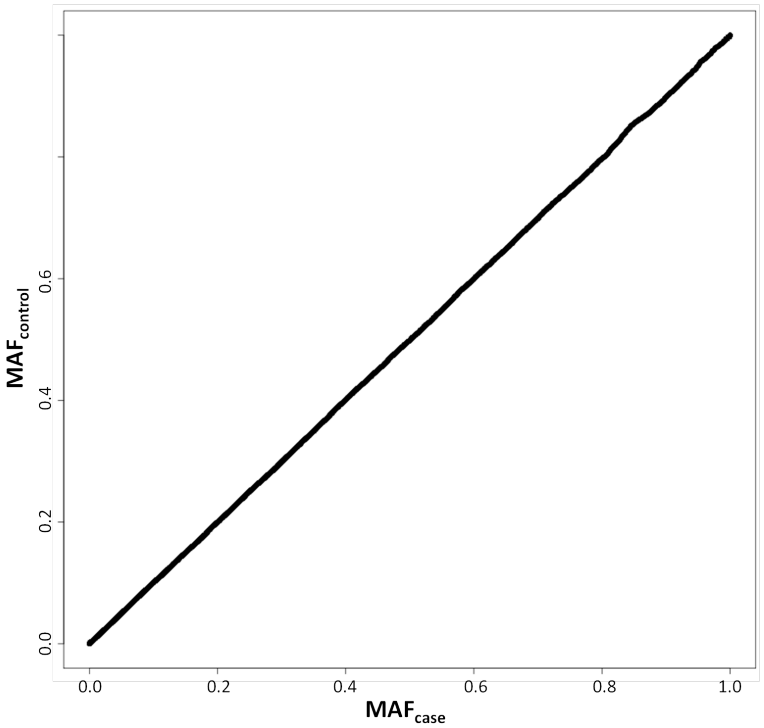


Figure 5.22: Q-Q plot of AF distribution in cases and controls - The x-axis (MAF_{case}) indicates allele frequencies in cases while the y-axis ($\text{MAF}_{control}$) indicates allele frequencies in controls. Each dot represents a quantile of two distributions being compared. The quantile's position is determined by its allele frequency in case (x-coordinate) and that in control (y-coordinate). If the two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line $x = y$.

5.6.4 Q-Q plot of p-values for single variant association test

Here I made a Q-Q plot to evaluate the significance of the observed associations of detected variants with OCD, compared to the identical association statistics that are expected under the null hypothesis of no association. Observed $-log_{10}P$ values that were calculated from t-statistics were ranked incrementally on the y-axis and plotted against the expected distribution of the $-log_{10}P$ under the null hypothesis of no association (i.e. uniform distribution) on the x-axis. Divergence from the identity line indicates either that the assumed null distribution is inaccurate or that the observed statistics includes values emerging in some other way, such as by a true association.

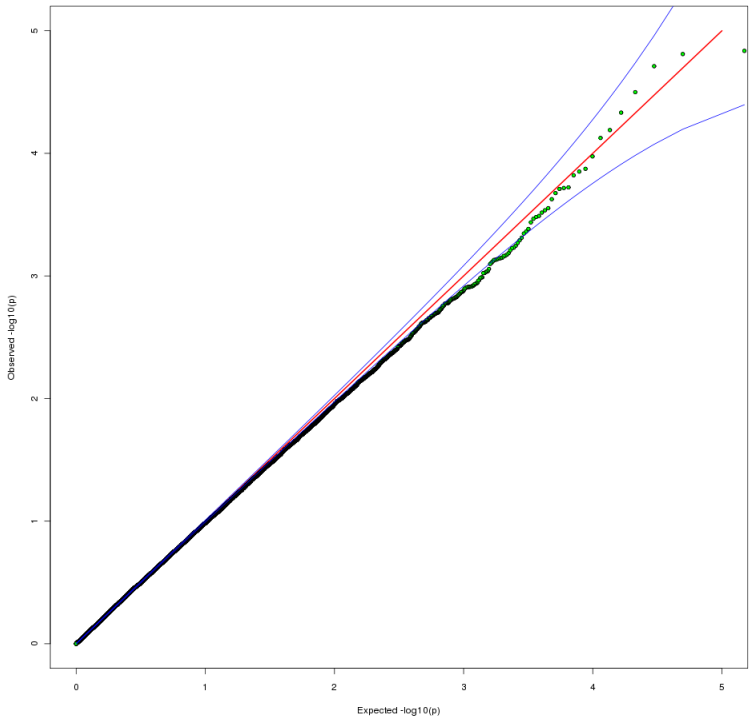


Figure 5.23: Q-Q plot for the distribution of p-values from single variant association test - Red line indicates the situation when the expected distribution (x-axis) and the observed distribution (y-axis) are identical. The 95% confidence intervals are shown as blue lines.

5.6.5 Assessing groupwise association tests

Initially, I employed three different groupwise association tests in order to select the most suitable association test:

- Burden test: test for an excess of non-reference alleles per gene in cases
- Weighted test: test for an excess of non-reference alleles (with a weight on rare alleles) per gene in cases
- Protective test: test for a deficit of non-reference alleles per gene in cases

	# associated genes		
Test variant type	**Burden**	**Weighted***	**Protective**
all	36	26	14
non-synonymous (ns)	13	16	18
coding	17	15	20
conserved	26	30	19
Conserved & ns	7	10	6
MAF < 10%	32	32	20
MAF < 1%	29	27	18
MAF < 10% & conserved	24	21	22
MAF < 1% & conserved	18	21	29
Total # assoc. genes	**123**	**111**	**109**

* weight applied on variants that are rare in control

Table 5.7: Number of associated genes ($p < 0.05$) from three different gene-based association tests - Three different gene-based association tests, namely burden, weighted, and protective tests, were employed. The first column shows the types of variants that were tested for each of the three association tests. MAF stands for minor allele frequency.

For each test, I tested nine categories of variant types that were classified based on mutation type, allele frequency, and evolutionary conservation status. The nine categories include: *nonsynonymous variants, coding variants, variants in conserved region, nonsynonymous variants in conserved region, low-frequency variants (MAF<10%), rare variants (MAF<1%), low-frequency variants in conserved elements*, and *rare variants in conserved elements*. The tests on different variant types for each of the three association tests yielded 7-36, 10-32, 6-29 associated ($p < 0.05$) genes for each category, making up 123, 111, and 109 investigative genes from burden test, weighted test, and protective test, respectively (Appendix Table 5.7).

As expected, 60% (72) of the investigative genes from weighted test overlaps with those from burden test (67%), while only 13.8-22.5% (17-25) of the investigative genes from protective genes overlap with those from the other two tests (15.6%, burden test; 23%, weighted test; Appendix Table 5.8).

In order to evaluate the significance of the observed gene associations with OCD from the three groupwise association tests, I compared the observed associations with the identical association statistics that are expected under the null hypothesis of no association, by making Q-Q plots for each association test. The Q-Q plot for burden test nicely follows the identity line up to $-log_{10}P < 1.5$, then shows an inflation in a higher association level along the borderline of the upper 95% confidence interval (Appendix Figure 5.24 panel A). Such deviation towards the

% overlap* (# genes)	Burden	Weighted	Protective
Burden		66.67% (74)	15.60% (17)
Weighted	60.16% (74)		22.94% (25)
Protective	13.82% (17)	22.52% (25)	
Total	100% (123)	100% (111)	100% (109)

* Percentage to column headings

Table 5.8: Overlap of associated genes among three association tests - The % overlap is shown as a percentage of the total number of investigative genes from the test indicated in the column headings. The numbers in bracket show the number of genes that are common in the two tests indicated in the column and the row headings.

upper boundary arises perhaps due to true associations with OCD. The Q-Q plot for weighted test also follows the identity line in the low level associations ($-log_{10}P < 1.5$), but then stays between the identity line and the lower 95% confidence interval (Appendix Figure 5.24 panel B). The protective test, on the other hand, follows the lower 95% confidence interval up to the association level at $-log_{10}P < 1$, then shows a notable deflation in a higher association level (Appendix Figure 5.24 panel C).

Such a strong deflation in the Q-Q plot of protective test might have occurred because the actual distribution of p-values under the null of this test is not uniform. Given that there were more controls than cases among the QC+ samples, there is inevitably a higher chance for a variant to have more allele counts in controls than in cases. Thus, the actual null distribution for the protective test in OCD case-control data set may show a surfeit of small p-values approaching 0, rather than uniform distribution. If the expected null distribution was corrected for this non-uniformity, the Q-Q plot of protective test might have followed the identity line better. In contrast, the actual null distribution of p-values for burden or weighted tests may show a surfeit of large p-values approaching 1, as it is less likely that a variant has higher allele counts in cases than in controls given the same cohort. If this opposite non-uniformity was accounted for the expected null distribution, the Q-Q plots for burden and weighted tests might have shown more inflated deviations from the identity line compared to the current Q-Q plots.

While all of the three groupwise association tests did not yield confident excess of strong associations, the results from burden test appeared to be most promising, given the deviation showing the borderline significance at the upper 95% confidence interval.

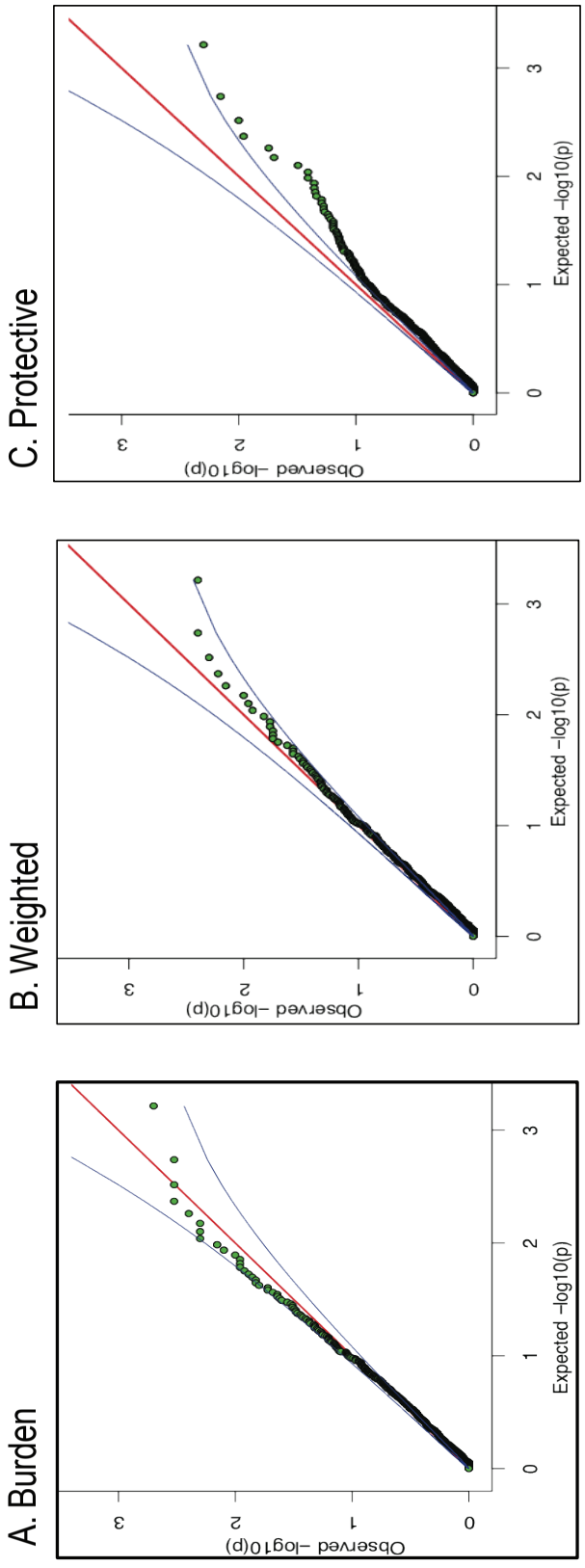


Figure 5.24: Q-Q plot for the distribution of p-values from three different groupwise association tests - Panel A-C shows Q-Q plots for burden test, weighted test, and protective test, respectively.

Figure 5.25: Q-Q plot for the distribution of $-log_{10}P$ from burden test on genic regions and their 50kb up- & downstream regions - For the comparison with a Q-Q plot from the same test on genic only regions, see Appendix Figure 5.24 panel A.

5.6.6 Top associated variants

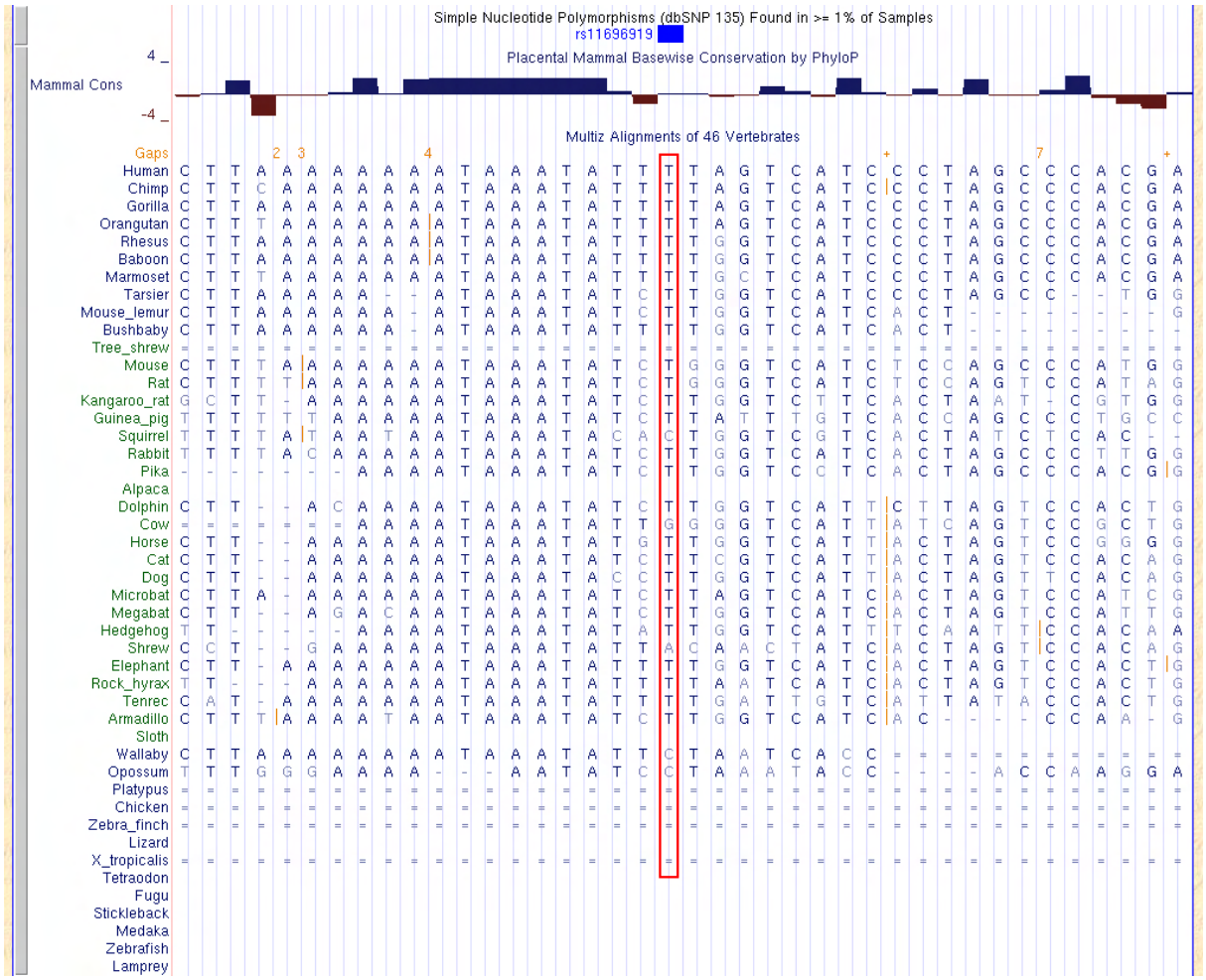


Figure 5.26: Chr20:23028063 position is evolutionarily conserved - The reference T allele at Chr20:23028063 (shown as rs11696919) is evolutionarily conserved across the primates and most mammals. See Appendix Figure 5.62 for their phylogeny.

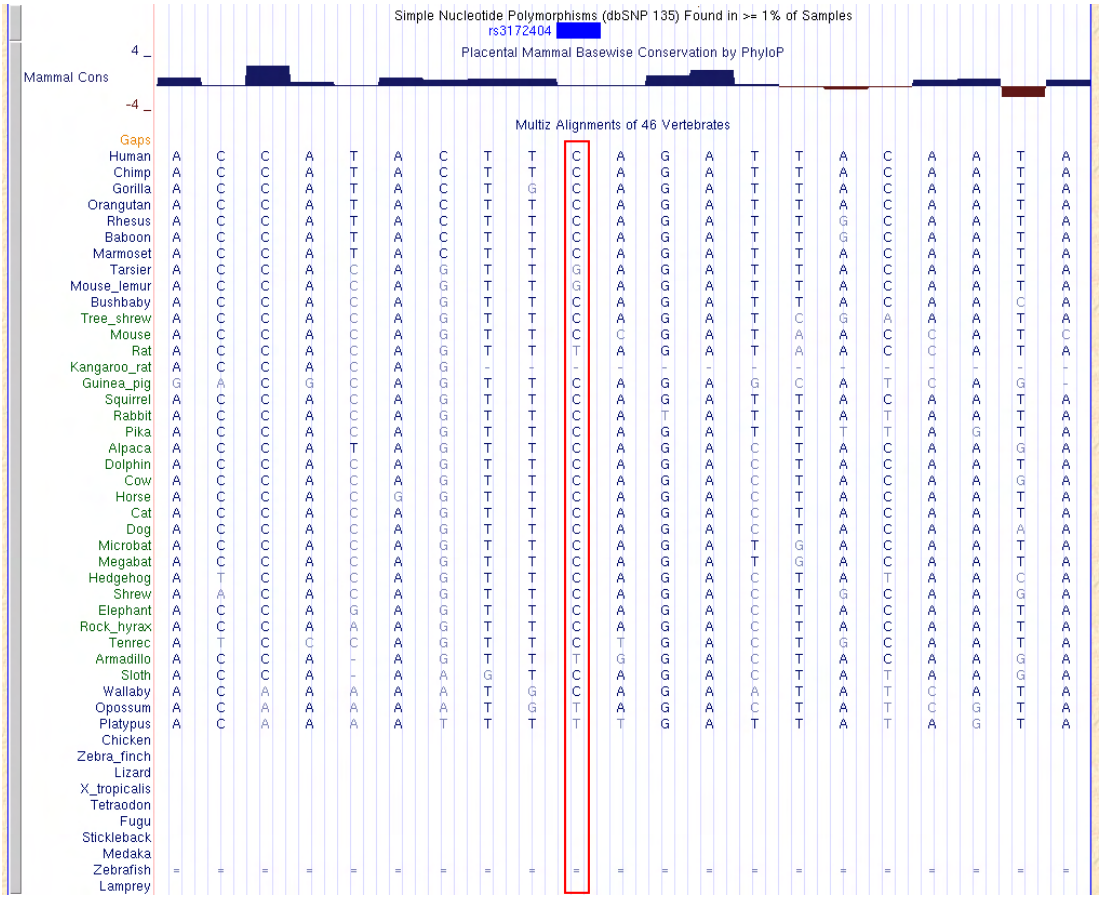


Figure 5.27: Reference allele at Chr3:190025964 is evolutionarily conserved - The reference C allele at the variant's (shown as rs3172404) position is conserved among 29 vertebrates. See Appendix Figure 5.62 for their phylogeny.

5.6.7 Samples and target regions

Table 5.9: List of targeted genes - The chromosome, start, and end coordinates are in hg19.

chr	start	end	gene
chr	start	end	gene
1	1950767	1962192	Gabrd
1	9005921	9035146	CA6
1	15736390	15756839	Efhd2
1	19665273	19811992	Capzb
1	35331037	35370984	DLGAP3
1	37261127	37499844	GRIK3
1	38275238	38325292	MTF1
1	39547117	39952789	Macf1
1	41086351	41131324	RIMS3
1	57463578	59012446	DAB1
1	70032867	70589169	Lrrc7
1	75594118	75627216	Lhx8
1	97543301	98386615	DPYD
1	107682628	108024471	NTNG1
1	110230417	110260883	GSTM1
1	154127779	154164609	Tpm3
1	154679916	154842754	KCNN3
1	155916629	155959864	Arhgef2
1	156611739	156629318	Bcan
1	156669409	156675375	Crabp2
1	156785541	156851642	NTRK1
1	164528801	164854300	Pbx1
1	165171104	165325475	Lmx1a
1	175913966	176176370	RFWD2
1	181452715	181770713	Cacna1e
1	186640944	186649559	PTGS2
1	204797781	204991949	Nfasc
1	209859549	209908294	HSD11B1
1	220701567	220837798	MARK1
1	231664398	232177016	DISC1
1	247920765	247921708	OR1C1
1	98511626	98511727	MIR137
1	39456915	39471735	AKIRIN1
2	946554	1371382	Sntg2
2	10861774	10925228	Atp6v1c2
2	37075472	37193615	Strn
2	50145643	51259674	NRXN1
2	54683453	54898582	Spnb2
2	65454828	65498385	Actr2
2	68405988	68479651	Ppp3r1
2	70834749	70995329	Add2

Continued on next page

chr	start	end	gene
2	101436612	101613287	NPAS2
2	105977284	106055230	Fhl2
2	114647536	114716167	Actr3
2	115199898	116601936	DPP10
2	122513120	122525426	TSN
2	154728425	155310488	GALNT13
2	157180945	157189287	Nr4a2
2	162272619	162281573	Tbr1
2	166095911	166248818	SCN2A
2	166845670	166930149	SCN1A
2	168810530	169104105	STK39
2	172639914	172750813	SLC25A12
2	172950207	172954399	DLX1
2	172964166	172967478	DLX2
2	173600524	173917618	RAPGEF4
2	182321618	182402466	ITGA4
2	191208195	191236390	INPP1
2	200134223	200329831	Satb2
2	206547223	206662856	NRP2
2	210288770	210598834	Mtap2/MAP2
2	219081873	219119070	Arpc2
2	222282748	222438922	Epha4
2	233344536	233352532	Ecel1
2	236402735	237034114	CENTG2
2	30945639	31030311	CAPN13
2	31395923	31456724	CAPN14
2	79412356	80875904	CTNNA2
2	31456879	31491259	EHD3
2	31133333	31378068	GALNT14
2	80515480	80531487	LRRTM1
3	2140549	3099644	CNTN4
3	8792096	8811300	OXTR
3	9834226	9848789	Arpc4
3	14186649	14220172	XPC
3	41236400	41281939	Ctnnb1
3	42299317	42307662	Cck
3	49394610	49395791	GPX1
3	53529030	53846491	Cacna1d
3	59735037	61237133	FHIT
3	62355346	62359190	FEZF2
3	67425144	67705038	SUCLG2
3	78646387	79817059	ROBO1
3	113847556	113918254	DRD3
3	121312169	121349139	FBXO40
3	129149792	129158852	MBD4
3	142984064	143567346	SLC9A9
3	143690639	143711209	C3orf58

Continued on next page

chr	start	end	gene
3	173116243	174001116	Nlgn1
3	179113876	179169371	Gnb4
3	182660558	182698326	DCUN1D1
3	182733007	182833860	MCCC1
3	182840003	182880667	LAMP3
3	182895830	183145855	MCF2L2
3	183205320	183273499	KLHL6
3	183415605	183530413	YEATS2
3	183533663	183543393	MAP6D1
3	183547173	183602693	PARL
3	183637725	183735727	ABCC5
3	183749331	183757156	HTR3D
3	183770834	183778459	HTR3C
3	183814851	183824782	HTR3E
3	183852809	183863098	EIF2B5
3	183873283	183891313	DVL3
3	183892633	183901878	AP2M1
3	183903862	183911795	ABCF3
3	183960116	183967313	ALG3
3	183967444	184010818	ECE2
3	183977002	183979251	CAMK2N2
3	184017021	184026840	PSMD2
3	184032355	184053145	EIF4G1
3	184053716	184064062	FAM131A
3	184064339	184079391	CLCN2
3	184079559	184086363	POLR2H
3	184089772	184095932	THPO
3	184097860	184107619	CHRD
3	184279586	184300195	EPHB3
3	184428155	184429836	MAGEF1
3	184529930	184770402	VPS8
3	184908413	184971837	EHHADH
3	185000753	185200624	MAP3K13
3	185207388	185216845	TMEM41A
3	185225570	185270369	LIPH
3	185300506	185348883	SENP2
3	185361527	185542827	IGF2BP2
3	185632359	185655924	TRA2B
3	185764107	185826901	ETV5
3	185864990	186080023	DGKG
3	186256232	186264491	CRYGS
3	186263857	186288332	TBCCD1
3	186287951	186303588	DNAJB11
3	186330849	186339105	AHSG
3	186353757	186370796	FETUB
3	186383797	186396022	HRG
3	186435119	186461741	KNG1

Continued on next page

chr	start	end	gene
3	186501360	186507683	EIF4A2
3	186507683	186524484	RFC4
3	186560462	186576250	ADIPOQ
3	186648515	186796339	ST6GAL1
3	186838740	186857263	RPL39L
3	186915273	186919251	RTP1
3	186935937	187009752	MASP1
3	187086167	187089367	RTP4
3	187386695	187388201	SST
3	187416046	187420345	RTP2
3	187439164	187463475	BCL6
3	187871718	188597441	LPP
3	189674518	189840226	LEPREL1
3	190023502	190040215	CLDN1
3	190231890	190375843	IL1RAP
3	190930321	190967908	OSTN
3	190984944	191048325	UTS2D
3	191046873	191116458	CCDC50
3	191178951	191179243	PYDC2
3	191859683	192445388	FGF12
3	196769431	197026143	Dlgh1
3	119540803	119813264	GSK3B
3	172162950	172166203	GHSR
3	171318619	171528504	PLD1
3	172223464	172241269	TNFSF10
4	2845583	2931787	Add1
4	2965342	3042474	GRK4
4	3768074	3770251	ADRA2C
4	46920918	46995580	Gabra4
4	47033294	47428445	GABRB1
4	48485359	48491541	Slc10a4
4	93225549	94693648	GRID2
4	99799606	99851786	EIF4E
4	119202416	119273922	Prss12
4	134070469	134112731	PCDH10
4	158141294	158287224	Gria2
4	187508937	187644987	FAT1
4	186506598	186877870	SORBS2
5	1392909	1445538	SLC6A3/DAT1
5	9035137	9546233	SEMA5A
5	10277707	10308168	Cmbl
5	10971951	11904110	Ctnnd2
5	24487209	24644911	CDH10
5	26880708	27038689	CDH9
5	34656432	34832716	Rai14
5	50678957	50690557	Isl1
5	52776594	52781902	Fst

Continued on next page

chr	start	end	gene
5	71403117	71505396	Mtap1b
5	78669786	78809700	Homer1
5	112043217	112181935	APC
5	134363424	134370324	PITX1
5	138089106	138270722	Ctnna1
5	148206155	148208188	ADRB2
5	149980641	150038792	Synpo
5	152869174	153193428	Gria1
5	156607906	156682109	EMT
5	156693090	156822604	Cyfip2
5	158122923	158526788	Ebf1
5	174867675	174871163	Drd1a
5	174905513	174955621	Sfxn1
5	176883613	176900694	Dbn1
5	108528055	108530759	AK021888
5	110998322	111312628	C5orf13
5	110559647	110820746	CAMK4
5	111478137	111755010	EPB41L4A
5	108083522	108523373	FER
5	111755279	111756673	FLJ11235
5	111498314	111505972	hNBL4
5	109025155	109203429	MAN2A1
5	108670410	108745675	PJA2
5	112212080	112258031	REEP5
5	110073861	110098482	SLC25A46
5	111497181	111497314	SNORA13
5	112196992	112203603	SRP19
5	110834021	110848233	STARD4
5	111496225	111498003	tiga1
5	109624933	110062381	TMEM232
5	110405777	110413722	TSLP
5	110427869	110466201	WDR36
6	2833735	2842081	Serpinb1b
6	5998234	6007633	Nrn1
6	6144311	6320924	F13A1
6	24495196	24537434	ALDH5A1
6	29523389	29600962	GABBR1
6	29624808	29640147	MOG
6	29910330	29913653	HLA-A
6	31543349	31546110	TNF alpha
6	31982571	32003194	C4B
6	32485162	32557562	HLA-DRB1
6	33387846	33421465	Syngap1
6	38643703	38670952	GLO1
6	53659533	53788919	LRRC1
6	56911383	56920021	KIAA1586
6	78171947	78173120	HTR1B

Continued on next page

chr	start	end	gene
6	86159301	86205496	Nt5e
6	88849586	88875767	CNR1
6	101846904	102517957	GRIK2
6	110299508	110301921	Gpr6
6	114376750	114663540	HS3ST5
6	116262692	116381921	FRK
6	118869441	118881586	PLN
6	123100645	123105216	FABP7
6	132269318	132272518	Ctgf
6	135605111	135818903	AHI1
6	137243401	137246774	Slc35d3
6	146348781	146758731	Grm1
6	152011630	152424406	ESR1
6	153071932	153080898	Vip
6	154331635	154568001	MOR
6	161768590	163148834	PARK2
6	166822853	167275771	RPS6KA2
6	15523040	15663271	DTNBP1
6	152598636	152949689	SYNE1
6	16299343	16761721	ATXN1
6	13469510	13486415	C6orf114
6	13791020	13814789	CCDC90A
6	14117864	14137146	CD83
6	13363818	13487787	GFOD1
6	16238810	16295778	GMPR
6	15245733	15522252	JARID2
6	16129316	16148476	MYLIP
6	13615558	13621126	NOL7
6	12716887	13287528	PHACTR1
6	13621730	13711796	RANBP9
6	13925202	13980233	RNF182
6	12718541	12935067	RPEL
6	13574791	13614783	SIRT5
6	13266773	13328770	TBC1D7
7	6414125	6443597	Rac1
7	14184674	14942550	Dgkb
7	15239942	15601640	TMEM195
7	27132615	27135625	HOXA1
7	31377081	31380538	Neurod6
7	50344377	50472796	Ikzf1
7	69063904	70257884	AUTS2
7	73113540	73133988	Stx1a
7	86273229	86494191	Grm3
7	94214535	94285521	SGCE
7	94536948	94925725	Ppp1r9a
7	94927670	94953884	PON1
7	96635289	96640351	DLX6

Continued on next page

chr	start	end	gene
7	97361270	97369782	Tac1
7	98923509	98963885	Arpc1a
7	103112232	103629963	RELN
7	106505722	106547585	PIK3CG
7	106685177	106802255	Prkar2b
7	107564245	107643804	LAMB1
7	107788083	108096826	NrCAM
7	110303109	111202538	IMMP2L
7	113726381	114331092	FOXP2
7	116312458	116438439	MET
7	116593380	116870073	ST7
7	116916687	116963343	WNT2
7	117350707	117513561	Cttnbp2
7	121958478	122526554	Cadps2
7	126078651	126893147	GRM8
7	129470574	129592789	UBE2H
7	137074385	137531609	Dgki
7	145813452	148118086	CNTNAP2
7	153584418	154685994	DPP6
7	155250823	155257520	EN2
7	82383320	82792197	PCLO
7	83587664	83824217	SEMA3A
7	84624873	84816171	SEMA3D
7	82993221	83278324	SEMA3E
8	1449568	1656640	DLGAP2
8	6264120	6506025	MCPH1
8	24771273	24776605	Nef3
8	24808470	24814131	Nefl
8	41510744	41754280	Ank1
8	53535018	53627026	RB1CC1
8	57353512	57359282	Penk1
8	72109669	72274467	Eya1
8	80523379	80578394	Stmn2
8	82192784	82197008	FABP5
8	97505881	97624037	SDC2
8	108911544	109095913	Rspo2
8	120428551	120436678	Nov
8	143953773	143961236	CYP11B1
8	26605666	26722922	ADRA1A
8	1772148	1906806	ARHGEF10
8	27317278	27336813	CHRNA2
8	26371790	26515321	DPYSL2
8	22545173	22550090	EGR3
8	21900427	21906317	FGF17
8	28351772	28421959	FZD3
8	9760897	9760982	MIR124-1
8	22102474	22102556	MIR320A

Continued on next page

chr	start	end	gene
8	14710946	14711019	MIR383
8	1765396	1765473	MIR596
8	9599181	9599278	MIR597
8	10892715	10892812	MIR598
8	18248754	18258723	NAT2
8	31497267	32622558	NRG1
8	17829837	17887455	PCM1
8	22298595	22398635	PPP3CC
8	20002366	20040717	SLC18A1
8	97657498	98155721	PGCP
9	14510	29739	WASH1
9	116234	118417	FOXD4
9	121038	179075	CBWD1
9	214864	465259	DOCK8
9	470293	746104	KANK1
9	841689	969088	DMRT1
9	976963	991732	DMRT3
9	1049857	1057553	DMRT2
9	2015341	2193621	SMARCA2
9	2621792	2654485	VLDLR
9	2717525	2729757	KCNV2
9	3224648	3525983	RFX3
9	3824127	4300035	GLIS3
9	4490443	4587469	SLC1A1
9	4662297	4665256	PPAPDC2
9	4679565	4706593	CDC37L1
9	4711157	4742043	AK3
9	4792833	4861064	RCL1
9	5163864	5185618	INSL6
9	5231418	5233966	INSL4
9	5299867	5304580	RLN2
9	5334968	5339873	RLN1
9	5510544	5571281	PDCD1LG2
9	5765075	5833081	ERMP1
9	5890908	5909822	MLANA
9	6011019	6015640	RANBP6
9	6215806	6257982	IL33
9	6413150	6507049	UHRF2
9	6532465	6645692	GLDC
9	6716494	7175647	KDM4C
9	8314246	10612723	PTPRD
9	35681989	35690054	Tpm2
9	87283465	87638505	NTRK2
9	90112657	90323548	DAPK1
9	119187506	120177317	ASTN2
9	123963760	124095120	Gsn
9	130965662	131017527	Dnm1

Continued on next page

chr	start	end	gene
9	131314865	131395938	Spna2
9	132575221	132586441	DYT1
9	133320093	133376660	ASS
9	135766734	135820020	TSC1
9	136501484	136524467	DBH
9	140033608	140406529	Grin1
10	16555741	16564004	C1ql3
10	61788159	62493155	Ank3
10	64571759	64578927	EGR2
10	64926989	65225722	JMJD1C
10	65281122	65381971	REEP3
10	67679724	69455949	CTNNA3
10	75910964	76469059	Adk
10	78629358	79397577	KCNMA1
10	87359311	88126250	Grid1
10	89623194	89728531	PTEN
10	92500577	92617671	HTR7
10	102756964	102767585	LZTS2
10	112836789	112840660	ADRA2A
10	52750944	54055274	PRKG1
11	532242	535550	HRAS
11	637304	640703	DRD4
11	925840	1012239	Ap2a2
11	2185159	2193035	TH
11	18042083	18062335	TPH1
11	20691116	21597227	Nell1
11	27515970	27528326	Lin7c
11	27528398	27743605	BDNF
11	35160416	35253946	CD44
11	46354738	46402104	Dgkz
11	67205517	67211263	Coro1b
11	70313961	70935842	Shank2
11	71145458	71159477	DHCR7
11	77774906	77779396	Thrsp
11	83166055	85338314	Dlgh2
11	105480799	105852819	Gria4
11	113185250	113244016	Ttc12
11	113280317	113346413	Drd2
11	113775588	113817282	HTR3B
11	113845796	113861032	HTR3A/5_HT3A
11	115039952	115375241	CADM1
11	117707692	117747367	Fxyd6
11	128834955	129062093	Grit
11	63275458	63284246	GRIP1
11	63228875	63258680	HRASLS5
11	63273523	63284246	LGALS12
11	63341943	63381941	PLA2G16

Continued on next page

chr	start	end	gene
11	63137260	63177710	SLC22A9
12	2080228	2807115	Cacna1c
12	9585012	31257725	DDX11
12	13714409	14133022	GRIN2B
12	47158543	47219780	Slc38a4
12	53585107	53601000	ITGB7
12	63540215	63546590	AVPR1A
12	72332625	72426221	TPH2
12	109038888	109125295	Coro1c
12	110872706	110888158	Arpc3
12	120123597	120315092	Cit
12	120427647	120532299	CCDC64
13	28874482	29069265	FLT1
13	31287614	31338556	ALOX5AP
13	35516455	36246872	NBEA
13	42614171	42803888	Dgkh
13	42846288	42897402	Akap11
13	47407512	47471050	HTR2A
13	66876966	67804468	PCDH9
13	84451343	84456528	SLITRK1/GTS
13	107142097	107187337	Efnb2
13	109248499	109860355	MYO16
14	31363004	31495607	Strn3
14	39866877	39901704	FBXO33
14	42076763	42373750	LRFN5
14	47308829	48143988	MDGA2
14	52313951	52436517	Gng2
14	64693752	64805268	ESR2
14	76837689	76968178	ESRRB
14	77228234	77249363	VASH1
14	96671134	96710666	BDKRB2
15	25582396	25684128	UBE3A
15	25923861	26108349	ATP10A
15	26788694	27184686	Gabrb3
15	29131167	29410514	APBA2
15	29560358	29562017	NDNL2
15	37183232	37393500	Meis2
15	41786121	41795745	Itpka
15	43809805	43823817	Mtap1a
15	52599479	52821247	Myo5a
15	63334837	63364111	Tpm1
15	68871572	69020142	Coro2b
15	75117816	75124136	Cplx3
15	80696691	80890272	ARNT2
15	88419987	88799661	NTRK3
15	22892683	22993155	CYFIP1
15	27112057	27194003	GABRA5

Continued on next page

chr	start	end	gene
15	27216516	27778134	GABRG3
15	50534147	50557926	HDC
15	23043280	23086843	NIPA1
15	23005348	23034410	NIPA2
15	22833394	22873889	TUBGCP5
16	1203240	1271771	CACNA1H
16	2097989	2138712	TSC2
16	2227183	2246465	Caskin1
16	8768443	8878431	ABAT
16	19179637	19278553	SYT17
16	23847299	24231930	PRKCB1
16	29882481	29910580	SEZ6L2
16	49312210	49315715	Cbln1
16	55690021	55737698	SLC6A2
16	82660577	83830197	CDH13
16	89334034	89556969	ANKRD11
16	30064471	30081734	ALDOA
16	29912146	29930953	ASPHD1
16	29827527	29833814	C16orf53
16	29753785	29757340	C16orf54
16	29869678	29874578	CDIPT
16	30016836	30022401	DOC2A
16	30035748	30040857	FAM57B
16	30004316	30007387	HIRIP3
16	29916332	29937545	KCTD13
16	29817425	29822502	MAZ
16	29831786	29859342	MVP
16	29823408	29827201	PRRT2
16	29690440	29709309	QPRT
16	29985221	29999719	TAOK2
16	30097116	30103205	TBX6
16	30103635	30107521	YPEL3
17	7093211	7123369	Dlgh4
17	7311501	7323179	Nlgn2
17	7608519	7614691	Efnb3
17	8043788	8055753	PER1
17	17584786	17714763	RAI1
17	26083792	26220409	NOS2A
17	27900486	27916610	GIT1
17	28523379	28562954	SLC6A4/OCD1
17	29421994	29708905	NF1
17	37219555	37307902	Plxdc1
17	40118758	40129754	Cnp1
17	44668037	44834833	Nsf
17	45331207	45421658	ITGB3
17	46606806	46608272	HOXB1
17	46689708	46692301	HOXB8

Continued on next page

chr	start	end	gene
17	48211101	48227877	Ppp1r9b
17	48638448	48704542	CACNA1G
17	55162552	55198709	Akap1
17	56378595	56406152	BZRAP1
17	56597615	56618179	Sept4
17	71161159	71168060	Sstr2
17	79008946	79091232	Baiap2
18	3498836	4455266	DLGAP1
18	7567313	8406858	Ptprm
18	47793255	47808144	MBD1
18	25530929	25616539	CDH2
18	52889561	53257045	TCF4
18	26000000	28000000	chr18_27Mb
19	463346	474983	ODF3L2
19	1576677	1592710	MBD3
19	2100993	2151556	Ap3d1
19	17712137	17799401	Unc13a
19	19040009	19052041	Homer3
19	35629731	35645205	Fxyd7
19	42502472	42569957	Grik5
19	45409038	45412649	APOE
19	46272975	46285815	DMPK
19	47222769	47250251	Strn4
19	49617617	49621716	Lin7b
19	50270179	50310367	Ap2a1
19	51165083	51223040	Shank1
19	54385466	54410901	Prkcc
19	8130287	8212381	FBN3
20	1959402	1974703	Pdyn
20	3052265	3053162	OXT
20	3063202	3065370	AVP
20	23016056	23017314	Sstr4
20	43248163	43280376	ADA
20	43514343	43537160	Ywhab
20	44802375	44937137	CDH22
20	57393973	57425958	GNAS1
20	61472467	61493115	Tcfl5
20	61904164	61921141	Arfgap1
21	34398238	34401500	OLIG2
21	44473300	44496472	CBS
22	19929262	19957496	COMT
22	24813708	24838324	Adora2a
22	24890076	24932591	Upb1
22	36960103	37098903	Cacng2
22	37621311	37640305	Rac2
22	40742503	40762573	ADSL
22	42372930	42394223	Sept3

Continued on next page

chr	start	end	gene
22	51113069	51171641	Shank3
22	19957420	19978335	ARVCF
22	20004536	20053447	C22orf25
22	19833661	19842462	C22orf29
22	19467413	19508134	CDC45L
22	19510546	19512860	CLDN5
22	19166988	19279239	CLTCL1
22	19117792	19132190	DGCR14
22	18893540	18899600	DGCR6
22	20301801	20307608	DGCR6L
22	20067833	20099398	DGCR8
22	19775933	19842462	GNB1L
22	19708290	19712297	GP1BB
22	19136504	19137796	GSC2
22	19318223	19419219	HIRA
22	20020661	20020742	MIR185
22	19420035	19423594	MRPL40
22	18900294	18923840	PRODH
22	20105023	20113711	RANBP1
22	20228939	20255816	RTN4R
22	19701986	19712297	SEPT5
22	19163094	19166301	SLC25A1
22	19744225	19771110	TBX1
22	20099397	20104768	TRMT2A
22	19118320	19120136	TSSK2
22	19863039	19929359	TXNRD2
22	19437463	19466738	UFD1L
22	20119364	20133958	ZDHHC8
X	5808083	6146706	NLGN4X
X	16141423	16171640	GRPR
X	18443724	18671747	CDKL5
X	23352984	23414917	PTCHD1
X	25021814	25034065	ARX
X	28605680	29974017	IL1RAPL1
X	31137344	33357726	DMD
X	43515408	43606067	MAOA
X	49061523	49089833	CACNA1F
X	50111896	50213737	Dgkk
X	66763873	66944119	AR
X	69664723	69725337	Dlgh3
X	70338405	70362303	MED12
X	70364710	70391049	Nlgn3
X	110537007	110655406	DCX
X	113818550	114144624	HTR2C
X	115301957	115306223	AGTR2
X	119243024	119249847	RHOXF1
X	122318095	122624760	Gria3

Continued on next page

chr	start	end	gene
X	146993508	147032639	FMR1
X	152953751	152962048	SLC6A8
X	153287263	153363188	MECP2
X	153626570	153632038	RPL10
X	154487526	154493852	RAB39B
Y	1664276	1711973	ASMT
Y	16634487	16955848	NLGN4Y

Table 5.10: Case cohort - Sample index for the case cohort. Age and sex information is noted where available.

No.	Index code	Code Labor	Age	Sex
1	1_1	ZF-HST-001-01	68	male
2	10_1	ZF-HST-010-01	19	female
3	100_1	ZF-KE-100-01	22	male
4	101_1	ZF-KE-101-01	28	male
5	103_1	ZF-KE-103-01	35	male
6	104_1	ZF-KE-104-01		male
7	106_1	ZF-KE-106-01	43	male
8	107_1	ZF-KE-107-01	31	female
9	109_1	ZF-KE-109-01	31	female
10	110_1	ZF-KE-110-01	51	female
11	111_1	ZF-KE-111-01	39	female
12	112_1	ZF-KE-112-01	69	male
13	120_1	ZF-KE-120-01	24	male
14	123_1	ZF-KE-123-01	40	male
15	126_1	ZF-KE-126-01	23	female
16	127_1	ZF-KE-127-01	28	female
17	129_1	ZF-KE-129-01	52	female
18	131_1	ZF-KE-131-01	42	female
19	132_1	ZF-KE-132-01	57	male
20	134_1	ZF-KE-134-01	26	male
21	135_1	ZF-KE-135-01		female
22	137_1	ZF-KE-137-01	24	male
23	138_1	ZF-KE-138-01	29	female
24	139_1	ZF-KE-139-01	42	male
25	2_1	ZF-HST-002-01	28	male
26	202_1	ZF-KJP-202-01	13	male
27	204_1	ZF-KJP-204-01	14	female
28	205_1	ZF-KJP-205-01	13	male
29	207_1	ZF-KJP-207-01	16	male
30	215_1	ZF-KJP-215-01	13	female
31	24_1	ZF-HST-024-01	61	male
32	26_1	ZF-HST-026-01	48	female
33	28_1	ZF-HST-028-01	49	female
34	3_1	ZF-HST-003-01	31	male
35	309_1	ZF-BN-309-01		

Continued on next page

No.	Index code	Code Labor	Age	Sex
36	311_1	ZF-BN-311-01		
37	315_1	ZF-BN-315-01	27	female
38	317_1	ZF-BN-317-01	27	male
39	318_1	ZF-BN-318-01	57	female
40	32_1	ZF-HST-032-01	30	female
41	321_1	ZF-BN-321-01	28	male
42	322_1	ZF-BN-322-01	28	female
43	324_1	ZF-BN-324-01	19	female
44	326_1	ZF-BN-326-01	42	female
45	327_1	ZF-BN-327-01	29	male
46	329_1	ZF-BN-329-01	57	female
47	33_1	ZF-HST-033-01	38	female
48	331_1	ZF-BN-331-01	32	female
49	332_1	ZF-BN-332-01	34	female
50	333_1	ZF-BN-333-01	34	female
51	334_1	ZF-BN-334-01	33	female
52	337_1	ZF-BN-337-01	24	female
53	342_1	ZF-BN-342-01	26	female
54	345_1	ZF-BN-345-01	30	male
55	346_1	Z-BN-071-1		
56	349_1	ZF-BN-349-01	51	female
57	35_1	ZF-HST-035-01	36	female
58	350_1	ZF-BN-350-01	43	female
59	351_1	ZF-BN-351-01	31	male
60	356_1	Z-BN-096-1		
61	357_1	ZF-BN-357-01		
62	36_1	ZF-HST-036-01	39	female
63	365_1	ZF-BN-365-01	57	male
64	369_1	ZF-BN-369-01	32	female
65	37_1	ZF-HST-037-01	28	female
66	371_1	ZF-BN-371-01		
67	374_1	ZF-BN-374-01		
68	380_1	ZF-BN-380-01	28	male
69	382_1	ZF-BN-382-01	37	female
70	383_1	ZF-BN-383-01	24	male
71	384_1	ZF-BN-384-01	38	male
72	385_1	ZF-BN-385-01	36	female
73	386_1	ZF-BN-386-01		
74	387_1	ZF-BN-387-01		
75	388_1	ZF-BN-388-01		
76	389_1	ZF-BN-389-01		male
77	390_1	ZF-BN-390-01		
78	391_1	ZF-BN-391-01		
79	393_1	Z-BN-093-1		
80	394_1	ZF-BN-394-01		
81	395_1	ZF-BN-395-01	27	female
82	396_1	Z-BN-103-1		

Continued on next page

No.	Index code	Code Labor	Age	Sex
83	398_1	ZF-BN-398-01	28	male
84	4_1	ZF-HST-004-01	41	female
85	40_1	ZF-HST-040-01	51	female
86	402_1	ZF-BN-402-01		
87	406_1	ZF-BN-406-01	22	female
88	41_1	ZF-HST-041-01	27	male
89	411_1	ZF-BN-411-01		
90	412_1	ZF-BN-412-01		
91	413_1	ZF-BN-413-01		
92	414_1	ZF-BN-414-01	19	male
93	415_1	ZF-BN-415-01		
94	417_1	ZF-BN-417-01	21	male
95	418_1	ZF-BN-418-01	20	female
96	42_1	ZF-HST-042-01	37	male
97	420_1	ZF-BN-420-01		
98	422_1	Z-BN-054-1		
99	423_1	ZF-BN-423-01	53	female
100	428_1	ZF-BN-428-01	42	female
101	429_1	ZF-BN-429-01	48	male
102	431_1	ZF-BN-431-01	57	female
103	433_1	ZF-BN-433-01		
104	438_1	Z-BN-044-1		
105	439_1	ZF-BN-439-01		
106	440_1	ZF-BN-440-01	20	female
107	441_1	ZF-BN-441-01	38	female
108	443_1	ZF-BN-443-01		
109	445_1	Z-BN-026-1		
110	447_1	Z-BN-023-1		
111	448_1	ZF-BN-448-01		
112	449_1	Z-BN-025-1		
113	450_1	Z-BN-029-1		
114	451_1	Z-BN-033-1		
115	452_1	Z-BN-061-1		
116	453_1	Z-BN-068-1		
117	455_1	ZF-BN-455-01		
118	456_1	Z-BN-036-1		
119	457_1	ZF-BN-457-01		
120	458_1	Z-BN-020-1		
121	459_1	Z-BN-021-1		
122	461_1	ZF-BN-461-01		
123	462_1	Z-BN-059-1		
124	463_1	Z-BN-051-1		
125	464_1	Z-BN-031-1		
126	465_1	ZF-BN-465-01	36	
127	466_1	Z-BN-069-1		
128	467_1	Z-BN-062-1		
129	468_1	Z-BN-049-1		

Continued on next page

No.	Index code	Code Labor	Age	Sex
130	469_1	Z-BN-038-1		
131	470_1	Z-BN-091-1		
132	471_1	Z-BN-102-1		
133	472_1	Z-BN-108-1		
134	473_1	Z-BN-099-1		
135	474_1	ZF-BN-474-01		
136	475_1	ZF-BN-475-01		
137	476_1	Z-BN-042-1		
138	479_1	Z-BN-067-1		
139	523_1	ZF-HST-523-01	21	male
140	526_1	ZF-HST-526-01	26	female
141	528_1	ZF-HST-528-01	26	female
142	529_1	ZF-HST-529-01	40	female
143	537_1	ZF-HST-537-01	36	female
144	541_1	ZF-HST-541-01	32	female
145	543_1	ZF-HST-543-01	54	female
146	544_1	ZF-HST-544-01	47	male
147	548_1	ZF-HST-548-01	39	male
148	590_1	ZF-HST-590-01	54	male
149	592_1	ZF-HST-592-01	50	male
150	593_1	ZF-HST-593-01	51	female
151	594_1	ZF-HST-594-01	52	male
152	60_1	ZF-HST-060-01	43	female
153	61_1	ZF-HST-061-01	26	female
154	63_1	ZF-HST-063-01	42	female
155	700_1	ZF-HOM-700-01	18	male
156	701_1	ZF-HOM-701-01	38	female
157	703_1	ZF-HOM-703-01	32	female
158	705_1	ZF-HOM-705-01	21	female
159	707_1	ZF-HOM-707-01	45	male
160	708_1	ZF-HOM-708-01	39	female
161	709_1	ZF-HOM-709-01	31	female
162	710_1	ZF-HOM-710-01	33	male
163	713_1	ZF-HOM-713-01	36	female
164	715_1	ZF-HOM-715-01	22	male
165	717_1	ZF-HOM-717-01	29	male
166	900_1	Z-BN-048-1		female
167	901_1	Z-BN-050-1		male
168	902_1	Z-BN-053-1		female
169	903_1	Z-BN-055-1		female
170	904_1	Z-BN-056-1		
171	905_1	Z-BN-058-1		female
172	906_1	ZF-BN-906-01		
173	907_1	Z-BN-060-1		male
174	908_1	ZF-BN-908-01		
175	909_1	Z-BN-028-1		
176	910_1	ZBN-035-1		male

Continued on next page

No.	Index code	Code Labor	Age	Sex
177	911_1	ZBN-039-1		male
178	913_1	ZBN-045-1		female
179	914_1	ZBN-046-1		male
180	915_1	ZBN-047-1		male
181	916_1	ZBN-022-1		
182	920_1	ZBN-043-1		male
183	921_1	ZBN-052-1		female
184	923_1	ZF-BN-923-01		
185	924_1	ZBN-101-1		
186	925_1	ZBN-080-1		
187	927_1	ZBN-076-1		
188	928_1	ZF-BN-928-01		
189	929_1	ZF-BN-929-01		
190	931_1	ZF-BN-931-01		
191	932_1	ZBN-095-1		
192	933_1	ZBN-094-1		
193	935_1	ZF-BN-935-01		
194	937_1	ZBN-084-1		
195	938_1	ZBN-088-1		
196	946_1	ZF-BN-946-01	21	female
197	947_1	ZF-BN-947-01	56	male
198	948_1	ZF-BN-948-01	20	female
199	949_1	ZF-BN-949-01	33	male
200	950_1	ZF-BN-950-01	46	male
201	951_1	ZF-BN-951-01	23	female

Table 5.11: Control cohort - Sample index for the control cohort. Age and sex information is noted where available.

No.	Code	Age	Sex
1	P009	58	Female
2	P021	23	Male
3	P043	24	Female
4	P044	36	Male
5	P047	46	Female
6	P049	48	Female
7	P052	32	Male
8	P055	48	Female
9	P058	71	Female
10	P060	35	Male
11	P067	78	Female
12	P070	43	Female
13	P071	56	Female
14	P080	39	Male
15	P081	26	Female
16	P083	57	Male
17	P084	34	Male

Continued on next page

No.	Code	Age	Sex
18	P085	43	Female
19	P087	40	Male
20	P088	33	Male
21	P090	40	Female
22	P091	31	Female
23	P092	49	Male
24	P094	37	Male
25	P096	56	Male
26	P097	34	Male
27	P098	35	Male
28	P100	40	Female
29	P101	49	Female
30	P102	58	Male
31	P103	47	Male
32	P104	34	Female
33	P105	46	Male
34	P106	45	Female
35	P107	54	Male
36	P108	60	Male
37	P109	40	Female
38	P110	47	Female
39	P113	31	Male
40	P116	33	Male
41	P119	34	Male
42	P123	27	Female
43	P125	31	Female
44	P126	51	Female
45	P127	39	Female
46	P128	50	Female
47	P129	51	Female
48	P130	43	Male
49	P131	55	Female
50	P133	31	Female
51	P135	30	Male
52	P136	36	Male
53	P138	49	Female
54	P140	46	Male
55	P141	39	Male
56	P144	48	Male
57	P145	42	Female
58	P146	33	Female
59	P147	35	Female
60	P148	59	Female
61	P149	37	Male
62	P150	52	Female
63	P151	36	Male
64	P152	39	Male

Continued on next page

No.	Code	Age	Sex
65	P154	33	Male
66	P155	45	Male
67	P157	53	Female
68	P159	38	Male
69	P160	24	Female
70	P161	60	Male
71	P162	58	Male
72	P163	37	Female
73	P164	55	Female
74	P165	39	Male
75	P166	49	Female
76	P167	48	Male
77	P168	45	Female
78	P169	53	Male
79	P170	44	Male
80	P171	53	Male
81	P172	55	Male
82	P173	42	Female
83	P175	34	Female
84	P176	43	Female
85	P177	40	Male
86	P179	45	Male
87	P180	48	Female
88	P181	43	Female
89	P182	52	Male
90	P183	59	Male
91	P184	44	Female
92	P185	36	Male
93	P186	39	Female
94	P187	57	Female
95	P188	48	Male
96	P189	36	Female
97	P190	56	Female
98	P191	43	Male
99	P192	55	Male
100	P195	43	Male
101	P196	46	Female
102	P197	42	Female
103	P198	36	Female
104	P200	52	Female
105	P201	50	Male
106	P202	50	Male
107	P203	37	Female
108	P204	33	Male
109	P206	40	Female
110	P207	49	Male
111	P208	45	Female

Continued on next page

No.	Code	Age	Sex
112	P209	36	Male
113	P210	32	Male
114	P211	40	Female
115	P212	35	Male
116	P213	49	Male
117	P214	48	Male
118	P215	43	Female
119	P216	31	Female
120	P217	48	Female
121	P218	39	Female
122	P219	45	Female
123	P220	31	Male
124	P221	44	Female
125	P223	60	Female
126	P224	43	Male
127	ZF-BN-301-01	25	Female
128	ZF-BN-304-01	28	Female
129	ZF-BN-305-01	31	Male
130	ZF-BN-306-01	28	Male
131	ZF-BN-310-4	24	Female
132	ZF-BN-312-01	26	Female
133	ZF-BN-313-01	28	Male
134	ZF-BN-320-01	25	Male
135	ZF-BN-323-01	35	Female
136	ZF-BN-330-01	31	Female
137	ZF-BN-340-01	34	Female
138	ZF-BN-341-01	27	Male
139	ZF-BN-343-01	19	Female
140	ZF-BN-348-01	28	Female
141	ZF-BN-354-01	22	Female
142	ZF-BN-358-01	34	Female
143	ZF-BN-359-4	37	Male
144	ZF-BN-363-01	20	Male
145	ZF-BN-366-4	41	Female
146	ZF-BN-381-01	35	Male
147	ZF-BN-397-01	32	Female
148	ZF-BN-400-01	31	Female
149	ZF-BN-403-01	58	Male
150	ZF-BN-404-01	43	Female
151	ZF-BN-407-01	17	Male
152	ZF-BN-442-2	53	Male
153	ZF-BN-480-2	59	Male
154	ZF-HOM-804-01	31	Female
155	ZF-HST-014-01	79	Male
156	ZF-HST-034-5	38	Female
157	ZF-HST-056-01	45	Male
158	ZF-HST-510-4	59	Female

Continued on next page

No.	Code	Age	Sex
159	ZF-HST-531-01	37	Male
160	ZF-HST-533-5	23	Male
161	ZF-HST-535-5	72	Female
162	ZF-HST-539-01	58	Male
163	ZF-HST-540-01	70	Male
164	ZF-HST-566-7	47	Female
165	ZF-HST-581-01	44	Male
166	ZF-KE-0113-01	34	Female
167	ZF-KE-0115-01	24	Female
168	ZF-KE-0116-01	29	Female
169	ZF-KE-0122-01	37	Female
170	ZF-KE-0125-4	18	Female
171	AARI-MEK-282-4	45	Female
172	ADBS-MEK-292-4	52	Female
173	AEHS-MEK-559-4	52	Female
174	CESZ-MEK-291-4	42	Female
175	EHSN-MEK-285-4	53	Female
176	ERHH-MEK-279-4	41	Female
177	GAMR-MEK-296-4	63	Female
178	GARG-MEK-333-4	49	Female
179	HNUN-MEK-272-4	63	Male
180	HRBS-MEK-293-4	53	Male
181	INJE-MEK-280-4	56	Male
182	ISKR-MEK-268-4	46	Female
183	JNHR-MEK-302-4	25	Male
184	JSSR-MEK-328-4	53	Male
185	KAWG-MEK-281-4	50	Female
186	KSMR-MEK-283-4	64	Male
187	MAGR-MEK-273-4	41	Female
188	MAHZ-MEK-274-4	52	Female
189	OC-ADRR-UKB-926-4	56	Male
190	OC-CANS-UKB-901-4	50	Female
191	OC-CEER-UKB-934-5	38	Female
192	OC-CSHE-UKB-124-4	56	Male
193	OC-DABS-UKB-126-4	43	Female
194	OC-DLSS-UKB-137-5	27	Male
195	OC-DSKL-UKB-087-4	47	Female
196	OC-EABL-UKB-115-4	75	Female
197	OC-EDKN-UKB-154-4	57	Male
198	OC-EEGL-UKB-117-4	53	Female
199	OC-EEMR-UKB-026-4	45	Female
200	OC-EHDR-UKB-092-4	56	Female
201	OC-EHHN-UKB-091-4	69	Female

5.6.8 Coverage and experimental quality (per pool statistics)

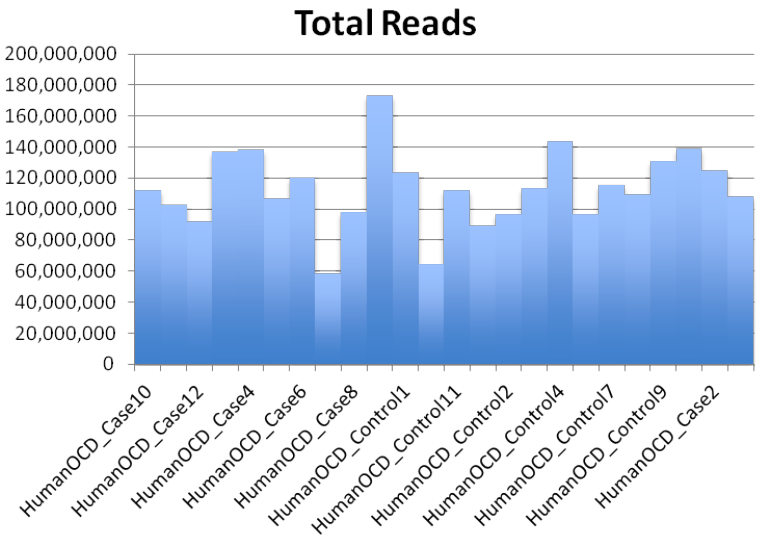


Figure 5.28: Number of total reads per pool - The x-axis shows each QC+ pool and the y-axis shows the number of total reads.

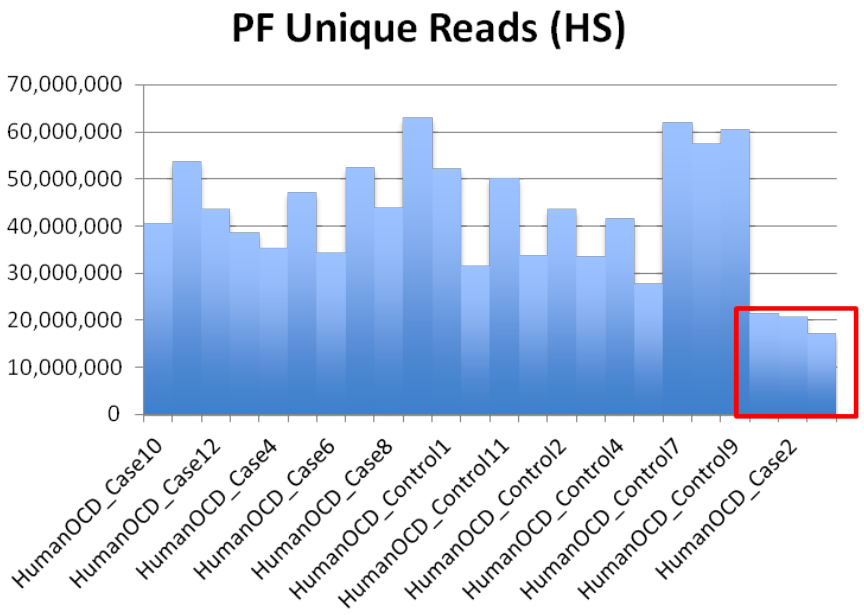


Figure 5.29: Number of unique reads per pool - This barplot shows the number of the unique reads after excluding duplicate reads. The three pools that are indicated with a red box have much fewer unique reads compared to the other pools.

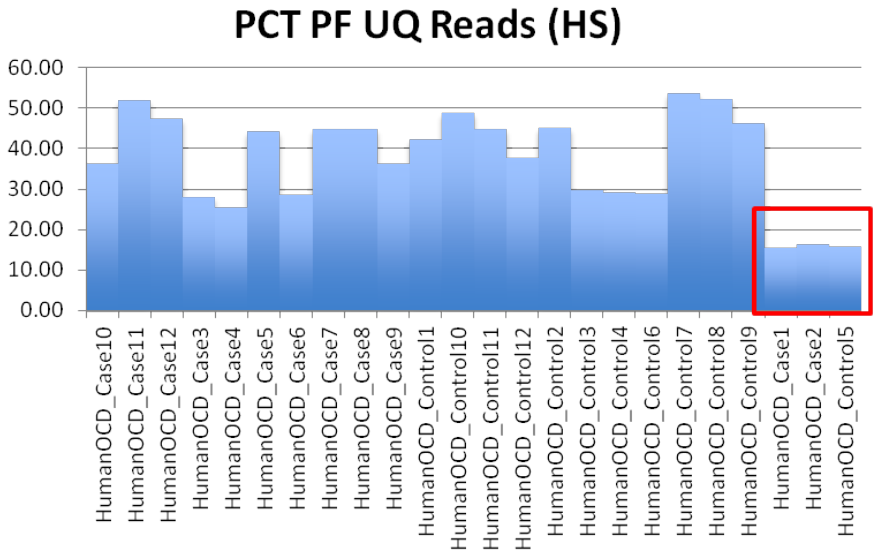


Figure 5.30: Percentage of unique reads to total reads per pool - The pools that have particularly low percentages of unique reads are indicated with a red box.

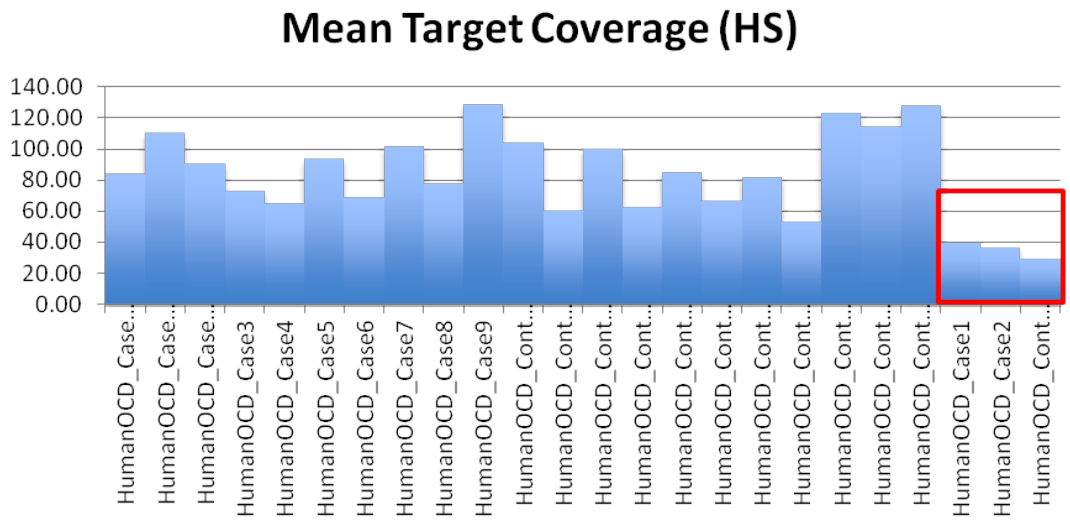


Figure 5.31: Mean target coverage per pool - Mean target coverage per pool ranges from 30x to 130x.

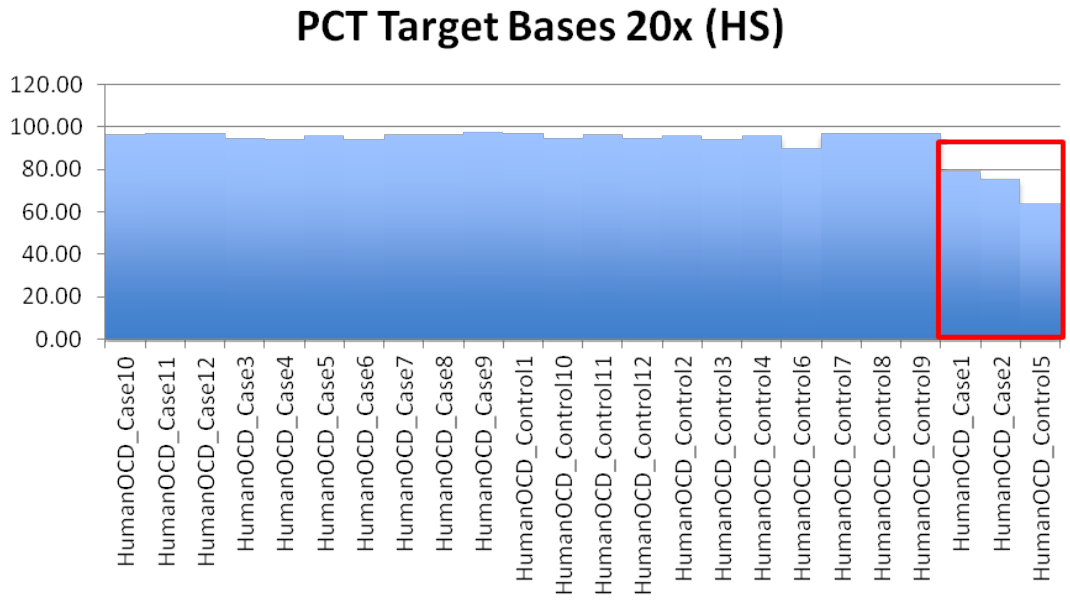


Figure 5.32: Percent target bases with >20x read depth - 21 of 24 pools achieved a coverage of at least 20x in most of the target region.

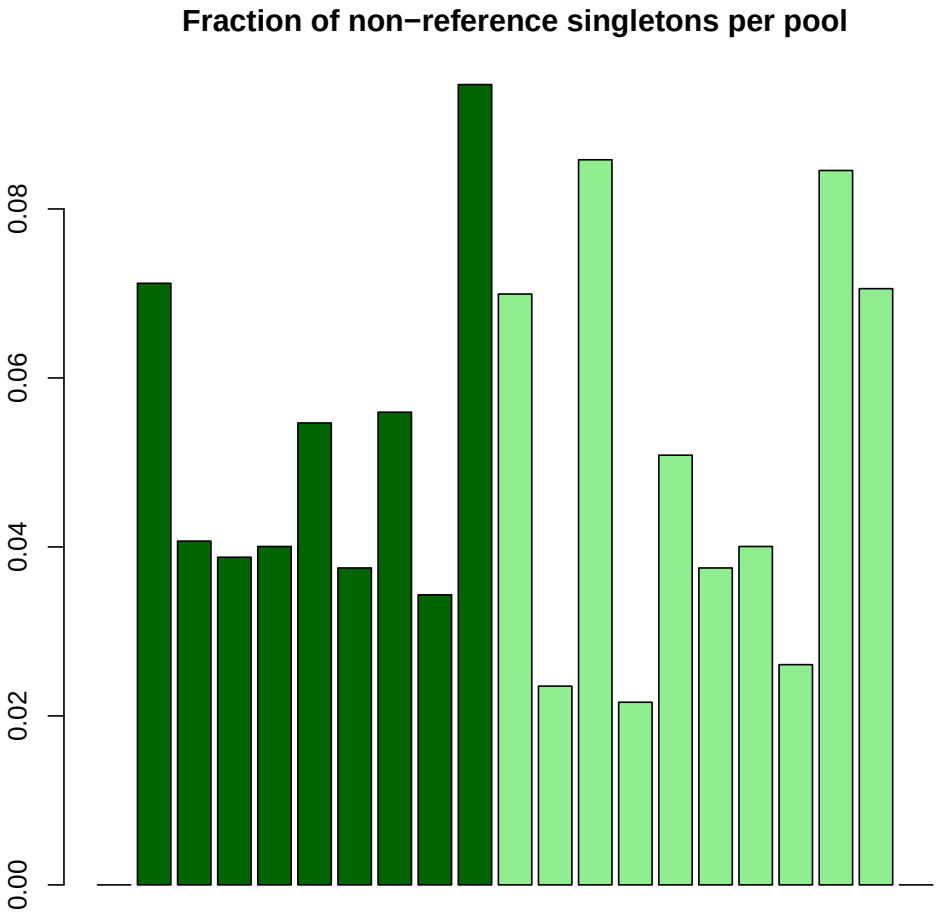


Figure 5.33: Fraction of detected singletons (per pool) to total number of discovered variants - Dark green bars indicate 10 QC+ case pools, while light green bars indicate 11 QC+ control pools.

5.6.9 Variant discovery per pool

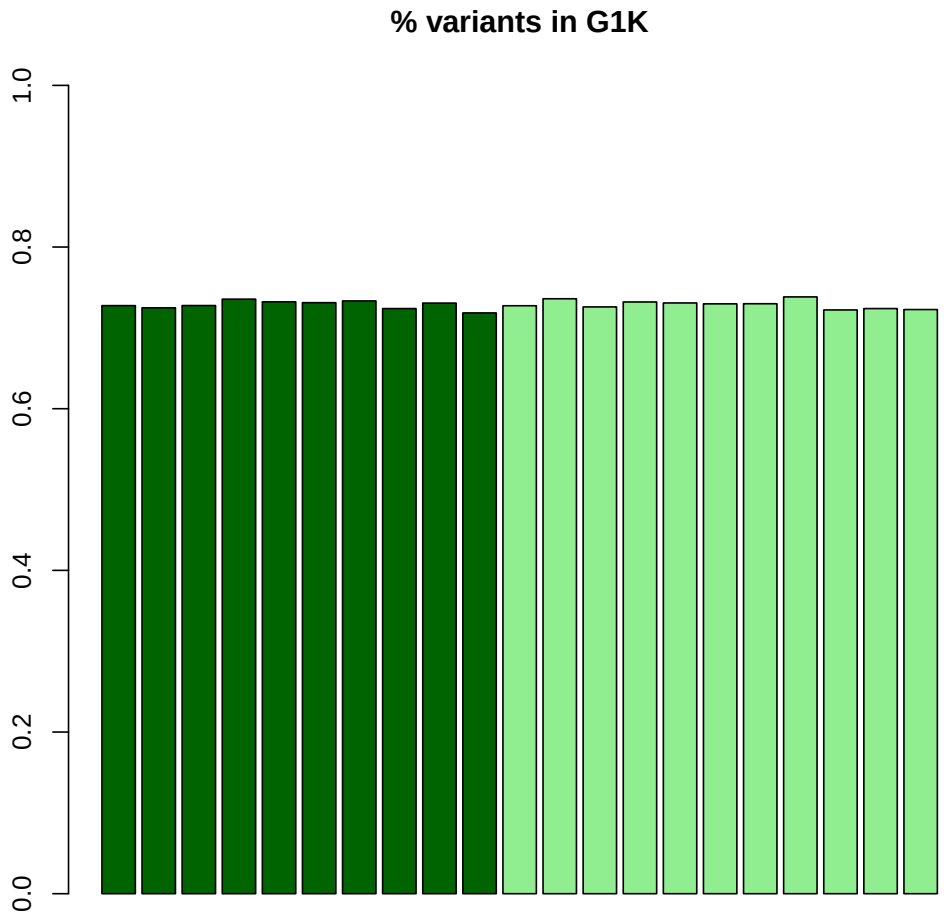


Figure 5.34: Fraction of known variants from the 1000 genome project to the total number of discovered variants per pool - see Appendix Figure 5.33.

5.6.10 Associated genes from burden test

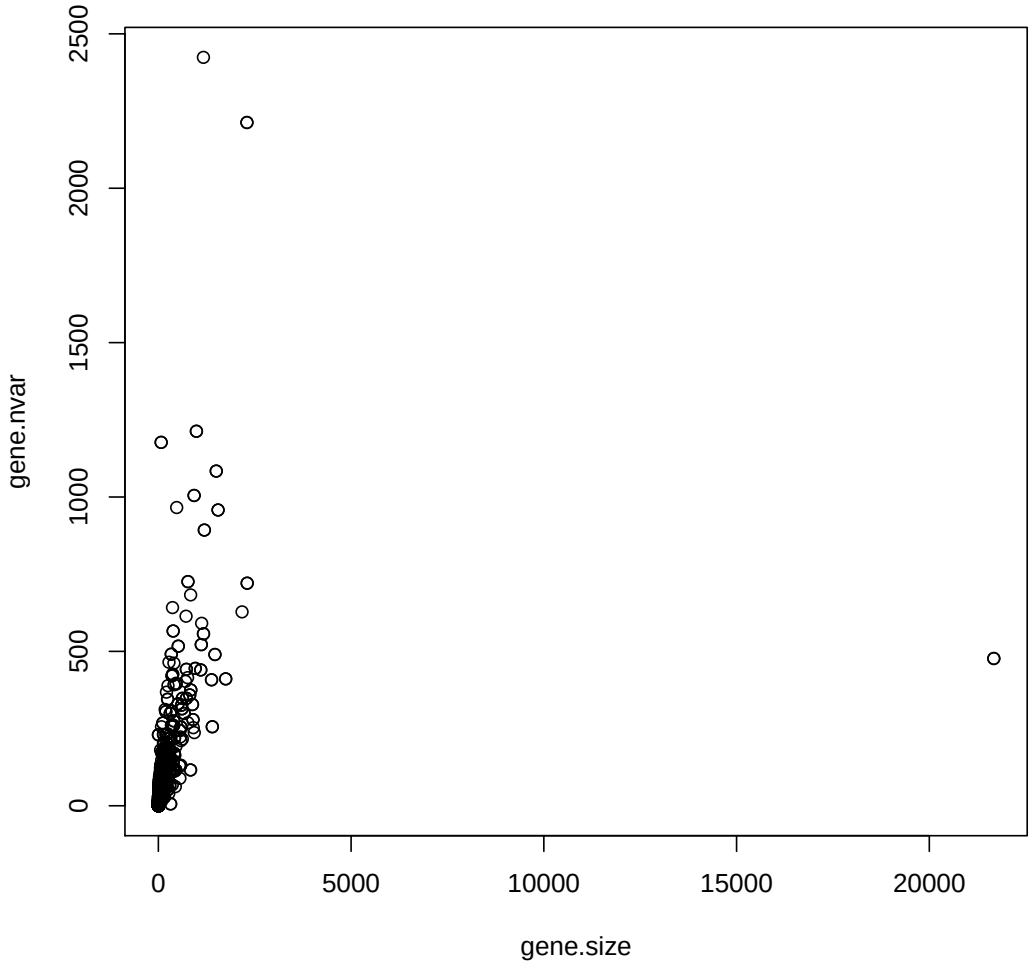


Figure 5.35: Number of variants found per gene and its size for all 611 genes - The bigger the gene, the more variants tend to be found.

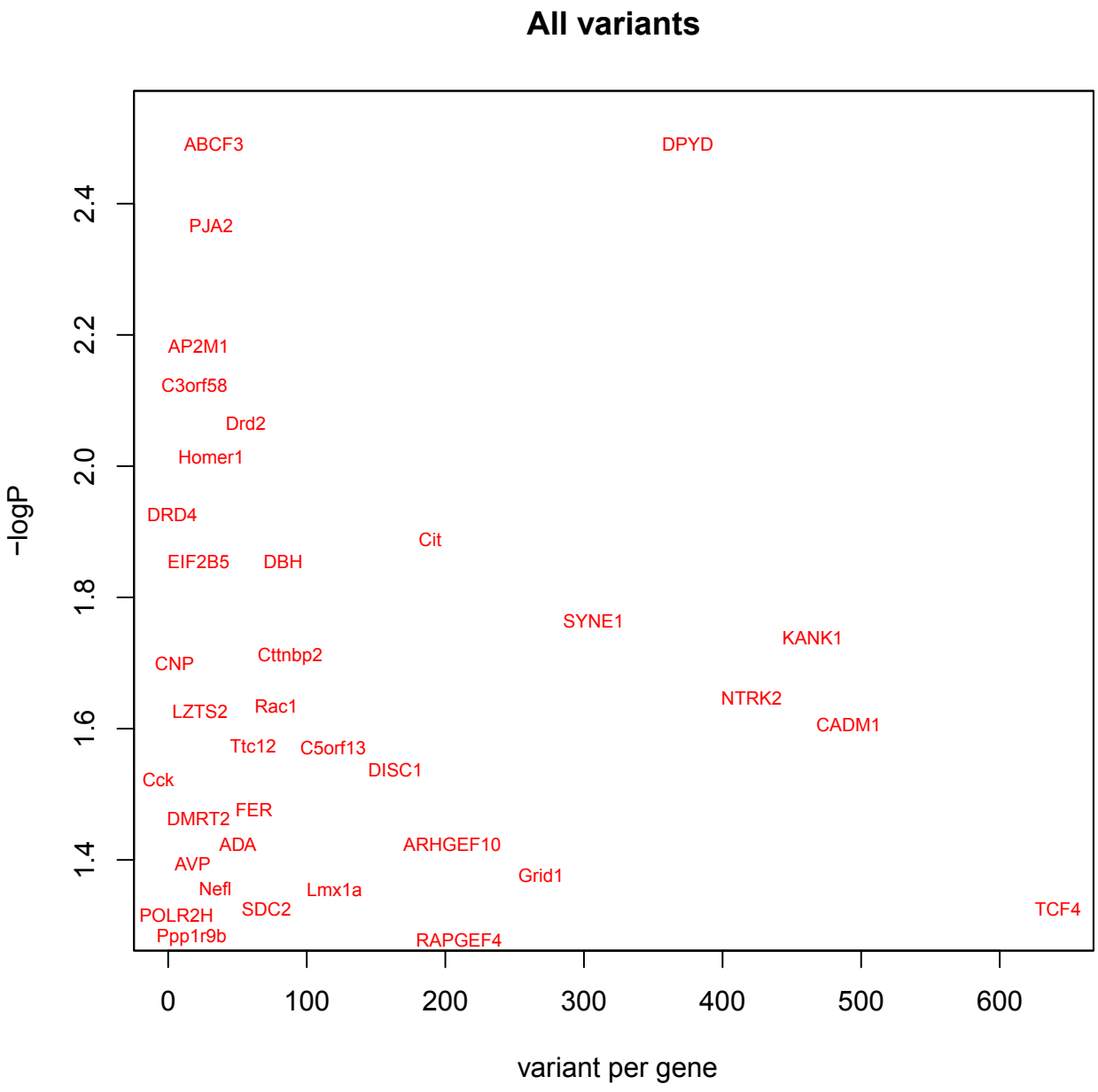


Figure 5.36: Associated genes from burden test (all variants) - The x-axis shows the number of variants per gene, while the y-axis shows $-log_{10}P$.

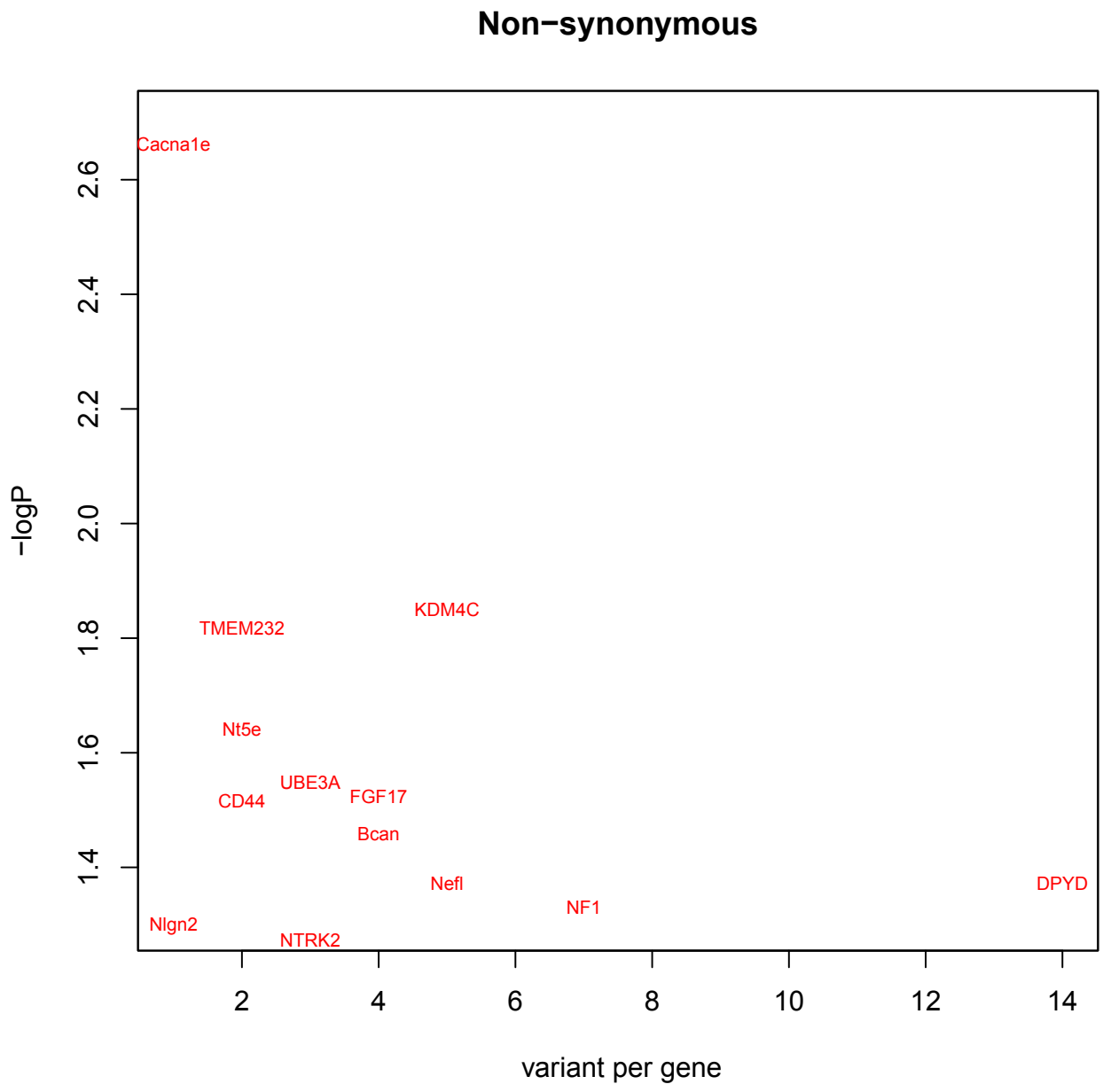


Figure 5.37: Associated genes from burden test (nonsynonymous variants) - See Appendix Figure 5.36.

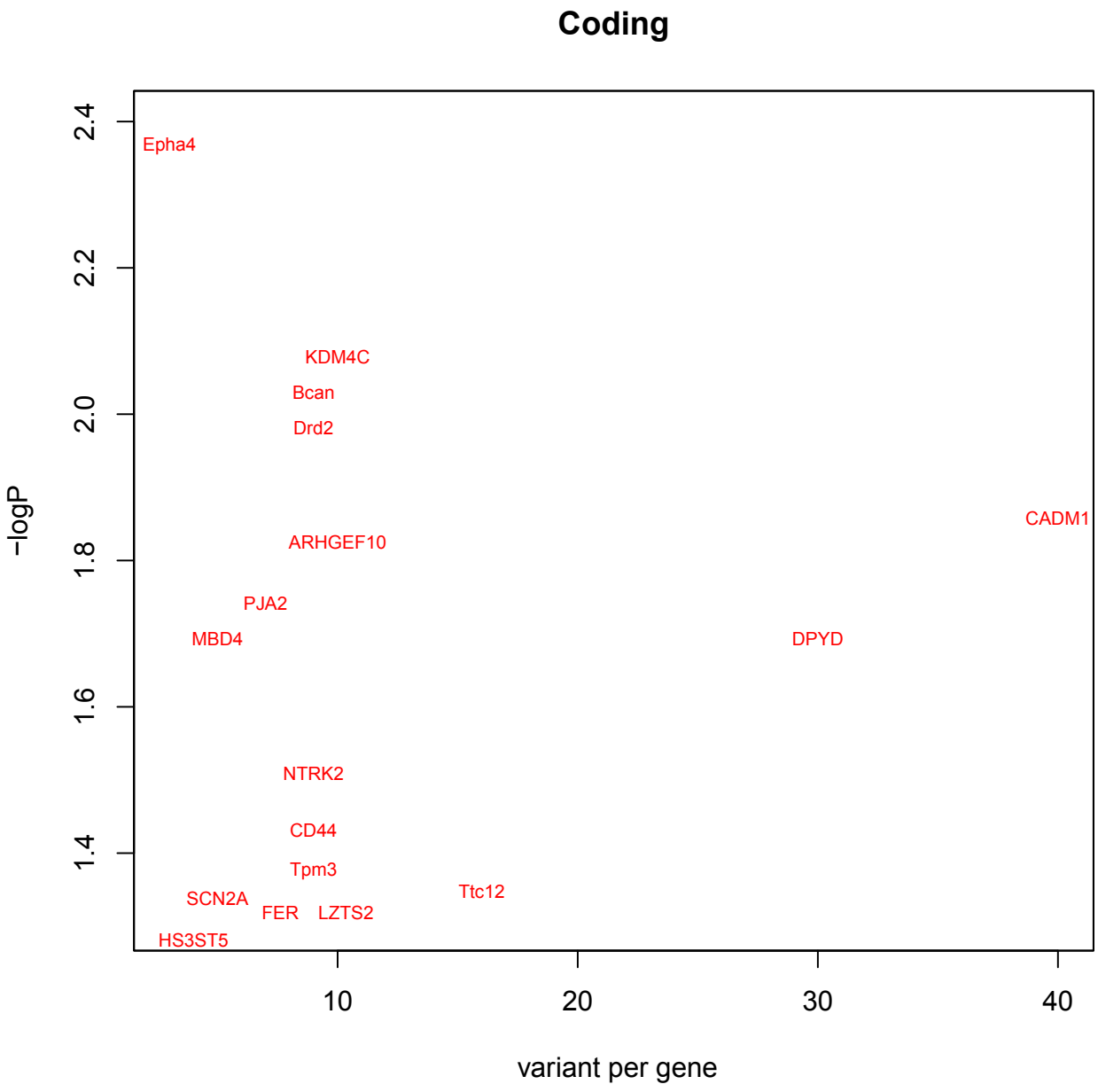


Figure 5.38: Associated genes from burden test (coding variants) - See Appendix Figure 5.36.

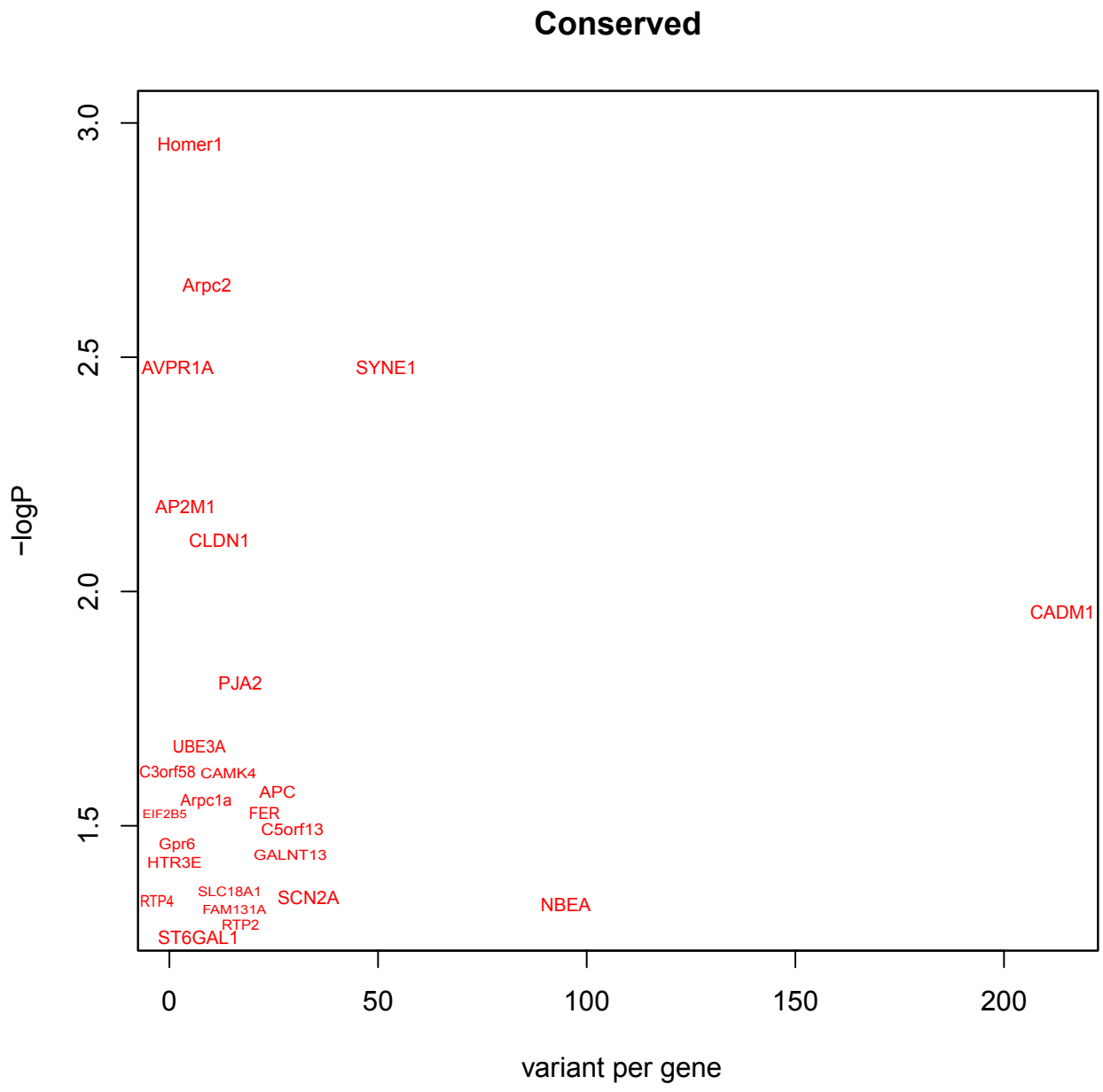


Figure 5.39: Associated genes from burden test (variants in evolutionarily conserved region) - See Appendix Figure 5.36.

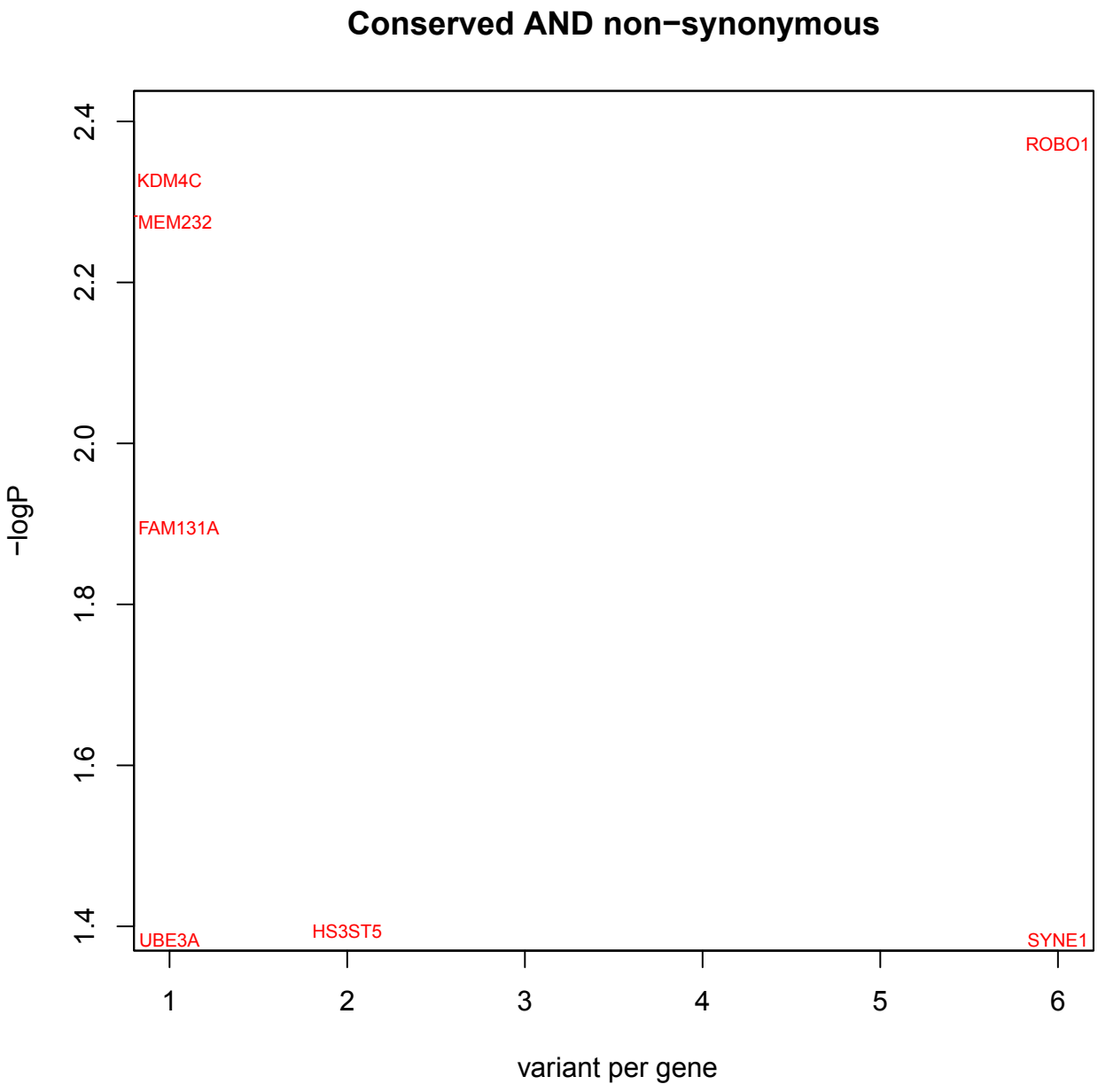


Figure 5.40: Associated genes from burden test (nonsynonymous variants in evolutionarily conserved region) - See Appendix Figure 5.36.

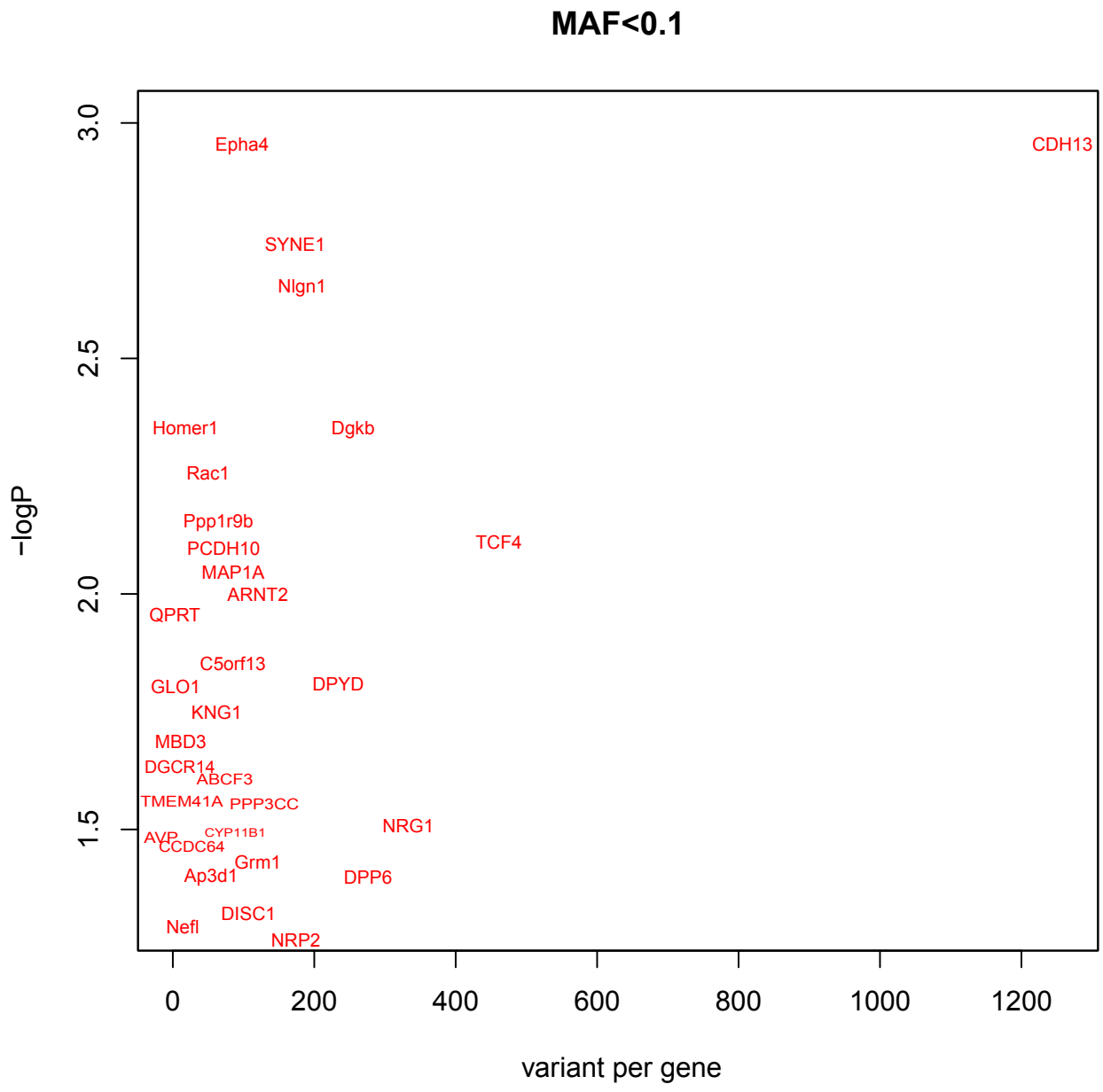


Figure 5.41: Associated genes from burden test (MAF< 10%) - See Appendix Figure 5.36.

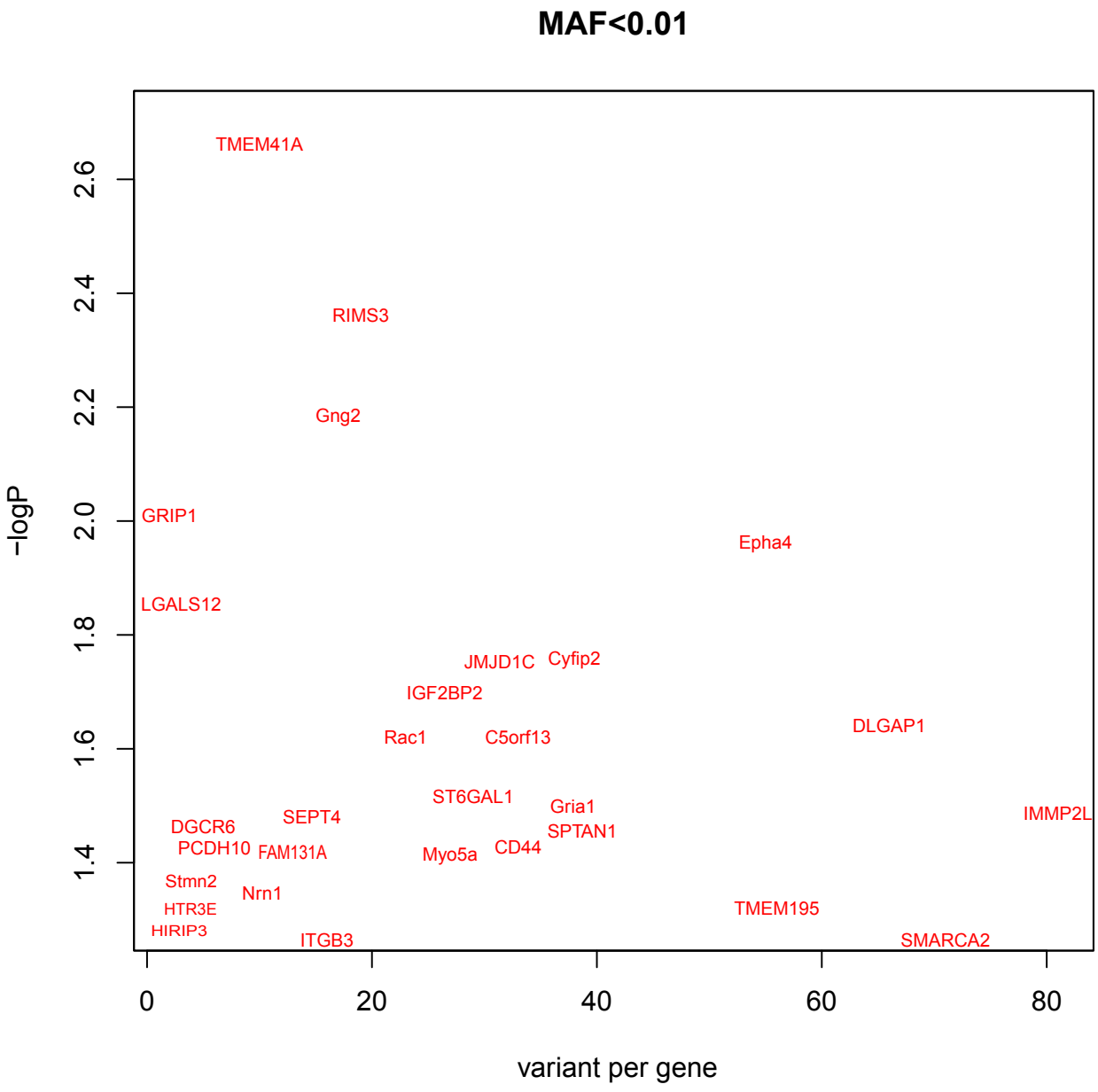


Figure 5.42: Associated genes from burden test (MAF< 1%) - See Appendix Figure 5.36.

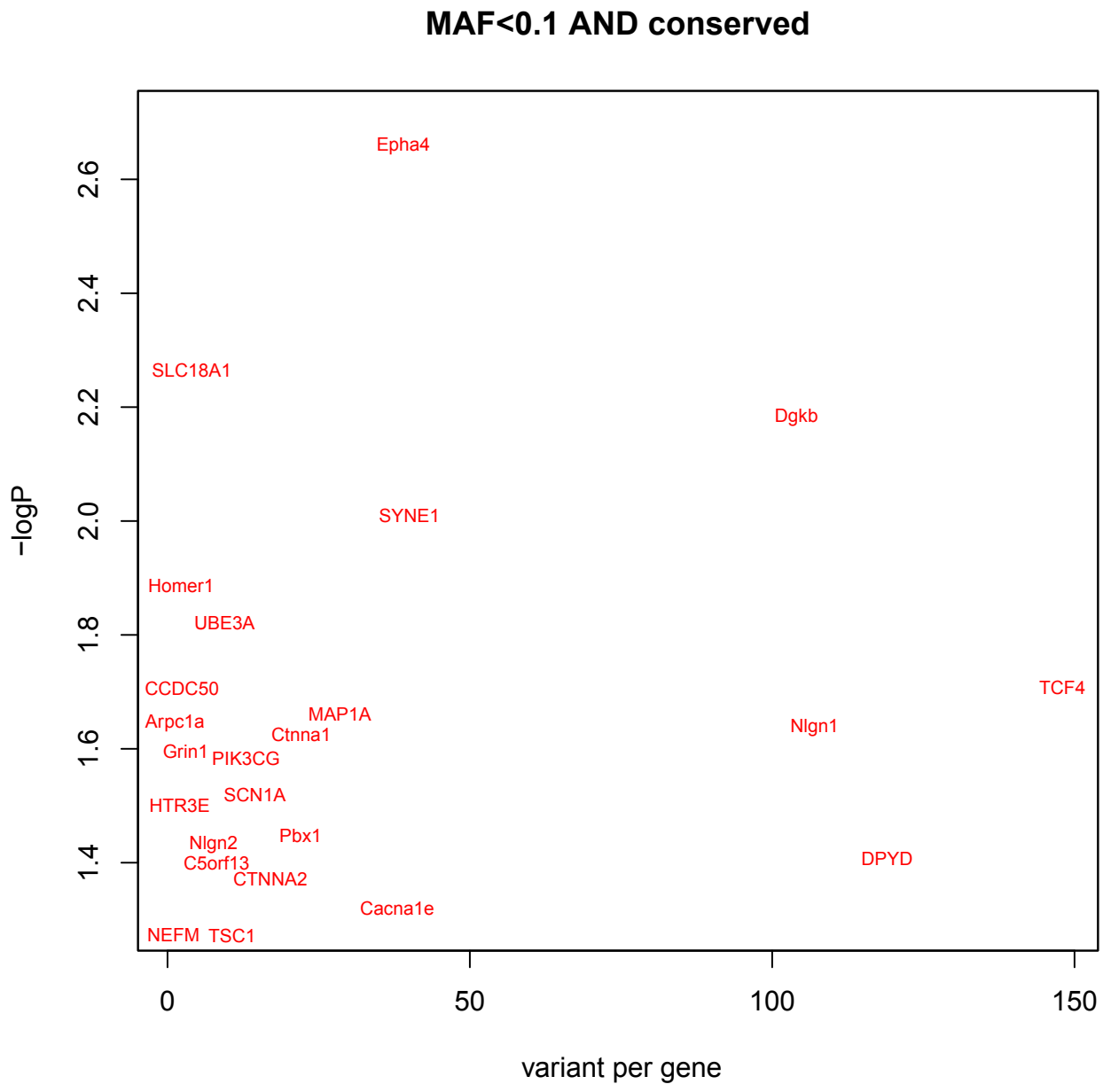


Figure 5.43: Associated genes from burden test (MAF<10% in conserved region) - See Appendix Figure 5.36.

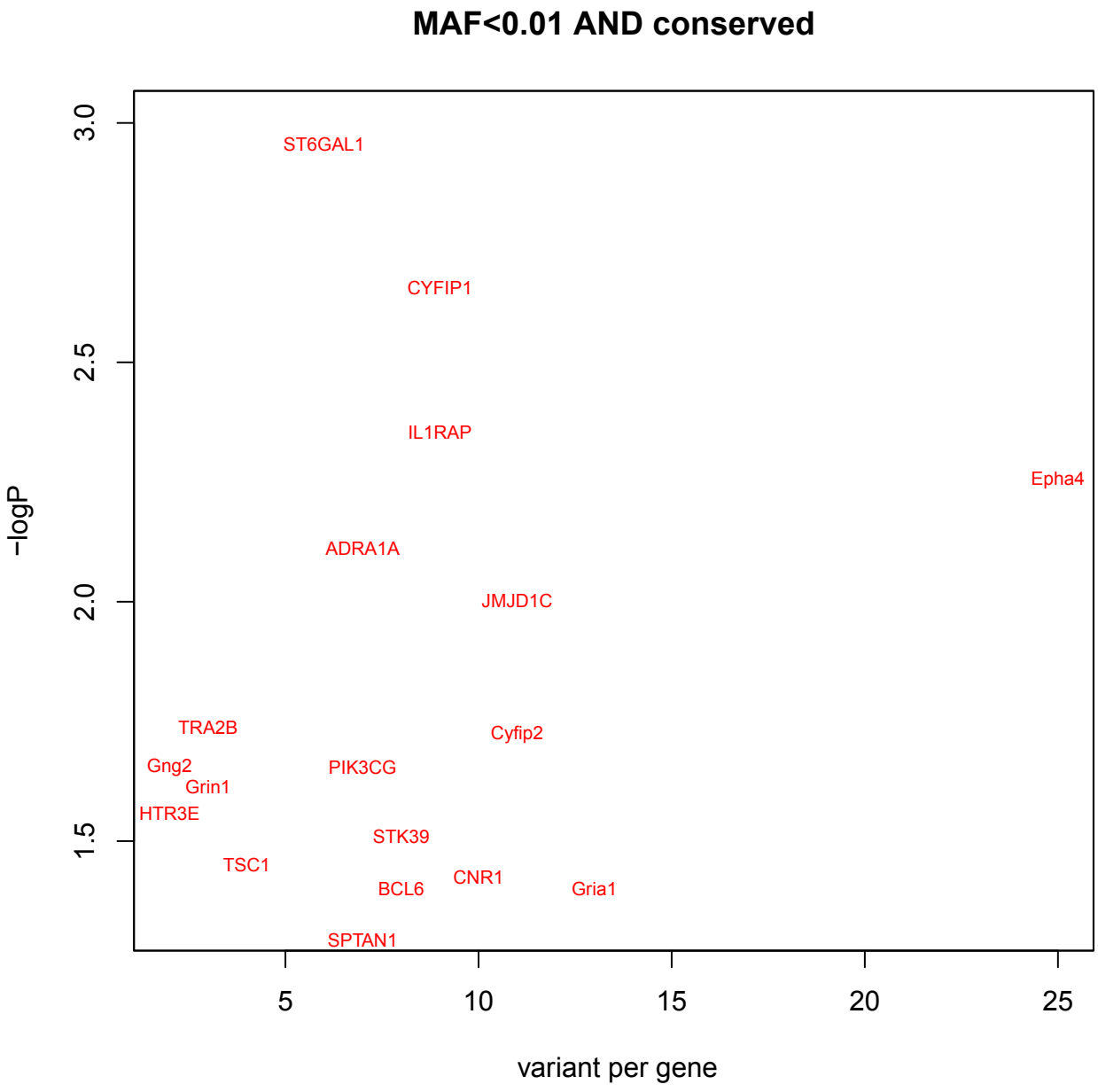


Figure 5.44: Associated genes from burden test (MAF<1% in conserved region) - See Appendix Figure 5.36.

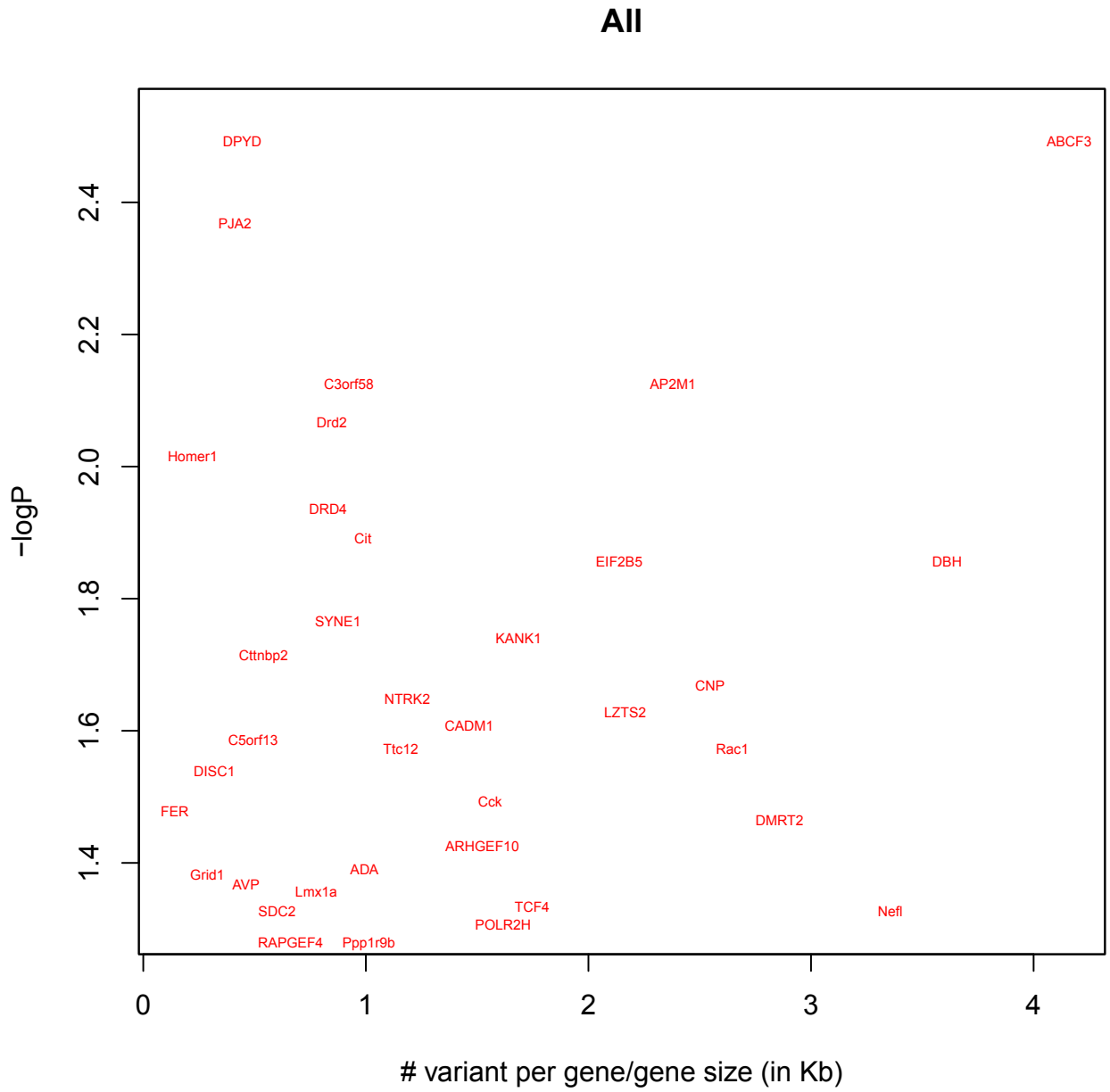


Figure 5.45: Associated genes from burden test (all variants, adjusted for gene size) - The x-axis shows the number of variants per 1kb per gene, while the y-axis shows $-log_{10}P$.

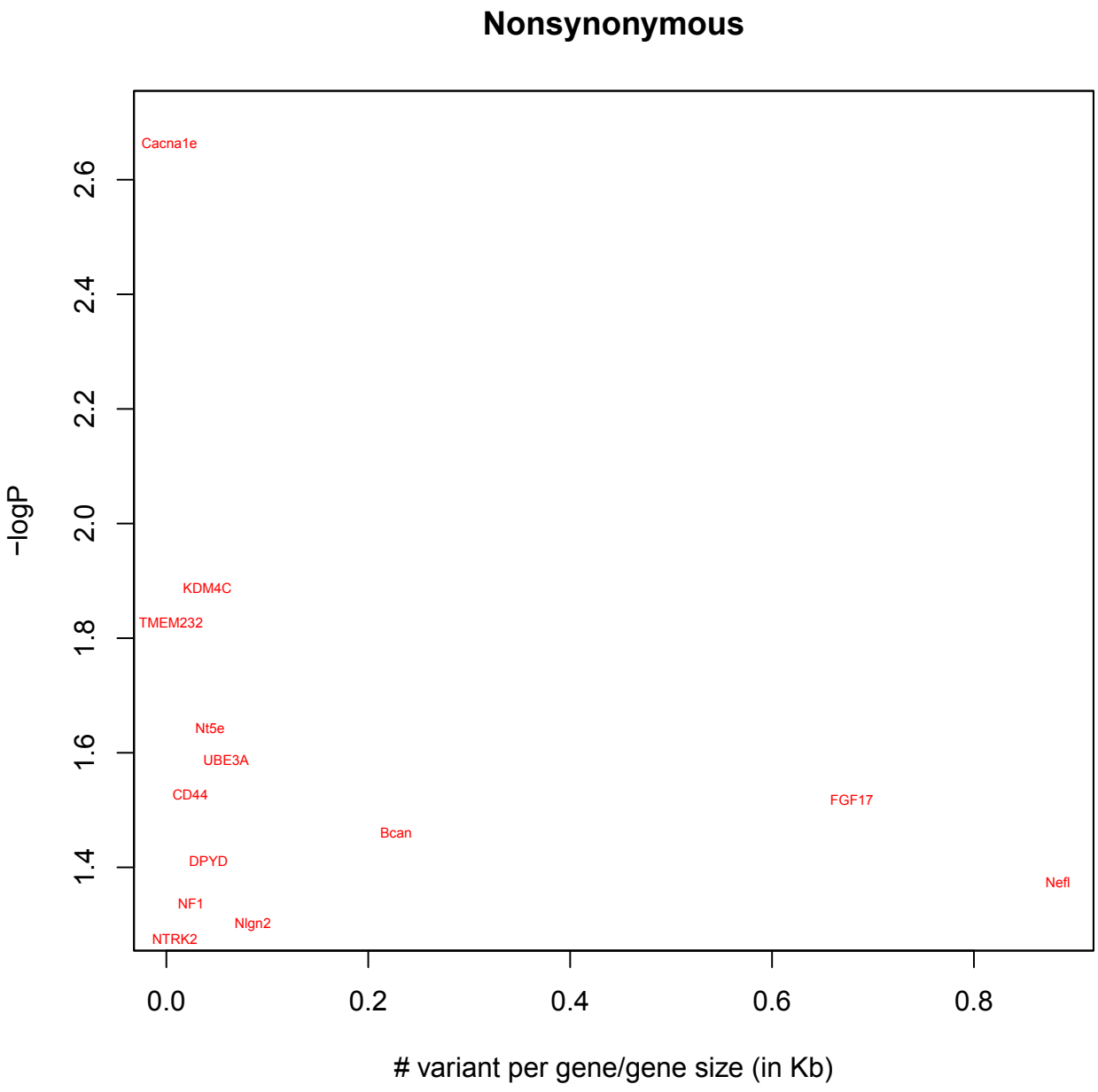


Figure 5.46: Associated genes from burden test (nonsynonymous variants, adjusted for gene size) - See Appendix Figure 5.45.

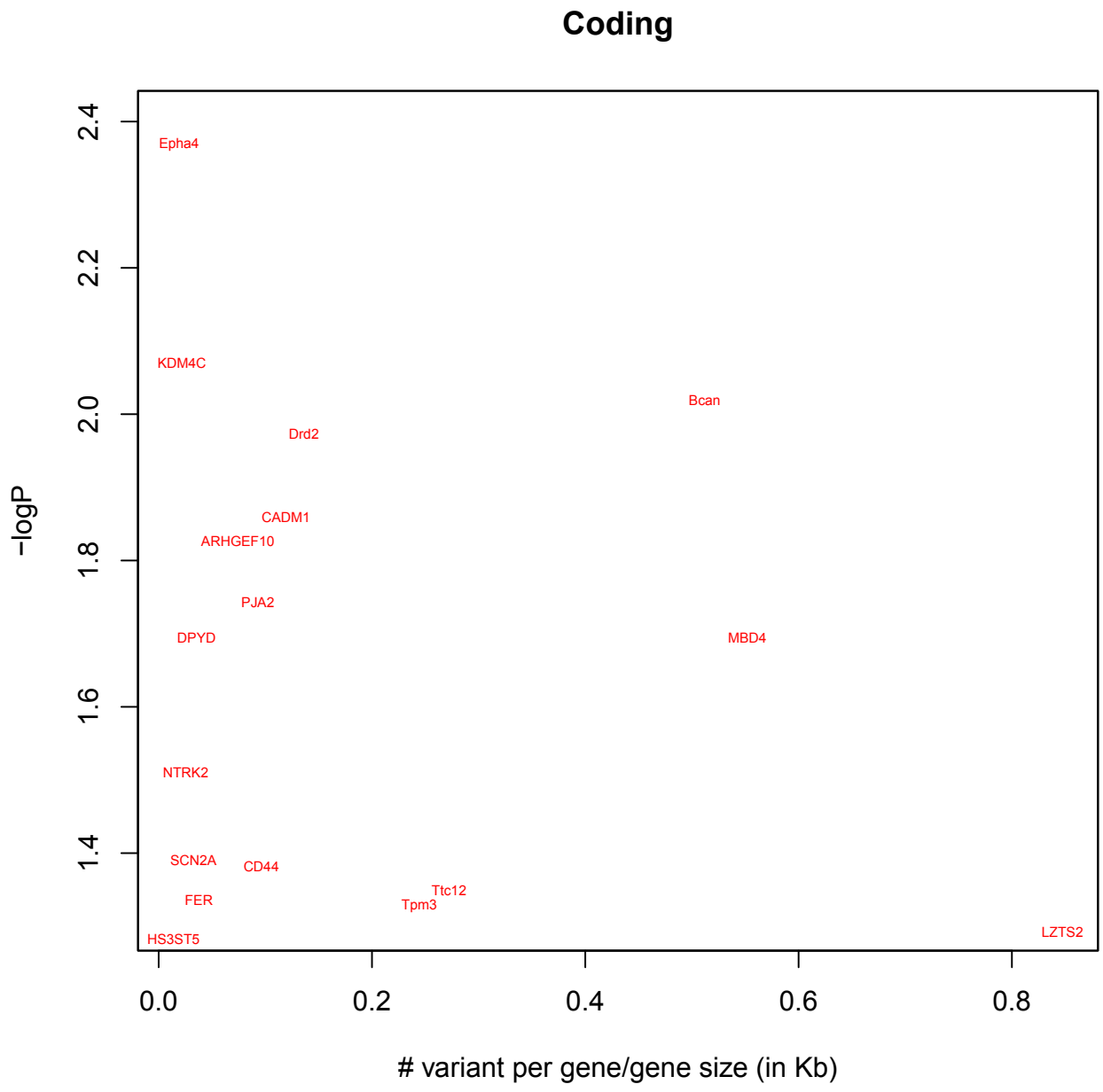


Figure 5.47: Associated genes from burden test (coding variants, adjusted for gene size) - See Appendix Figure 5.45.

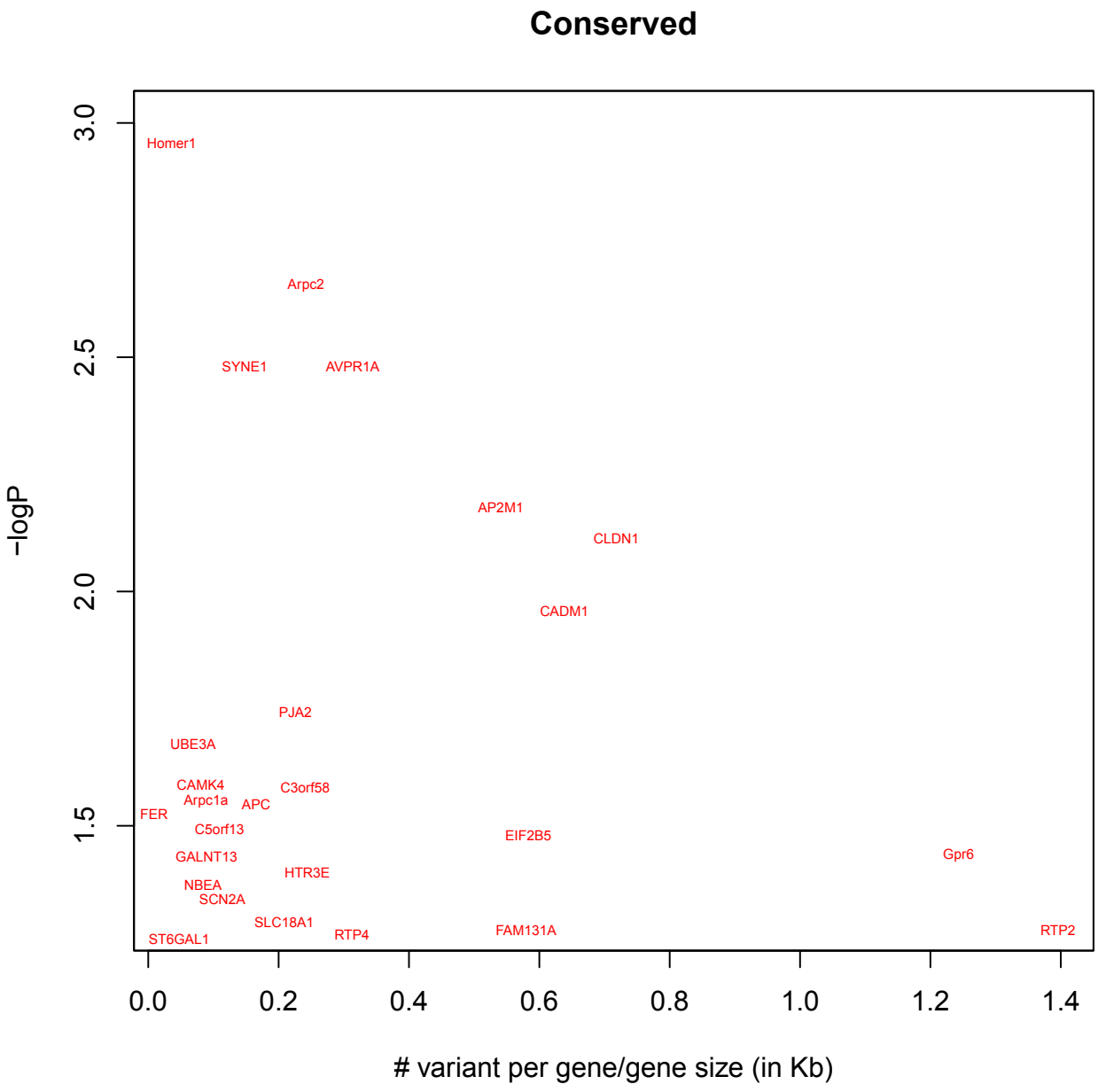


Figure 5.48: Associated genes from burden test (variants in evolutionarily conserved region, adjusted for gene size) - See Appendix Figure 5.45.

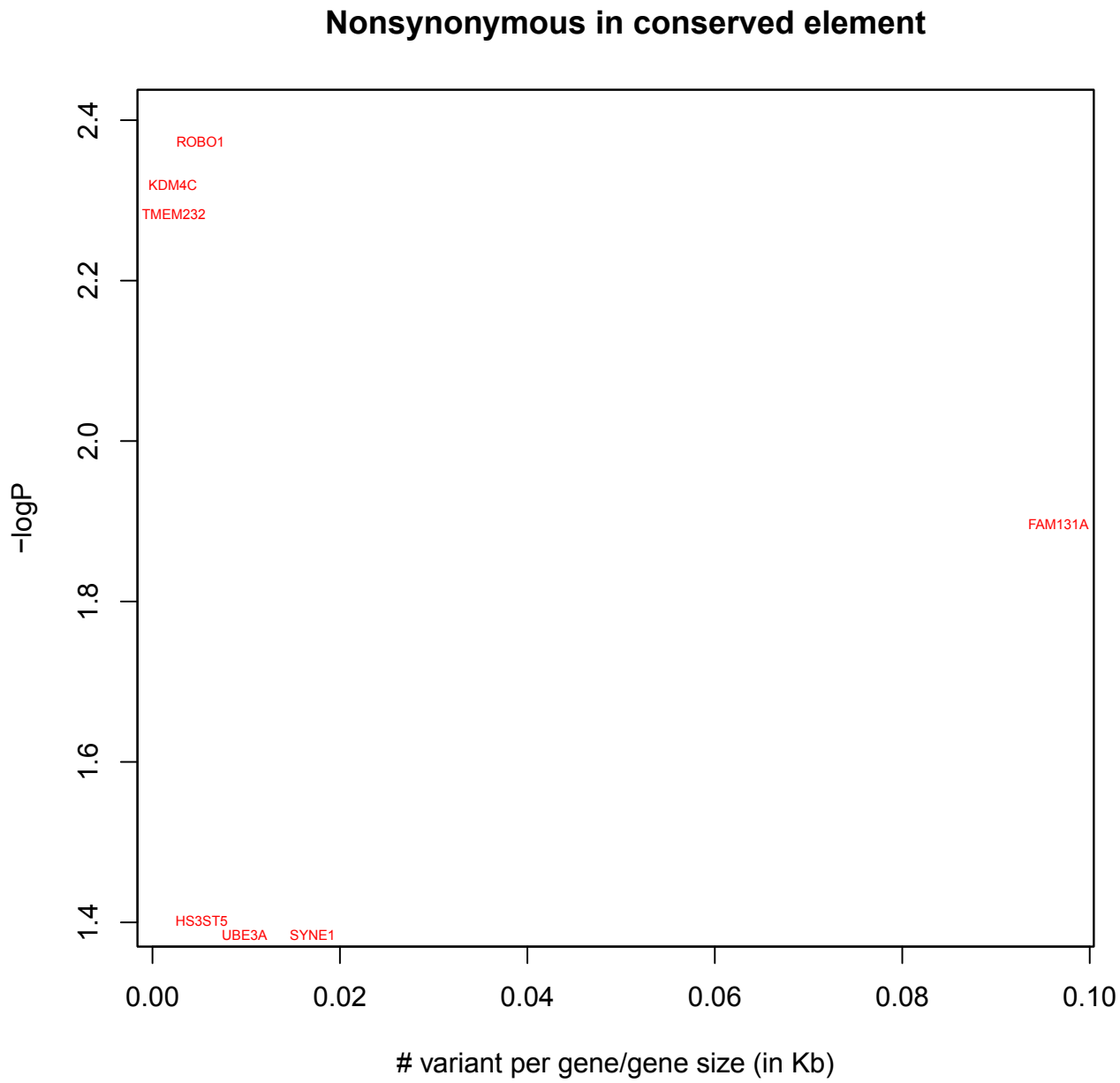


Figure 5.49: Associated genes from burden test (nonsynonymous variants in evolutionarily conserved region, adjusted for gene size) - See Appendix Figure 5.45.

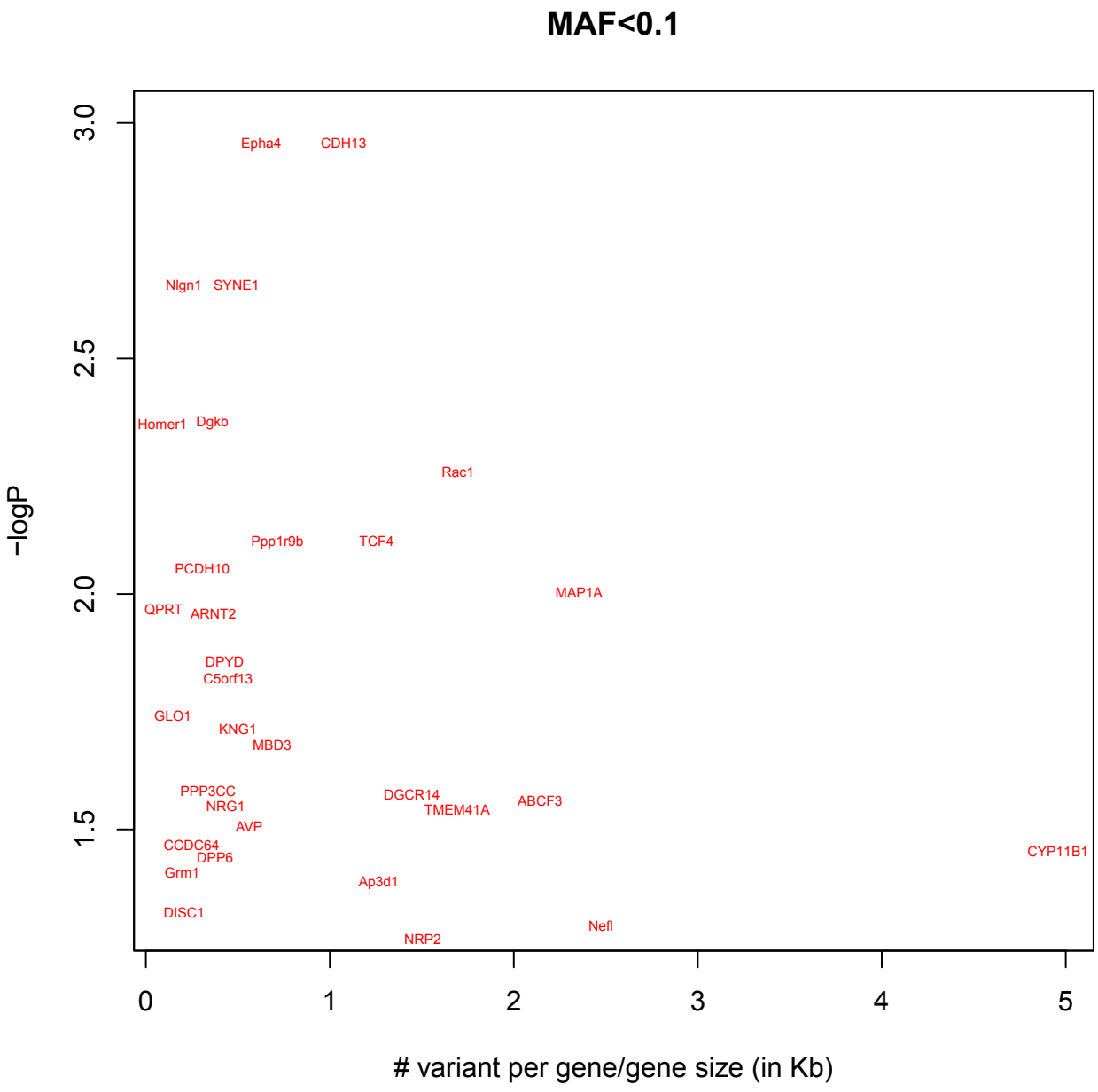


Figure 5.50: Associated genes from burden test (MAF<10%, adjusted for gene size) - See Appendix Figure 5.45.

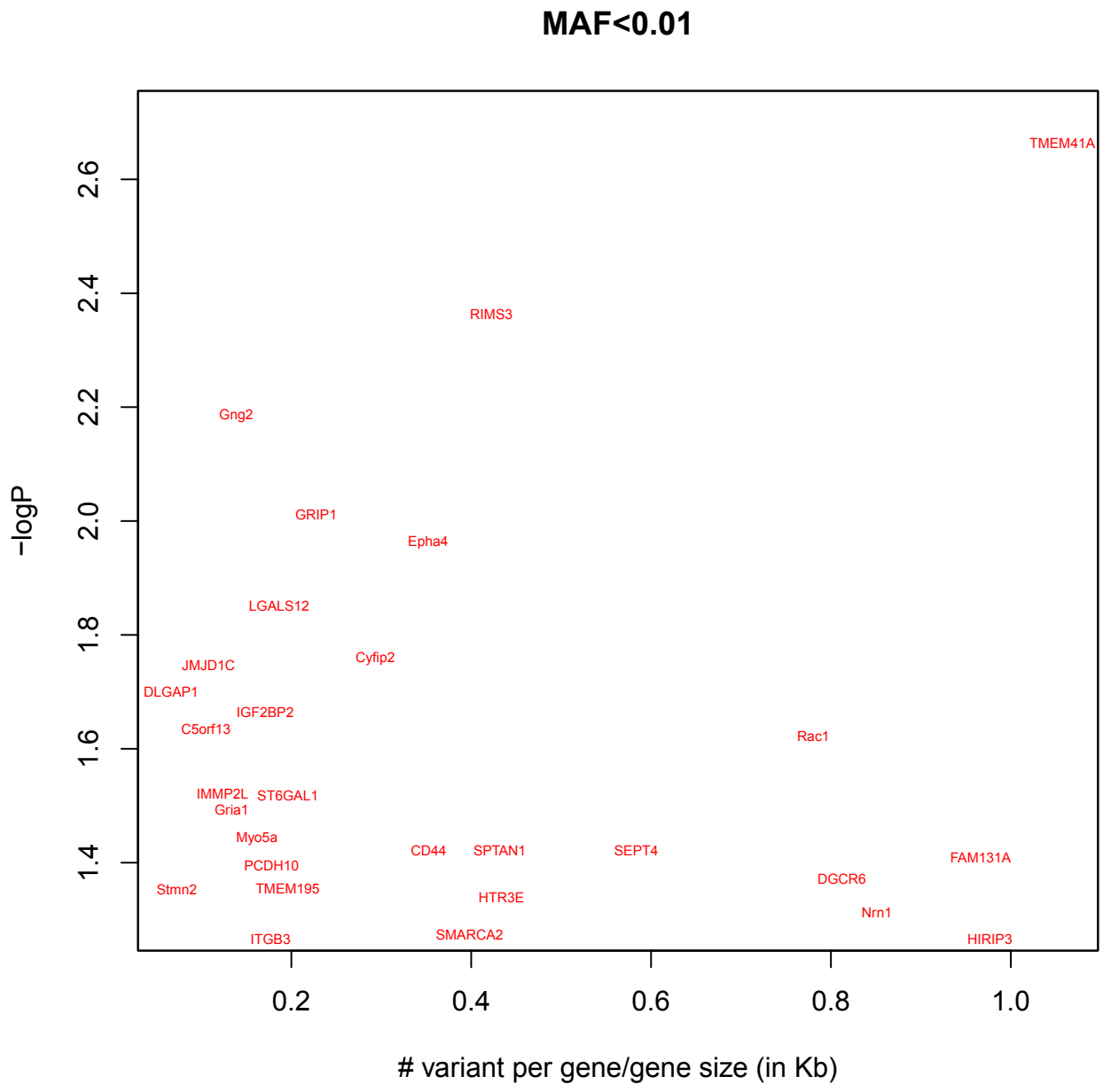


Figure 5.51: Associated genes from burden test (MAF<1%, adjusted for gene size) - See Appendix Figure 5.45.

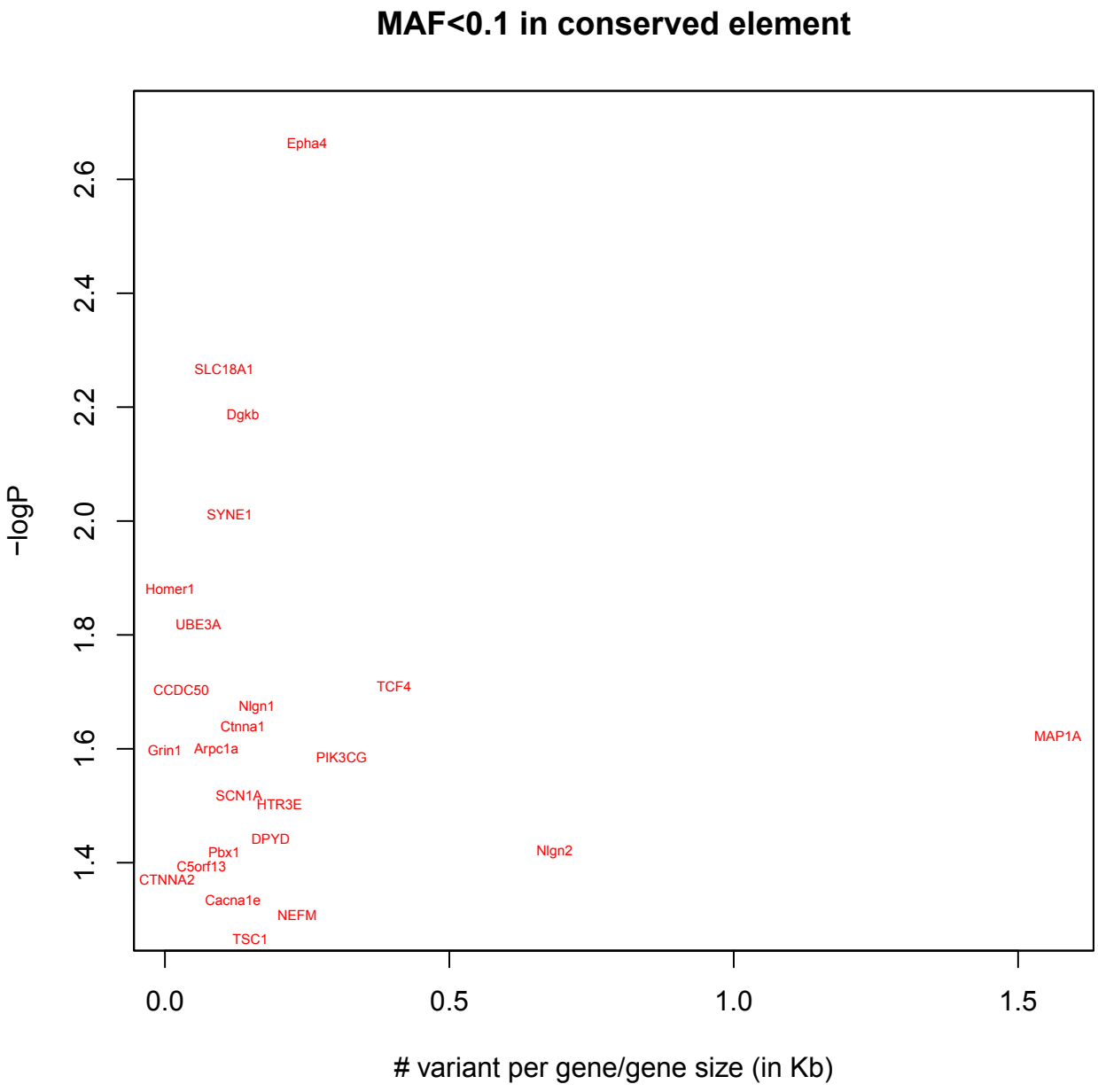


Figure 5.52: Associated genes from burden test (MAF<10% in conserved region, adjusted for gene size) - See Appendix Figure 5.45.

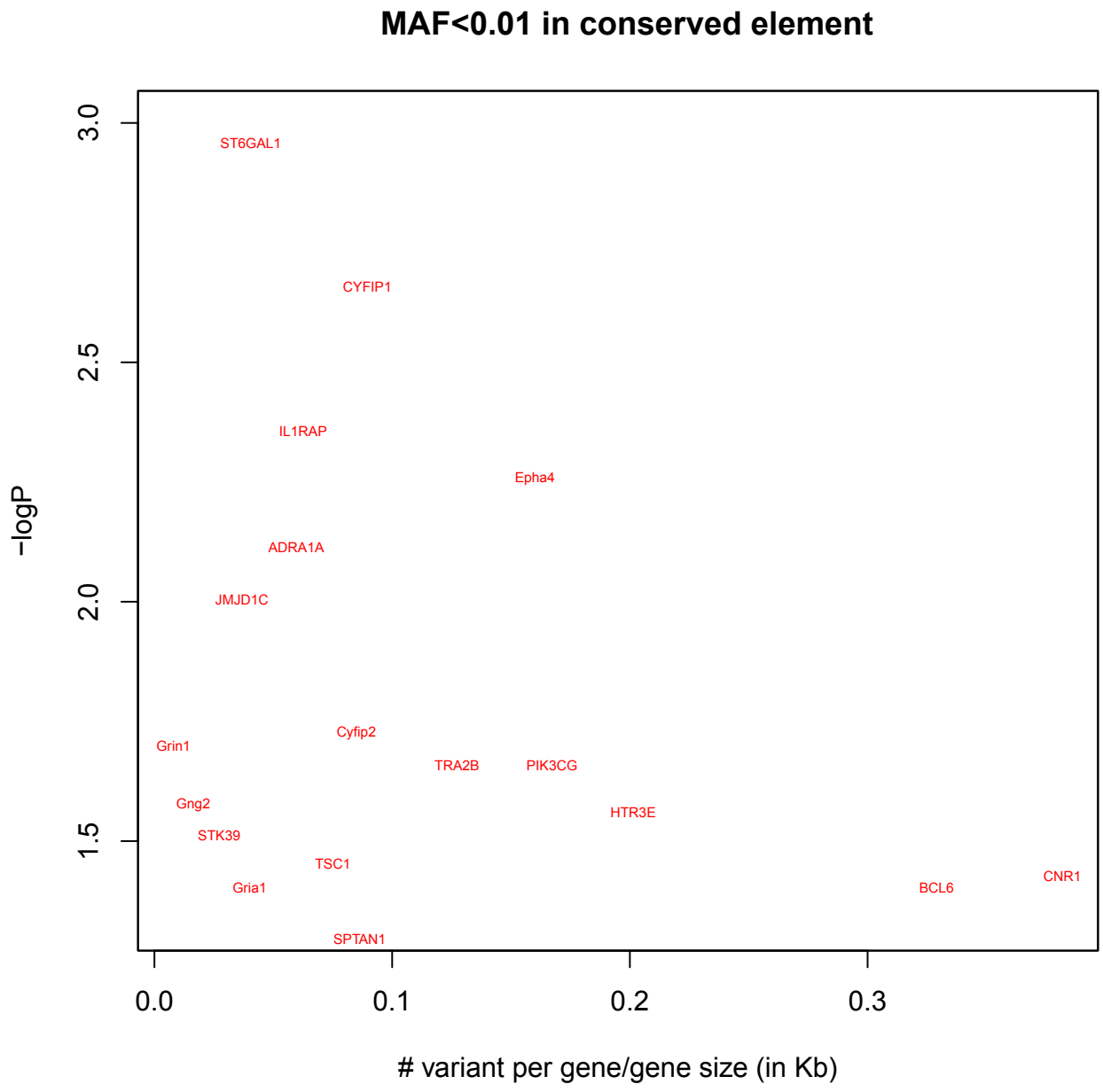


Figure 5.53: Associated genes from burden test (MAF$<$ 1% in conserved region, adjusted for gene size) - See Appendix Figure 5.45.

5.6.11 Pathway analyses

The columns in Appendix Tables 5.12 and 5.13 represent the following: *Category*, high-level GO term categories i.e. cellular component (CC), biological process (BP), and molecular function (MF); *Term*, GO ID; *Description*, description of GO term; *Count*, the number of genes that contribute to each enrichment; *%*, percentage of the number of genes annotated with each term to the total number of the input genes; *PValue*, hypergeometric probabilities; *List Total*, the number of genes in input list that are annotated with a given GO category (e.g. GOTERM_BP_FAT, not to be confused with specific GO terms); *Pop Hits*, the number of genes (from the whole genome) that are annotated with each GO term; *Pop Total*, the number of genes (from the whole genome) that are annotated with GO category; *Fold Enrichment*, fold-enrichment of associated genes over that expected by chance; *FDR*, p-values adjusted for false discovery rate.

Appendix Table 5.12 shows GO enrichments among 123 investigative genes, and Appendix Table 5.13 shows GO enrichments among 611 target genes.

Category	Term	Description	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	FDR
GOTERM_CC_FAT	GO:0043005	neuron projection	23	18.70	5.17E-14	107	342	12782	8.03	6.63E-11
GOTERM_CC_FAT	GO:0030424	axon	16	13.01	3.32E-12	107	159	12782	12.02	4.26E-09
GOTERM_CC_FAT	GO:0042995	cell projection	27	21.95	5.65E-11	107	697	12782	4.63	7.25E-08
GOTERM_CC_FAT	GO:0045202	synapse	20	16.26	8.22E-11	107	355	12782	6.73	1.06E-07
GOTERM_CC_FAT	GO:0030054	cell junction	23	18.70	1.94E-10	107	518	12782	5.30	2.49E-07
GOTERM_CC_FAT	GO:0044459	plasma membrane part	47	38.21	2.43E-10	107	2203	12782	2.55	3.12E-07
GOTERM_BP_FAT	GO:0019226	transmission of nerve impulse	19	15.45	3.99E-10	111	350	13528	6.62	6.75E-07
GOTERM_BP_FAT	GO:0007610	behavior	20	16.26	6.59E-09	111	469	13528	5.20	1.11E-05
GOTERM_CC_FAT	GO:0044456	synapse part	15	12.20	1.45E-08	107	246	12782	7.28	1.86E-05
GOTERM_BP_FAT	GO:0031175	neuron projection development	14	11.38	1.53E-07	111	256	13528	6.66	2.58E-04
GOTERM_CC_FAT	GO:0045211	postsynaptic membrane	11	8.94	1.66E-07	107	135	12782	9.73	2.14E-04
GOTERM_BP_FAT	GO:0030030	cell projection organization	16	13.01	2.79E-07	111	368	13528	5.30	4.72E-04
GOTERM_BP_FAT	GO:0031644	regulation of neurological system process	11	8.94	4.54E-07	111	153	13528	8.76	7.68E-04
GOTERM_BP_FAT	GO:0007409	axonogenesis	12	9.76	4.67E-07	111	193	13528	7.58	7.90E-04
GOTERM_BP_FAT	GO:0048666	neuron development	15	12.20	6.16E-07	111	339	13528	5.39	1.04E-03
GOTERM_BP_FAT	GO:0007268	synaptic transmission	14	11.38	8.67E-07	111	298	13528	5.73	1.47E-03
GOTERM_BP_FAT	GO:0021537	telencephalon development	8	6.50	9.16E-07	111	65	13528	15.00	1.55E-03
GOTERM_BP_FAT	GO:0048667	cell morphogenesis involved in neuron differentiation	12	9.76	1.03E-06	111	209	13528	7.00	1.74E-03
GOTERM_BP_FAT	GO:0048812	neuron projection morphogenesis	12	9.76	1.24E-06	111	213	13528	6.87	2.10E-03
GOTERM_BP_FAT	GO:0044057	regulation of system process	14	11.38	1.30E-06	111	309	13528	5.52	2.20E-03
GOTERM_BP_FAT	GO:0007267	cell-cell signaling	19	15.45	1.51E-06	111	600	13528	3.86	2.54E-03
GOTERM_CC_FAT	GO:0014069	postsynaptic density	8	6.50	1.90E-06	107	71	12782	13.46	2.44E-03
GOTERM_BP_FAT	GO:0032535	regulation of cellular component size	13	10.57	2.01E-06	111	271	13528	5.85	3.41E-03
GOTERM_BP_FAT	GO:0007626	locomotory behavior	13	10.57	2.26E-06	111	274	13528	5.78	3.82E-03
GOTERM_CC_FAT	GO:0005624	membrane fraction	22	17.89	2.44E-06	107	809	12782	3.25	3.14E-03
GOTERM_BP_FAT	GO:0030182	neuron differentiation	16	13.01	2.49E-06	111	438	13528	4.45	4.22E-03
GOTERM_CC_FAT	GO:0005886	plasma membrane	55	44.72	2.88E-06	107	3777	12782	1.74	3.70E-03
GOTERM_BP_FAT	GO:0051969	regulation of transmission of nerve impulse	10	8.13	3.02E-06	111	147	13528	8.29	5.11E-03
GOTERM_CC_FAT	GO:0005626	insoluble fraction	22	17.89	4.33E-06	107	839	12782	3.13	5.56E-03
GOTERM_BP_FAT	GO:0000904	cell morphogenesis involved in differentiation	12	9.76	4.64E-06	111	244	13528	5.99	7.85E-03
GOTERM_BP_FAT	GO:0048858	cell projection morphogenesis	12	9.76	4.83E-06	111	245	13528	5.97	8.17E-03
GOTERM_BP_FAT	GO:0032990	cell part morphogenesis	12	9.76	7.34E-06	111	256	13528	5.71	1.24E-02
GOTERM_CC_FAT	GO:0030425	dendrite	10	8.13	8.16E-06	107	163	12782	7.33	1.05E-02
GOTERM_BP_FAT	GO:0050767	regulation of neurogenesis	10	8.13	8.17E-06	111	166	13528	7.34	1.38E-02
GOTERM_BP_FAT	GO:0007155	cell adhesion	19	15.45	1.29E-05	111	700	13528	3.31	2.17E-02
GOTERM_BP_FAT	GO:0022610	biological adhesion	19	15.45	1.31E-05	111	701	13528	3.30	2.22E-02
GOTERM_BP_FAT	GO:0048878	chemical homeostasis	16	13.01	1.63E-05	111	512	13528	3.81	2.76E-02
GOTERM_BP_FAT	GO:0044087	regulation of cellular component biogenesis	9	7.32	2.01E-05	111	142	13528	7.72	3.40E-02
GOTERM_CC_FAT	GO:0000267	cell fraction	24	19.51	2.11E-05	107	1083	12782	2.65	2.71E-02
GOTERM_BP_FAT	GO:0042981	regulation of apoptosis	20	16.26	2.35E-05	111	804	13528	3.03	3.97E-02
GOTERM_BP_FAT	GO:0051960	regulation of nervous system development	10	8.13	2.62E-05	111	192	13528	6.35	4.43E-02
GOTERM_BP_FAT	GO:0043067	regulation of programmed cell death	20	16.26	2.69E-05	111	812	13528	3.00	4.55E-02
GOTERM_BP_FAT	GO:0010941	regulation of cell death	20	16.26	2.83E-05	111	815	13528	2.99	4.78E-02

Table 5.12: 43 GO FAT terms that are significantly enriched (FDR$<$ 5%) among 123 investigative genes from burden test - See Appendix text 5.6.11.

Category	Term	Description	Count	%	PValue	List Total	Pop Hits	Pop Total	Fold Enrichment	FDR
GOTERM_BP_FAT	GO:0019226	transmission of nerve impulse	82	13.60	2.80E-42	503	350	13528	6.30	5.05E-39
GOTERM_CC_FAT	GO:0045202	synapse	80	13.27	1.24E-39	481	355	12782	5.99	1.73E-36
GOTERM_CC_FAT	GO:0043005	neuron projection	78	12.94	5.33E-39	481	342	12782	6.06	7.43E-36
GOTERM_BP_FAT	GO:0007268	synaptic transmission	73	12.11	6.86E-39	503	298	13528	6.59	1.24E-35
GOTERM_CC_FAT	GO:0044456	synapse part	62	10.28	1.91E-33	481	246	12782	6.70	2.67E-30
GOTERM_CC_FAT	GO:0042995	cell projection	100	16.58	3.53E-32	481	697	12782	3.81	4.92E-29
GOTERM_CC_FAT	GO:0045211	postsynaptic membrane	47	7.79	4.25E-32	481	135	12782	9.25	5.92E-29
GOTERM_BP_FAT	GO:0007267	cell-cell signaling	92	15.26	5.36E-32	503	600	13528	4.12	9.66E-29
GOTERM_CC_FAT	GO:0044459	plasma membrane part	186	30.85	5.64E-30	481	2203	12782	2.24	7.86E-27
GOTERM_CC_FAT	GO:0030425	dendrite	48	7.96	4.38E-29	481	163	12782	7.83	6.11E-26
GOTERM_BP_FAT	GO:0044057	regulation of system process	63	10.45	1.23E-28	503	309	13528	5.48	2.21E-25
GOTERM_BP_FAT	GO:0007610	behavior	77	12.77	1.37E-28	503	469	13528	4.42	2.47E-25
GOTERM_CC_FAT	GO:0030054	cell junction	80	13.27	1.20E-27	481	518	12782	4.10	1.67E-24
GOTERM_BP_FAT	GO:0031644	regulation of neurological system process	44	7.30	2.35E-26	503	153	13528	7.73	4.24E-23
GOTERM_BP_FAT	GO:0051969	regulation of transmission of nerve impulse	42	6.97	4.92E-25	503	147	13528	7.68	8.86E-22
GOTERM_CC_FAT	GO:0030424	axon	42	6.97	1.97E-23	481	159	12782	7.02	2.74E-20
GOTERM_CC_FAT	GO:0014069	postsynaptic density	28	4.64	9.54E-21	481	71	12782	10.48	1.33E-17
GOTERM_BP_FAT	GO:0030182	neuron differentiation	63	10.45	2.90E-20	503	438	13528	3.87	5.22E-17
GOTERM_BP_FAT	GO:0031175	neuron projection development	48	7.96	3.25E-20	503	256	13528	5.04	5.85E-17
GOTERM_BP_FAT	GO:0048666	neuron development	55	9.12	4.09E-20	503	339	13528	4.36	7.37E-17
GOTERM_BP_FAT	GO:0030030	cell projection organization	56	9.29	3.67E-19	503	368	13528	4.09	6.62E-16
GOTERM_CC_FAT	GO:0005886	plasma membrane	233	38.64	5.03E-19	481	3777	12782	1.64	7.01E-16
GOTERM_BP_FAT	GO:0048812	neuron projection morphogenesis	41	6.80	9.99E-18	503	213	13528	5.18	1.80E-14
GOTERM_BP_FAT	GO:0007409	axonogenesis	39	6.47	1.28E-17	503	193	13528	5.43	2.30E-14
GOTERM_BP_FAT	GO:0048858	cell projection morphogenesis	43	7.13	4.71E-17	503	245	13528	4.72	8.48E-14
GOTERM_BP_FAT	GO:0048667	cell morphogenesis involved in neuron differentiation	39	6.47	2.05E-16	503	209	13528	5.02	4.00E-13
GOTERM_BP_FAT	GO:0000904	cell morphogenesis involved in differentiation	42	6.97	2.32E-16	503	244	13528	4.63	4.00E-13
GOTERM_BP_FAT	GO:0032990	cell part morphogenesis	43	7.13	2.39E-16	503	256	13528	4.52	4.00E-13
GOTERM_BP_FAT	GO:0044087	regulation of cellular component biogenesis	27	4.48	1.21E-11	503	142	13528	5.11	2.18E-08
GOTERM_BP_FAT	GO:0007626	locomotory behavior	37	6.14	3.58E-11	503	274	13528	3.63	6.45E-08
GOTERM_BP_FAT	GO:0051960	regulation of nervous system development	30	4.98	1.06E-10	503	192	13528	4.20	1.92E-07
GOTERM_CC_FAT	GO:0005624	membrane fraction	68	11.28	5.86E-10	481	809	12782	2.23	8.17E-07
GOTERM_CC_FAT	GO:0005626	insoluble fraction	69	11.44	1.08E-09	481	839	12782	2.19	1.50E-06
GOTERM_BP_FAT	GO:0021537	telencephalon development	17	2.82	1.23E-09	503	65	13528	7.03	2.22E-06
GOTERM_BP_FAT	GO:0050767	regulation of neurogenesis	26	4.31	2.24E-09	503	166	13528	4.21	4.04E-06
GOTERM_CC_FAT	GO:0000267	cell fraction	76	12.60	1.24E-07	481	1083	12782	1.86	1.72E-04
GOTERM_BP_FAT	GO:0032535	regulation of cellular component size	29	4.81	9.73E-07	503	271	13528	2.88	1.75E-03
GOTERM_BP_FAT	GO:0007155	cell adhesion	53	8.79	1.32E-06	503	700	13528	2.04	2.38E-03
GOTERM_BP_FAT	GO:0022610	biological adhesion	53	8.79	1.34E-06	503	701	13528	2.03	2.41E-03
GOTERM_BP_FAT	GO:0048878	chemical homeostasis	42	6.97	2.93E-06	503	512	13528	2.21	5.28E-03

Table 5.13: 40 GO FAT terms that are significantly enriched (FDR$<5\%$) both in 123 investigative genes and 611 target genes - See Appendix text 5.6.11.

Figure 5.54: REVIGO-summary of the significantly associated (FDR <5%) Biological Process GO FAT terms among investigative genes - The summary includes biological themes related to cell projection organisation, chemical homeostasis, behavior, cell-cell signaling, and adhesion.



Figure 5.55: REVIGO-summary of the significantly associated ($FDR < 5\%$) Cellular Component GO FAT terms among investigative genes - The summary includes biological themes related to plasma membrane part, neuron projection, synapse, and cell junction/projection.

5.6.12 Power calculation

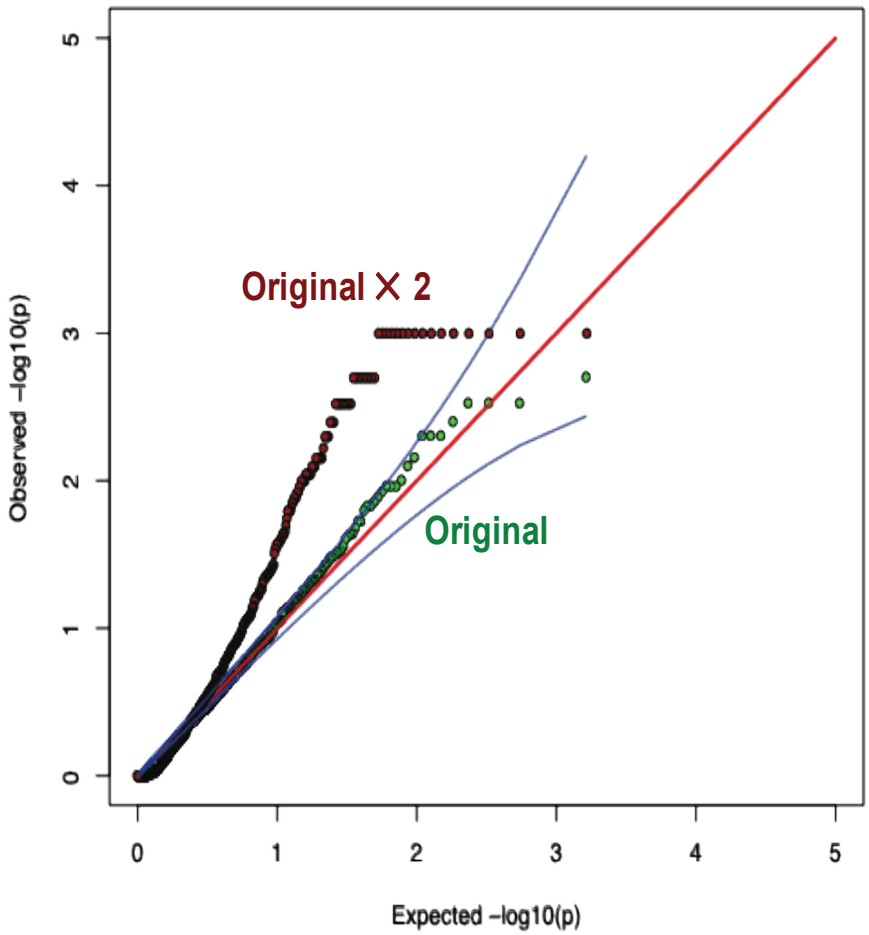


Figure 5.56: Q-Q plot for original dataset and doubled dataset - The green dots represent the original Q-Q plot while the red dots represent the Q-Q plot of an artificial dataset where the original dataset is doubled.

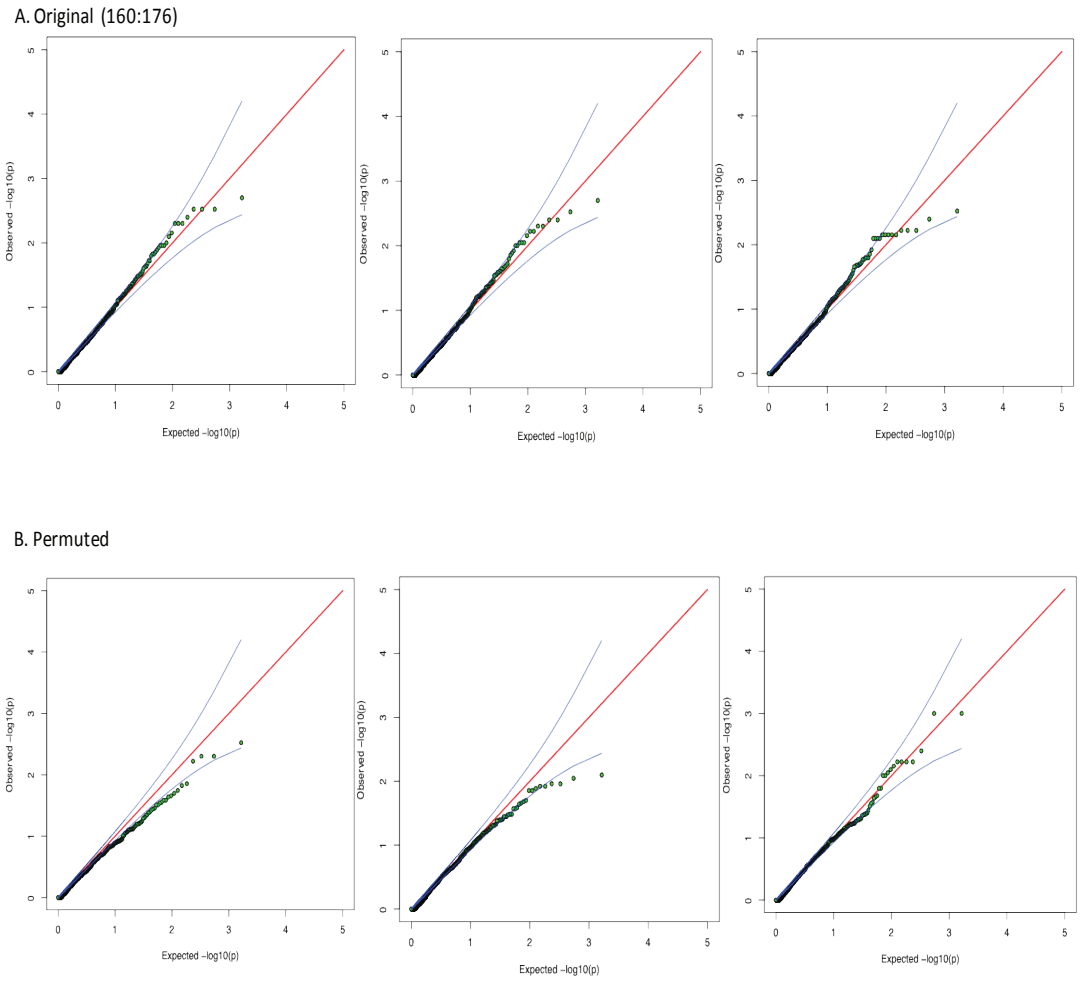


Figure 5.57: Q-Q plots for original dataset and permuted datasets where the sample size matches to the original data - For the panel A, The gene-based association test (burden test) with 1,000 permutations are repeated for each Q-Q plot. For the panel B, the case/control pools are randomly permuted for each Q-Q plot.

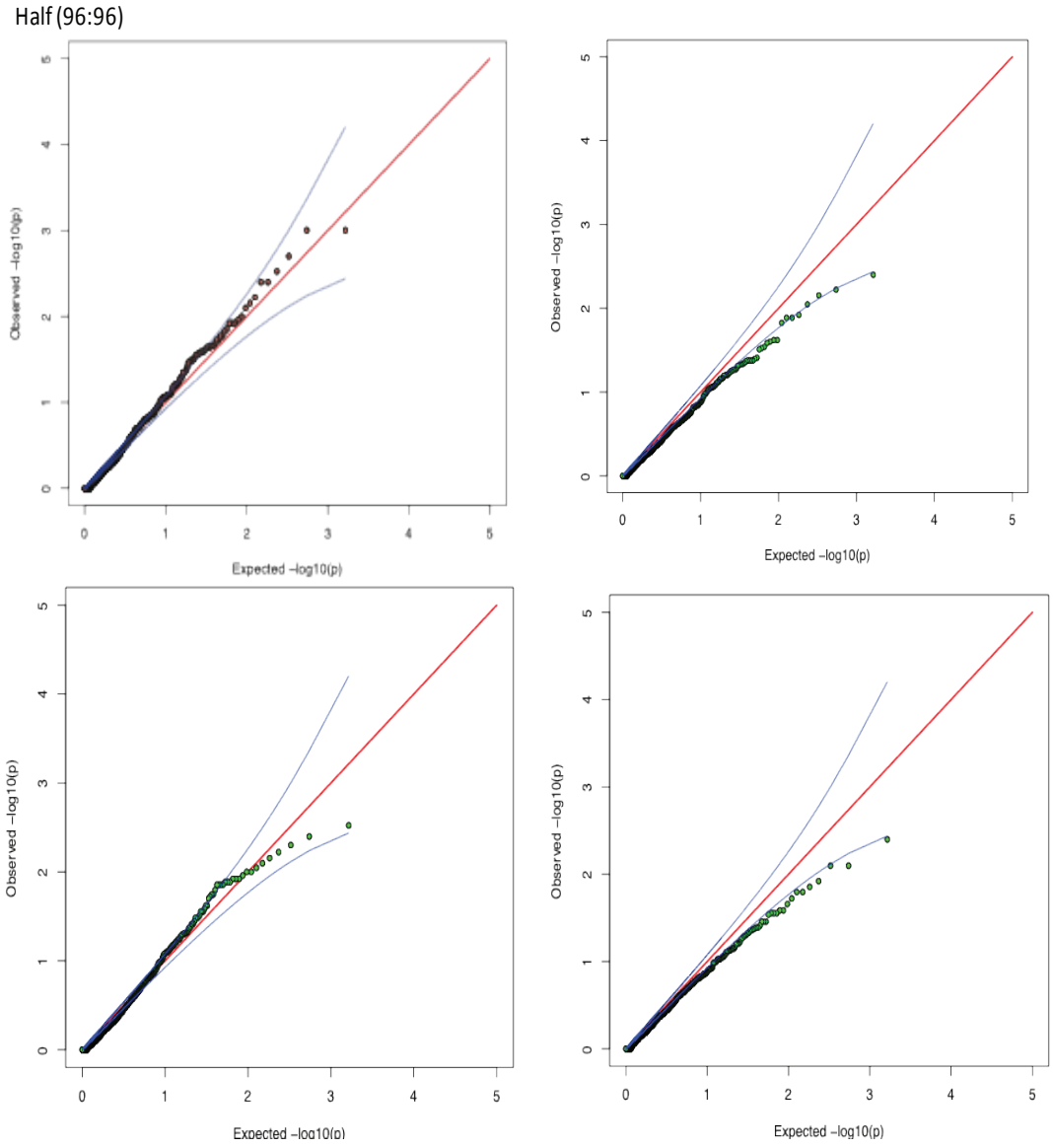


Figure 5.58: Q-Q plots for sub-sampled datasets consisting of 96 individuals for each case and control from the original dataset - Each Q-Q plot represents different sub-samples of the original dataset.

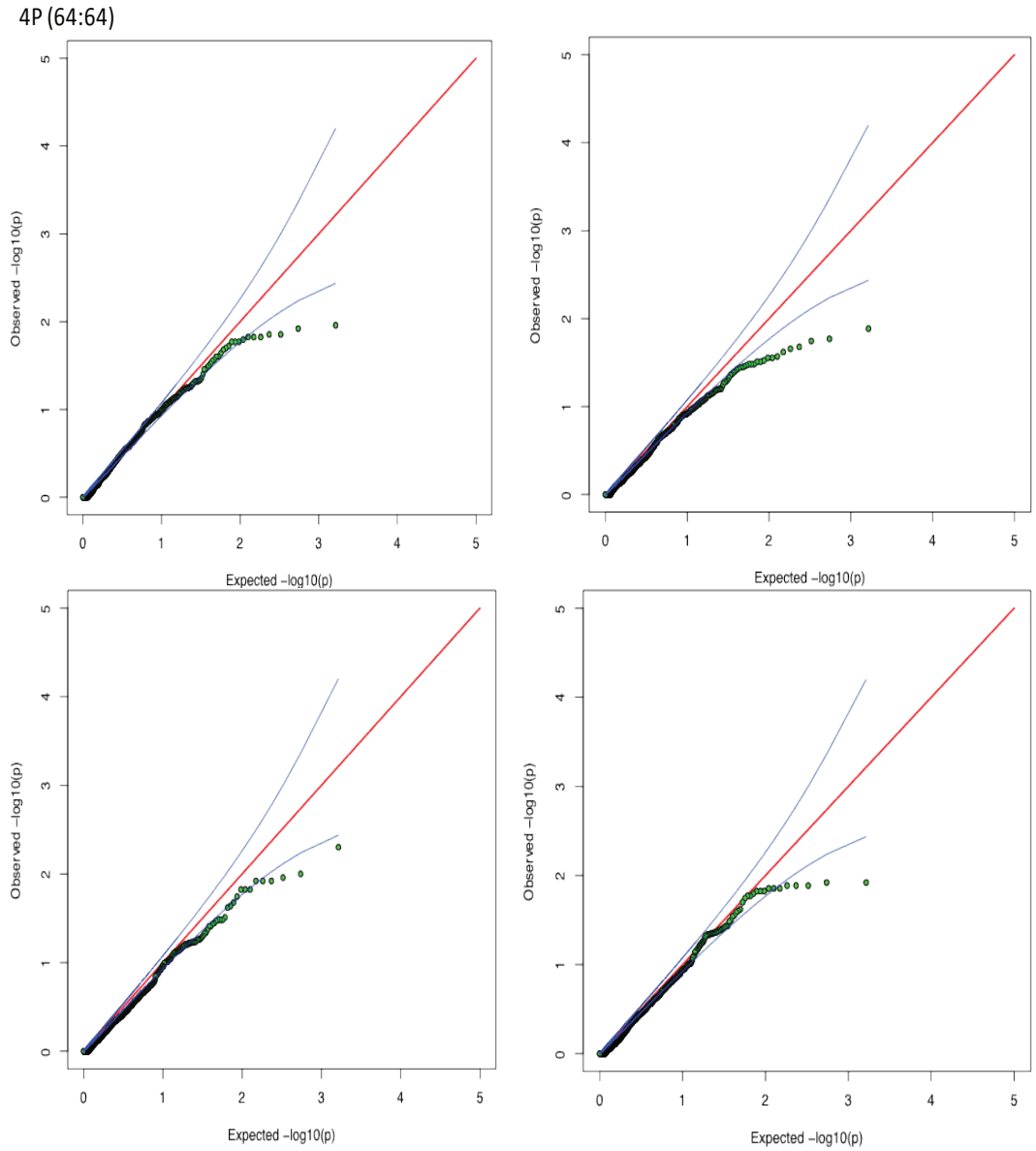


Figure 5.59: Q-Q plots for sub-sampled datasets consisting of 64 individuals for each case and control from the original dataset - Sub-sampling was repeated for each Q-Q plot.

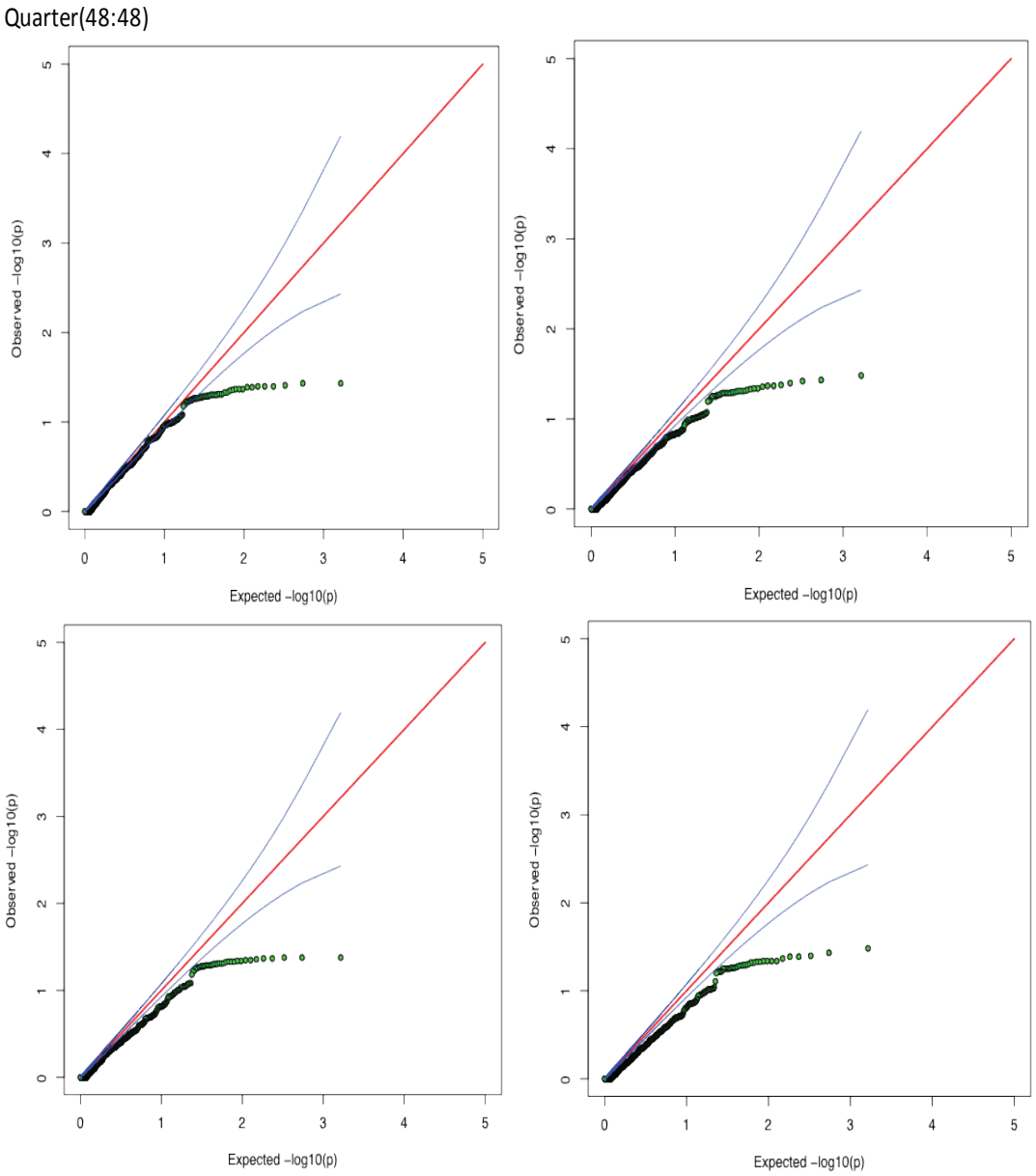


Figure 5.60: Q-Q plots for sub-sampled datasets consisting of 48 individuals for each case and control from the original dataset - Sub-sampling was repeated for each Q-Q plot.

5.6.13 Tracks for regulatory marks

Tracks for regulatory marks used herein were obtained from the ENCODE project (5) available via the UCSC genome browser (128). The *Transcription* track shows the transcription levels assayed by sequencing of poly-adenylated RNA from a variety of cell types (Appendix Table 5.14). The overlayed *H3K4Me1* and *H3K27Ac* tracks show the modification of histone H3 proteins i.e. suggestive of enhancer and other regulatory activities. These histone modifications, particularly H3K4Me1, are quite broad. The actual enhancers are typically just a small portion of the area marked by these histone modifications. The overlayed *H3K4Me3* track shows the histone marks associated with promoters. The *DNase1 Clusters* track shows the regions where the chromatin is hypersensitive to cutting by the DNase1 enzyme, assayed in various cell types (Appendix Table 5.14). This track is used based on the assumption that the regulatory regions, in general, tend to be DNase sensitive. The *Txn Factor ChIP* track shows the DNA regions where the transcription (Txn) factors (i.e. the proteins responsible for modulating gene transcription) bind, assayed by chromatin immunoprecipitation (ChIP) with antibodies specific to the transcription factors followed by sequencing of the precipitated DNA (ChIP-seq).

These tracks are complementary to each other and together can shed much light on regulatory regions (128). The histone marks are informative at a high level, but they have a resolution of just ∼200 bases and do not provide much in the way of functional detail. The DNase hypersensitive assay is higher in resolution at the DNA level and can be done on a large number of cell types. At the functional level, DNase hypersensitivity suggests that a region is very likely to be regulatory in nature, but provides little information beyond that. The transcription factor ChIP assay has a high resolution at the DNA level, and, due to the very specific nature of the transcription factors, is often informative with respect to functional details. However, since each transcription factor must be assayed separately, the information is only available for a limited number of transcription factors on a limited number of cell lines. Though each assay has its own strengths and weaknesses, the fact that all of these assays are relatively independent of each other gives greater confidence when multiple tracks are suggestive of regulatory function for a region simultaneously (text adapted from (128)).

Colour code	Cell	Tissue	Lineage	Description	Karyotype	Sex
	GM12878	blood	mesoderm	B-lymphocyte, lymphoblastoid, International HapMap Project - CEPH/Utah - European Caucasion, Epstein-Barr Virus	normal	F
	H1-hESC	embryonic	inner cell mass	embryonic stem cells	normal	M
	HSMM	muscle	mesoderm	skeletal muscle myoblasts	normal	U
	HUVEC	blood vessel	mesoderm	umbilical vein endothelial cells	normal	U
	K562	blood	mesoderm	leukemia, "The continuous cell line K-562 was established by Lozzio and Lozzio from the pleural effusion of a 53-year-old female with chronic myelogenous leukemia in terminal blast crises." - ATCC	cancer	F
	NHEK	skin	ectoderm	epidermal keratinocytes	normal	U
	NHLF	lung	endoderm	lung fibroblasts	normal	U

Table 5.14: Cell line information assayed in the ENCODE project - The table shows the information about the 7 cell lines used for the "Integrated Regulation" from the ENCODE Tracks in the UCSC genome browser at the time of this study being conducted.

5.6.14 Transcription factor binding DNA motif



Figure 5.61: The CCAAT enhancer binding protein gamma (CEBPG) motif - CEBPG is a transcription repressor that plays a role in DNA-dependent regulation of transcription. The position Chr3:190025964 is indicated by a red arrow. The smaller the letters are, the less constrained bases for the motif.

5.6.15 Tracks for evolutionary conservation

The *Vertebrate Multiz Alignment and Conservation* track shows the multiple alignments of 46 vertebrate species and their evolutionary conservation (397)(398). The multiple alignments were generated using the Multiz and other tools implemented in the UCSC/Penn State Bioinformatics comparative genomics alignment pipeline. PhastCons is a hidden Markov model-based method that estimates the probability that each nucleotide belongs to a conserved element based on the multiple alignment. It considers not just each individual alignment column, but also its flanking columns. By contrast, PhyloP separately measures the conservation at individual columns, ignoring the effects of their neighbours. As a consequence, the PhyloP plots have a less smooth appearance than the PhastCons plots, with more "texture" at individual sites.

The two methods have different strengths and weaknesses. PhastCons is sensitive to "runs" of conserved sites, and is therefore effective at picking out conserved elements. PhyloP, on the other hand, is more appropriate for evaluating signatures of selection at particular nucleotides or classes of nucleotides, such as third codon positions or the first positions of miRNA target sites. Another important difference is that PhyloP can measure acceleration (faster evolution than expected under neutral drift) as well as conservation (slower than expected evolution). In PhyloP plots, sites predicted to be conserved are assigned positive scores (i.e. conservation, shown in blue), while sites predicted to be fast-evolving are assigned negative scores (i.e. acceleration, shown in red). The absolute values of the scores represent -$log_{10}p-values$ under a null hypothesis of neutral evolution. The PhastCons scores, by contrast, represent the probabilities of negative selection and range between 0 and 1.

Both PhastCons and PhyloP treat alignment gaps and unaligned nucleotides as missing data, and both were run with the same parameters for each species set (i.e. vertebrates, placental mammals, and primates; Appendix Figure 5.62). Thus, in regions in which only primates appear in the alignment, all three sets of scores will be the same, but in the regions in which additional species are available, the mammalian and/or vertebrate scores may differ from the primate scores. These alternative plots help to identify sequences that are under different evolutionary pressures in, say, primates and non-primates, or mammals and non-mammals (text adapted from (127)).

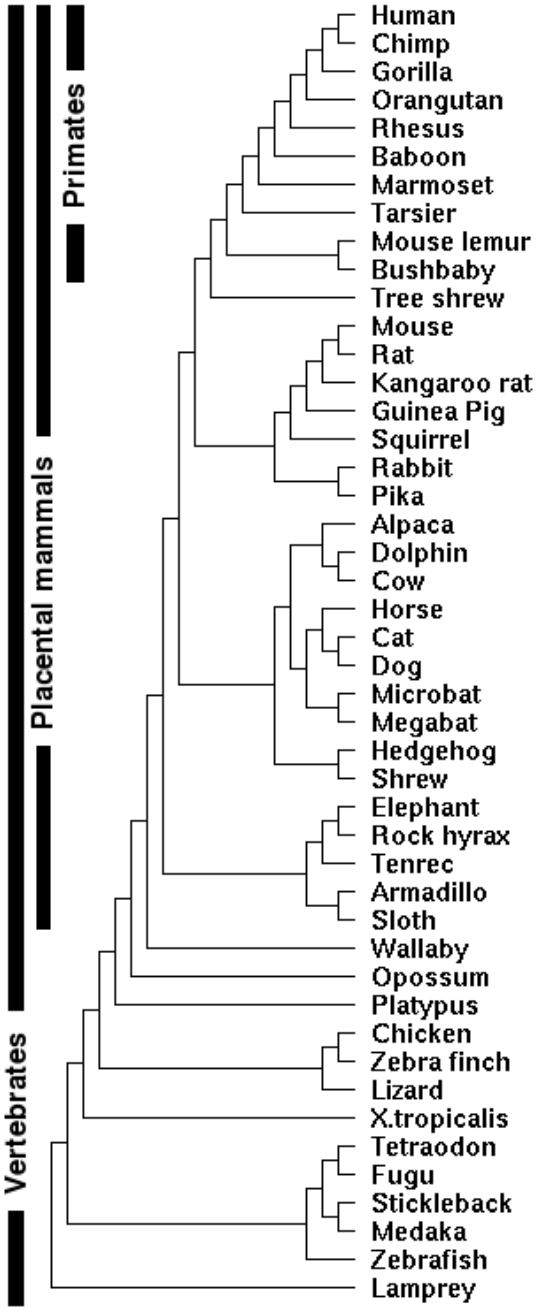


Figure 5.62: Phylogeny of the 46 species in the Vertebrate Multiz Alignment and Conservation track - The Vertebrate Multiz Alignment and Conservation track was used in the main text to evaluate whether OCD candidate variants are evolutionarily conserved, and hence are likely to be functional.

5. TARGETED RESEQUENCING OF HUMAN GENOMIC REGIONS FOR OBSESSIVE-COMPULSIVE DISORDER

6

Resequencing of genomic regions for compulsive disorder in dogs

6.1 Abstract

Obsessive-compulsive disorder (OCD), a critical mental condition exhibited as time-consuming repetition of behaviours, affects 1-3% of the human population. Its canine counterpart, canine compulsive disorder (CCD) strikingly resembles OCD, with normal activities such as grooming, predation and suckling developed to tail chasing, and flank/blanket sucking. As in humans, approximately half of affecteds are responsive to selective serotonin reuptake inhibitors (SSRIs). While the complexity of OCD makes human genetic studies difficult, the canine model offers advantages due to the limited genetic complexity within breeds. Indeed, a GWAS and selective sweep test for CCD identified several candidate regions in which many genes are potentially involved with the CSTC brain circuit that may be responsible for compulsive behaviour. Here, I analysed the novel data generated by resequencing 26 candidate genomic regions (∼5Mb) for CCD in 12 affected dogs and 12 matched controls that were selected from multiple breeds. To this end, I generated high-quality sequencing alignments of over half a billion reads, and identified >25,000 SNPs, >7,800 micro-indels and >170 deleted CNVs. Of these, ∼300 variants were found to reside within known protein coding sequences, intronic splice sites, or untranslated regions of the genome, and over 400 variants overlapped evolutionarily conserved elements. The detected variants were prioritised, considering their phenotypic concordance and functional impact when disrupted. Finally, case-control comparison of multiple variants per gene identified candidate genes for CCD. These candidate genes' functions and contributing biological pathways include synaptic formation and cell-cell adhesion that may be comparable to those identified from human OCD study (Chapter 5).

6.2 Introduction

6.2.1 The domestic dog: its unique population history

The unique population history of the domestic dog (*Canis familiaris*) places the modern dog in a special position in genomics. Dog is the oldest domesticated species (399) that can be traced back at least 15,000 years to possibly as far back as 100,000 years ago, to its initial domestication from the grey wolf in East Asia (400)(401)(402)(403) (Figure 6.1). Ever since domestication, humans and dogs have been sharing common living space and nutritional resources. Over the last few centuries, humans have selectively bred dogs for their behaviours (e.g. herding flocks, guarding homes, hunting by sight and scent, retrieving wild animals, pulling sledges and obedience), in order to aid humans' daily activities (Figure 6.1). Throughout this artificial selection, dog breeds have become rich in behaviours that mimic human behaviours and support our needs (404). More recently, dogs have also been bred for their morphology such as size, coat colour and skull shape, which in turn generated dogs with greatly diverse physical properties across breeds (405)(406). As a result, each of the modern dog breeds is now defined by its specific behavioural and physical features.

6.2.2 Dog as a model organism for human disease

Rigorous breeding schemes and population bottlenecks have led to unintended health problems of purebred dogs with a high prevalence of specific diseases (408), reflecting the enrichment of risk alleles due to random fixation during population bottlenecks, hitch-hiking of mutations near desired traits, or pleiotropic effects of selected variants (409)(407). Notably, most of these diseases are also shared with humans, and the clinical manifestations in two species are often strikingly similar (409). Just as in humans, canine diseases occur spontaneously over the course of their lives, and include common human diseases such as cancers (410)(411), diabetes (412), heart disease, epilepsy (413), and even neuropsychiatric disorders such as OCD (414). The clinical manifestations of diseases are also similar in dogs and humans. For instance, both in dogs and humans, osteosarcoma appears first as the primary tumour in the metaphyseal region of the long bones, then metastasises to the lungs, leading to death usually due to diffuse pulmonary metastasis (415).

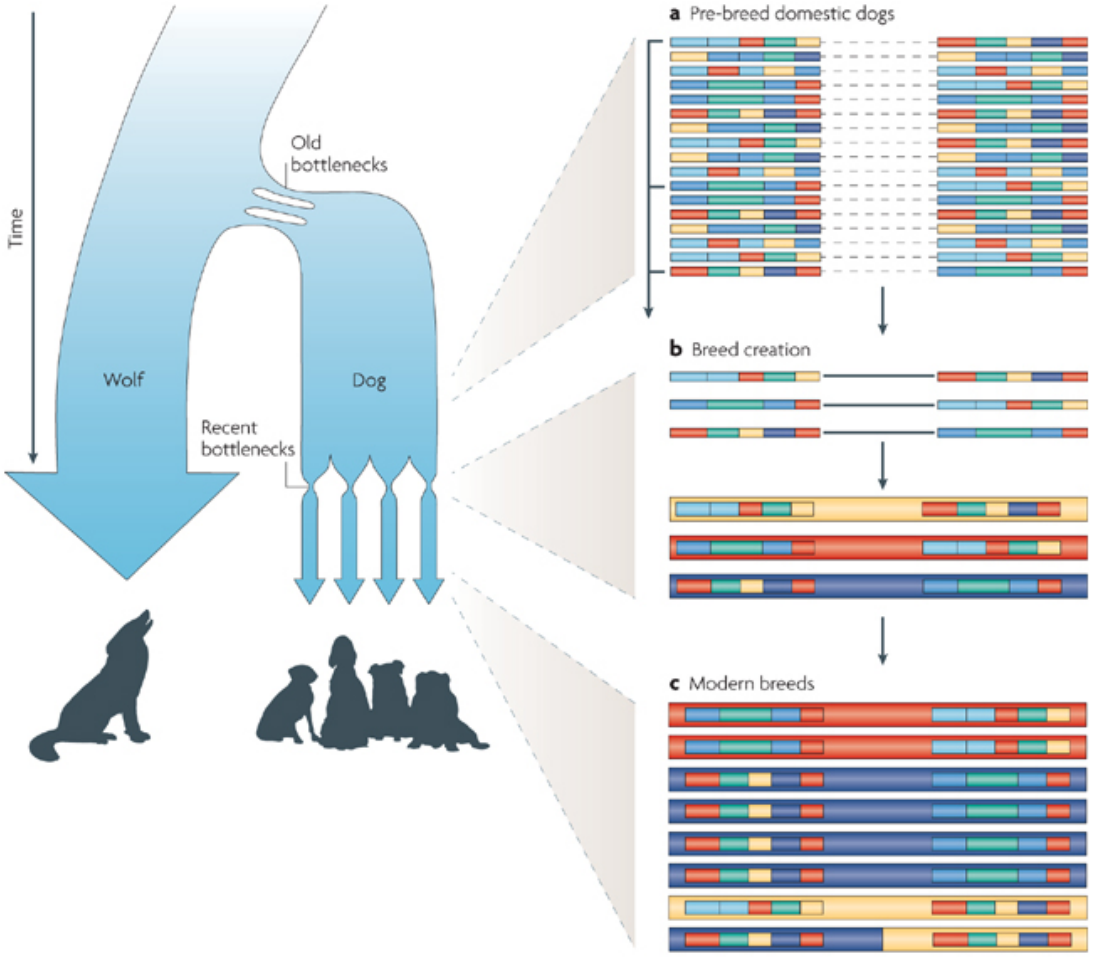


Figure 6.1: Population history and haplotype structure of *Canis familiaris* - The haplotype structure of modern dog breeds was largely shaped by two populations bottlenecks, namely, ancient domestication (at least ∼15,000 years ago) and recent inbreeding programme (few hundred years ago). a) Since the domestication event, the dog population grew large in its size and existed for long time. This led to the short-range linkage disequilibriums (LDs) in domestic dogs, before the creation of modern breeds. b) During the stringent breeding programme, a subset of chromosomes was selected from the pool of domestic dogs. The long-range patterns that were carried on these chromosomes became common within the breed, thereby creating long-range LDs. c) Since the breed creation, these long-range patterns have not yet been substantially broken down by recombination as the breeds are relatively young (∼200 years). These long-range haplotypes within breeds consist of the short ancestral haplotype blocks from the pre-breed domestic dog population, and these are revealed when the chromosomes are examined across many breeds. Adapted from (407).

Furthermore, the human genome is more similar to the dog than to the mouse genome, the foremost model organism, with an average of ∼0.35 nucleotide substitutions, compared to ∼0.51 for the mouse (407). The genome is relatively compact (2.4Gb), with fewer repeat insertions and segmental duplications than many other mammals. Lastly, the number of canine genes is estimated to be almost the same as in humans, most of which are 1:1 orthologues (404).

Family history is one of the strongest risk factors for most human diseases (416). Similarly, the exceptionally high prevalence of particular diseases in certain dog breeds attributes canine diseases to genetic components (407). Such greatly increased risks in breeds with relatively little genetic diversity (Figure 6.1) imply that just a few loci are involved, allowing more tractable genetic examination in dogs compared to in humans.

Finally, dogs receive the most medical intervention after the human. As a dog's lifespan is approximately seven times shorter than humans, progression of disease is much faster, allowing a shorter time period (1-3 years) for clinical trial compared to human (5-15 years) (417)(418). Therefore, dogs may serve as an effective testing ground for novel therapies.

6.2.3 Trait mapping strategies for the dog genome

Most breeds have been created less than 200 years ago via strong artificial selection, keeping their linkage disequilibrium (LD, i.e. non-random association of alleles at two or more loci) and exceptionally long haplotype blocks (i.e. regions in which little recombination has occurred) (Figure 6.1). These unique genetic features make pure-bred dogs particularly amenable to genome-wide association (GWA) mapping, with fewer markers and fewer individuals required compared to humans (419)(420). Such a long-range genetic pattern has been experimentally confirmed, typically with several megabase-long LDs, 40-100 times longer than in humans (404)(421). The domestic dog population, however, is ancient, hence pattern of long LD has been broken down over time (422). In turn, the LD pattern of the whole domestic dog population is comparable to that of humans (404) (Figure 6.1). Thus, a mapping strategy that takes into account both the long-range LD within a breed and the short-range ancestral haplotypes that are shared across breeds provides a powerful tool for genetic studies (Figure 6.2).

Two-stage mapping, in which a disease locus may be identified within a breed, then narrowed down using multiple breeds, can take full advantage of the bimodal distribu-

tion of LDs in the domestic dog population (Figure 6.2). As dog breeds are young, they have had insufficient time to accumulate substantial number of novel mutations. Thus if multiple breeds are predisposed to the same disease, they are likely to share the same disease mutations that arose from the ancient haplotype that were carried into these breeds. However, identifying these ancestral blocks harbouring causative mutations in a cross-breed study requires a highly dense SNP array, approximately one SNP every 1-2kb. Therefore, a genome-wide scan may be most efficient in a single breed, with additional multiple breeds with the shared trait to be mapped in a fine-mapping step (Figure 6.2). Below I will discuss this two-stage strategy in the context of the two major mapping methods used in dogs, i.e. association mapping and selection mapping.

6.2.3.1 Association mapping

In the first stage, GWA in a breed can be performed to identify disease-associated loci that are ∼1Mb long. The dog genome project has proven that at least 15,000 SNPs are sufficient for association mapping genome-wide (404). The number of samples that are required varies depending on the inheritance mode, from dozens [for Mendelian recessive traits] to hundreds [for complex traits] (404). In the second stage, dogs from multiple breeds that manifest the same disease should be included. Examining the associated regions identified in the first stage across multiple breeds with a much denser array of SNPs, then permits the identification of smaller regions of 10-100kb for the disease (Figure 6.2).

6.2.3.2 Selection mapping

In addition to the association mapping of disease genes that leverages the power of genetically isolated dog breed populations (420)(421)(404), selection mapping may provide additional power to detect disease-underlying regions. As domestic dog breeds arose from the fixation of separate mutations via strong artificial selection, we would expect to observe an increased frequency of a genetic variant linked to the genes that are selected for breed-defining traits (i.e. genetic hitchhiking), which in turn reduces the genetic variation of the region in the population (i.e. selective sweep). If several breeds share the same trait or predispose to the same disease, the genes that are responsible for these traits or diseases may lie in the regions that have been subjected to selective sweeps that occurred at the same genomic locations across these breeds.

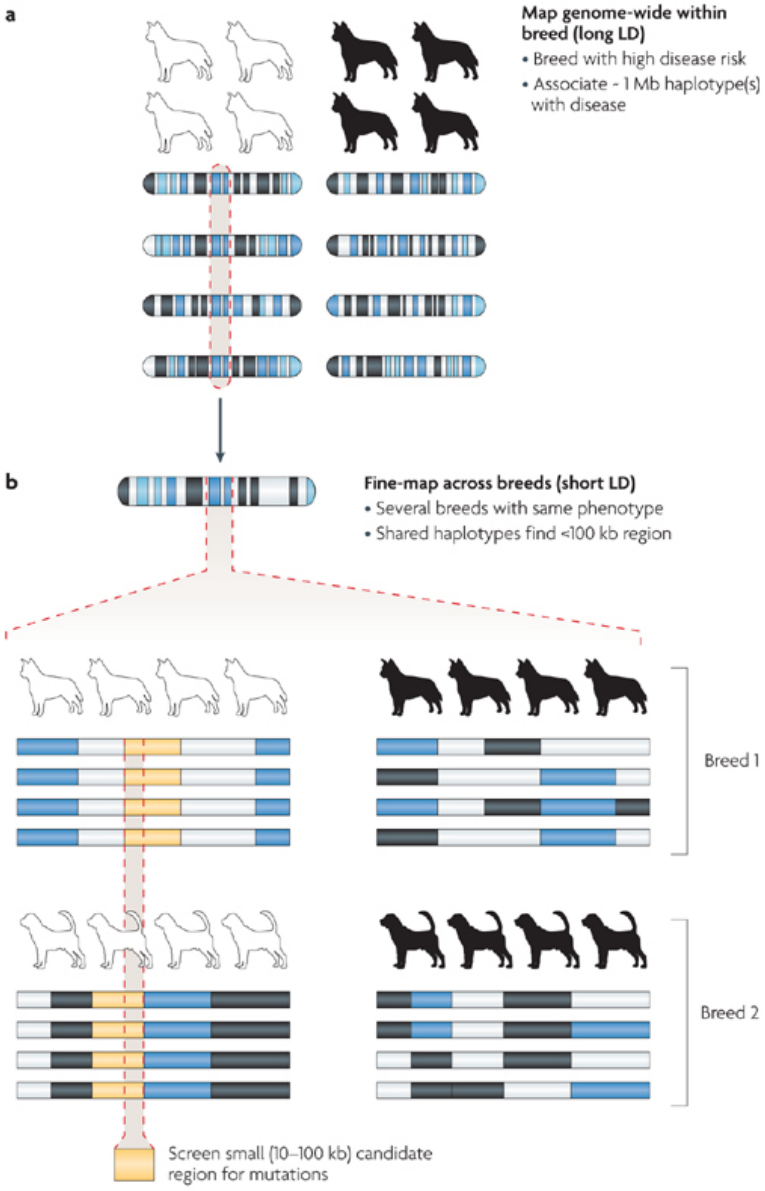


Figure 6.2: Two-stage trait mapping strategy in dogs - In order to identify genetic mutations underlying a disease phenotype, dogs with disease phenotype (white) and the normal population (black) are compared. a) In stage one, genome-wide association analysis within a disease-predisposed breed with dozens to hundreds of samples and ≥15,000 SNPs may identify an associated region(s) of ∼1Mb long. b) In stage two, fine-mapping with a much denser set of SNPs in multiple breeds that share the same phenotype refines the associated region(s) to regions of 10-100kb. This candidate region, which corresponds to the shared ancestral haplotype that carries the causative mutation, can then be screened for functional variants that underlie the disease. Adapted from (407).

In summary, across-breed approaches (including both selection and association mapping) offer great advantages in mapping traits that are fixed in breeds. Moreover, due to a high prevalence of a certain phenotype within a breed, only a small number of individuals are required to represent the whole population of the breed. However, it is important to remember that multi-breed approaches assume that breeds with the same phenotype have the same causative mutations arising from the pre-breed domestic dogs. Therefore, phenotypes with heterogeneous origins (e.g. where separate random mutations or standing variants cause the phenotype) will be more challenging to map.

6.2.4 Canine Compulsive Disorder (CCD)

Among several neuropsychiatric diseases, of which counterparts exist in both dogs and humans (Table 6.1), my study particularly concerns compulsive disorder. Therefore, in this section, I will discuss the validity of Canine Compulsive Disorder (CCD) as a model for human OCD. In addition, I will describe where my research lies in the context of the trait mapping strategies mentioned above.

Dogs	Humans
Canine compulsive disorder	OCD
Canine cognitive dysfunction	Alzheimer's disease
Dominance, Aggression	Impulse control disorders
Panic disorders, Noise phobia	Panic disorders
Separation anxiety	Social and attachment anxieties, Separation anxiety

Table 6.1: Human neuropsychiatric disorders and its canine counterparts - The table shows human psychiatric conditions and potentially analogous or homologous conditions in domestic dogs. Adapted from (414).

6.2.4.1 Face validity of CCD

CCD is usually recognised by ritualistic, stereotypic behaviours of affected dogs. Such behaviours include circling, tail-chasing, flank sucking (particularly in Dobermans), fence running, fly-biting, self-mutilation (e.g. acral lick granuloma, neurotic dermatitis), hair/air biting, pica, pacing/spinning, staring and vocalising, wool/fabric-sucking and self-directed vocalising (414)(423)(424).

Despite little understanding of diseases aetiology for either dogs or humans, symptomatology is remarkably similar in the two species (i.e. face validity). Both human

OCD and CCD are characterised by repetitive, ritualistic behaviours that are otherwise required for normal function, the extreme execution of which interferes with normal, daily activities and functioning (414). Over the course of the disease, these behaviours become exaggerated in form as well as in duration in both species. Moreover, just as in humans where the patients may restrain their behaviours in the presence of others, flank sucking or tail chasing dogs [who have nothing wrong with their flank or tail] may, after frequent rebukes, remove themselves from the sight of the owners and commit this behaviour elsewhere, or cease the behaviour upon approach or physical restraint then resume when such controls disappear (414). This particular component of CCD is perhaps the translation of human patients' verbal/written response to an assessment of their obsession (414).

6.2.4.2 Predictive and construct validity of CCD

The drug response in CCD is strikingly similar to that in human OCD. This is also known as predictive validity which requires an analogous response to treatments that prevent or reverse the symptoms in the human disease. As for human patients, ∼50% of CCD dogs respond to a SSRI, fluoxetine (425), or a tricyclic antidepressant, clomipramine (426)(427)(428)(424). In addition, as in the treatment of human OCD, in the absence of pharmacological or behavioural treatments, CCD relapses in many cases (429). Also, symptoms may worsen in a stressful environment or anxiety producing occasions in both humans and dogs (140)(429).

Given the lack of neuroanatomical examination of CCD affected dogs and poorly understood genetic basis of compulsive disorder in both dogs and humans, it is difficult to assess the construct validity (i.e. having comparable aetiology) of CCD. However, similarities in disease progression and its prevalence between the two species suggest that CCD and OCD may involve the same biological dysfunction. As with childhood onset OCD in humans, CCD starts between pre-pubescence and early social maturity, and more males are affected than are females (430)(426). As is true for humans, dogs that are first degree relatives of CCD dogs are usually at higher risk and manifest endophenotypes of CCD, suggesting a strong heritable component for disease aetiology in both species (414). In addition, both CCD and OCD manifest in a range of diverse symptoms (e.g. washers, checkers [for lights, gas jets, locks etc.], and trichotillomania [i.e. hair-pulling] for humans; suckers, chasers, biters, etc. for dogs), suggesting

complex, heterogeneous aetiologies for both. Taken together, CCD is an excellent analogous and possibly even homologous model for human OCD. Thus, investigating the genetics of CCD, in which the genetic study may be more tractable due to the unique population history of the domestic dog, may provide insights into the aetiologies of both CCD and human OCD.

6.2.4.3 Previous genetic mappings for CCD

Previously, the GWAS on CCD with Doberman Pinscher dogs identified a genome-wide significant association with the neural cadherin gene, *CDH2* ($p = 7 \times 10^{-7}$, $p_{genome} = 0.01$) (357). Other regions that also showed strong associations include further genes (unpublished; Appendix Tables 6.16 and 6.17) that are potentially linked to the CSTC brain circuit that may be responsible for OCD (see section 5.2.1.2). Given the unique population history from the strong artificial selection for behaviours in dog breeds, it is plausible that the ancestral haplotypes harbouring mutations underlying CCD were transmitted to the modern breeds that are predisposed to CCD. Thus, a selective sweep test across multiple breeds that have a high prevalence of CCD (e.g. Doberman Pinschers, German Shepherds, and Pit Bull Terrier) was also performed and identified further genes in the same pathway (Appendix Table 6.17; unpublished).

In my study, the regions identified from the association mapping in Doberman Pinschers and those from selection mapping using several CCD predisposed breeds were resequenced, which permitted the screening of functional variants. In order to efficiently define these 26 CCD candidate regions encompassing a total of $\sim$ 5Mb of sequence, affected dogs from several breeds and the breed-matched healthy controls were employed (Figure 6.2 and Appendix Table 6.15) to be resequenced. To this end, I sought to identify functional variants, genes and/or biological pathways that are commonly disrupted among affected dogs across multiple breeds that may share an aetiology with human OCD.

6.3 Materials and Methods

6.3.1 Sample collection

In this study, 12 case and 12 control dogs were employed (Appendix Table 6.15). Specifically, the cohort includes 4 Doberman Pinschers, 2 German Shepherds, 1 West Highland White Terrier, 1 Shiba Inu, 2 Bull Terriers, 1 Shetland Sheepdog, and 1 Jack Russell Terrier with CCD and the same number of healthy control dogs for each breed (Figure 6.5). In particular, the 4 affected Doberman Pinschers were selected because of their possession of a known risk haplotype for CCD, i.e. each dog with identifier *15260* and *15255* to be homozygous for risk allele T at Chr7:63867472 and *15237* and *15259* to be heterozygous at this position. Blood (3-10ml depending on the size of the dog) or cheek swab (2 swabs) samples were acquired at local veterinary surgeries, veterinary schools or at dog shows and breed club meetings. Blood or cheek swabs were obtained by trained and insured veterinary staff under standard procedure (i.e. blood samples drawn from a vein on the front of the leg or from the vena cephalica). All dogs were privately owned and samples were collected with informed owner consent. The canine behavioural phenotyping was conducted by a veterinarian with a neurology specialisation.

6.3.2 Target resequencing and raw data processing

26 genomic regions of the dogs covering ∼5Mb was selected for resequencing. The selected regions include: *(i)* 9 genomic regions showing strong association ($p-value < 10^{-5}$) with CCD in Doberman Pinschers that together encompass ∼3.9Mb (Appendix Figure 6.22 and Appendix Table 6.16) (357), and *(ii)* 17 fixed regions showing selective sweep signals among multiple dog breeds that are predisposed to CCD that together cover ∼1.3Mb (Appendix Table 6.17).

The 24 samples were individually barcoded and the ∼5Mb target regions were captured using Nimblegen's customised Hybrid Selection array. The captured samples were then pooled into 4 batches, and each batch was sequenced in 2 lanes of the Illumina GA2. The generated sequence reads were mapped to the CanFam2.0 assembly and, duplicates marked, realigned, and recalibrated as described in section 5.3.3.

6.3.3 SNPs and micro-indels discovery and genotyping

In general, I followed the framework for best practice variant detection that was developed by the Genome Analysis Toolkit (GATK) team (367). Briefly, once I processed the raw sequence data and produced analysis-ready BAM files, I called genomic sites that appear to possess non-reference alleles using the Unified Genotyper of GATK (366). This then produced a multi-sample VCF file, with polymorphic sites discovered and genotyped for each sample in the cohort. Once this raw VCF file was generated, I applied the variant quality score recalibration (VQSR), which builds an adaptive error model using known variant sites and applies the error model to estimate the probability of each SNP to be a true genetic variant or else a machine artefact.

As there were insufficient data to build a robust error model for micro-indels, these I manually filtered using VariantFiltrationWalker. To do this, I first separated SNPs and micro-indels using the SelectVariants module, recalibrated the SNPs (using VQSR) and hand-filtered the micro-indels in parallel, and combined high quality SNPs and micro-indels back into one analysis-ready VCF file, using the CombineVariants module. In order to validate variant calls, all samples were genotyped using the CanineHD Whole-Genome Genotyping array containing 170,000 markers placed on the CanFam2.0 reference sequence.

Finally, I summarised variants in terms of their novelty to previously reported dog SNPs CanFam2.0 (http://www.broadinstitute.org/mammals/dog/snp2) and Ti/Tv ratio using the VariantEval module, and annotated the detected variants with Ensembl build CanFam2.0 using the VariantAnnotator for downstream analyses.

6.3.4 Structural variant discovery

GenomeSTRiP (431) was used to discover larger structural variants from the processed BAM files. GenomeSTRiP identifies putative deletions using two classes of information: *(i)* technical features of the sequence within a single genome, and *(ii)* population-scale patterns of putative structural variants across multiple genomes (Figure 6.3 and Appendix Figures 6.23 and 6.24). Within-genome sequence features include breakpoint-spanning reads (432)(433), aberrant read-pairs (434)(435)(436), and local variation in the depth of coverage (437)(438)(439). GenomeSTRiP then combines

these within-genome sequence features with the population-scale patterns across multiple genomes. Population-scale patterns that are examined include allele sharing among multiple genomes, sequence heterogeneity in a population, substitution of alternative structural alleles, and the haplotype structure of genomes, when available (Figure 6.3). In practice, I identified structural variants using 4 modules of GenomeSTRiP (Ap-

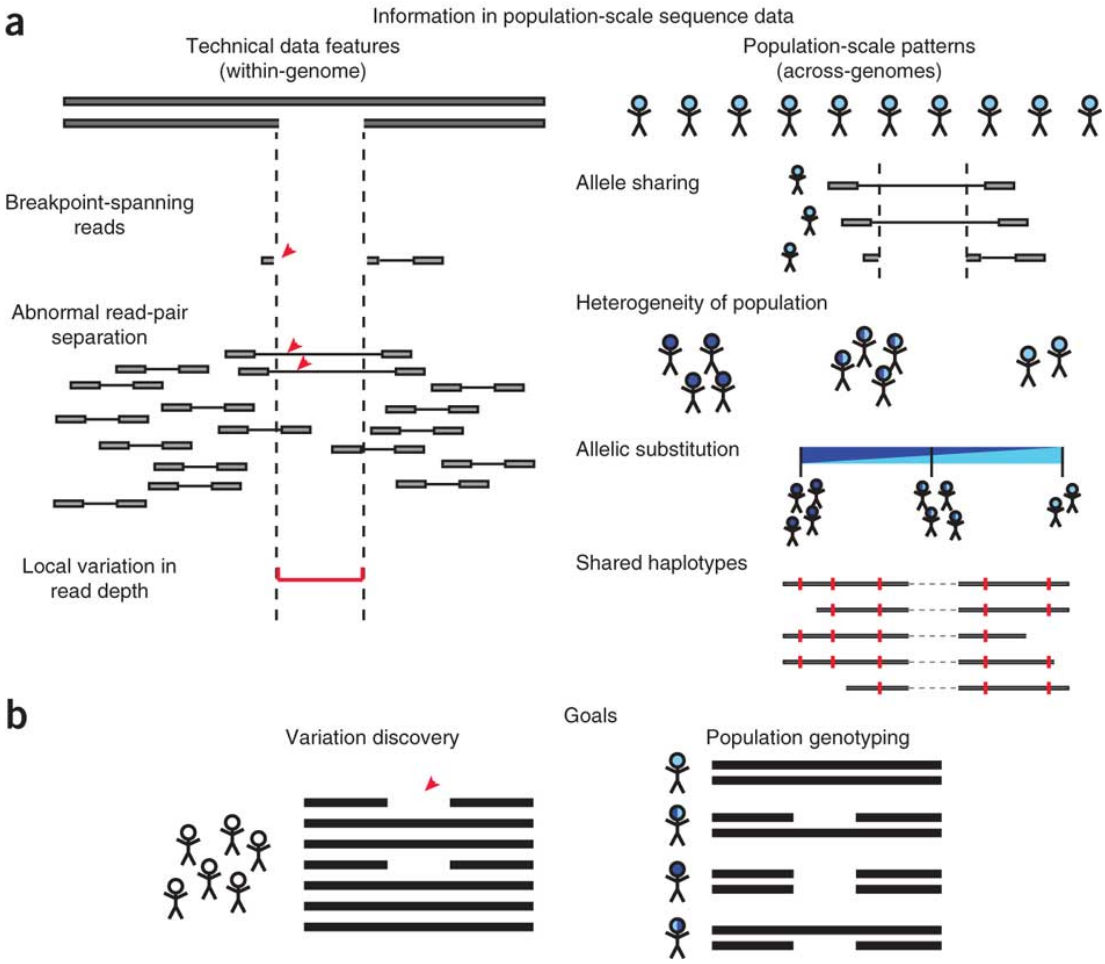


Figure 6.3: Analytical framework for Genome STRucture in Populations (GenomeSTRiP) - (a) Two types of information are used in GenomeSTRiP to discover structural variations, namely, technical-data features within a genome and population-scale pattern across multiple genomes. (b) Once structural variants are identified in a population, the genotypes of variants for each individual are determined. Adapted from (431).

pendix Figure 6.25), i.e. SVPreprocess to prepare input BAM files for other modules, SVDiscovery to discover putative deletions which produces a VCF file of potential vari-

ant sites, SVGenotyper to genotype identified deletions from SVDiscovery in a given set of genomes, and SVAltAlign to realign previously unmapped reads against putative alternate alleles. I then manually checked alignments of all discovered deletion sites for aberrant read-pairs and/or read-depth to ensure the reliability of the calls. I used the Integrative Genomics Viewer (IGV) (440) to visualise the pileup information of all alignments. Currently, GenomeSTRiP generates structural variant calls only for deletions.

6.3.5 Phenotype correlation analyses

6.3.5.1 SEQscoring for individual variants

Due to the small number of samples and multiple dog breeds, conventional case-control comparison was not possible in this study. Thus, I employed SEQscoring in order to evaluate detected variants based on their concordance with the individual's phenotype (Figure 6.4).

I first scored detected variants based on evolutionary conservation status at the genomic position of each variant using the Scoring module. In this module, data from any species can be employed, so long as constraint score data are available for the genome of the given species. For this study, I used constraint scores calculated from 29 mammals (3) that were lifted over to the CanFam2.0. For each variant, the scoring module checks whether it is located within a constrained element, and if not the distance to the closest such element is calculated.

Next, I ranked the variants with their constraint scores according to their concordance with the expected inheritance pattern, using the case/control module. The expected pattern can be chosen according to the most likely mode of inheritance, recessive or dominant/complex. For a recessive trait, homozygous cases are considered important, while for a dominant/complex trait, homozygosity for the wildtype in control individuals is primarily considered. In practice, the case/control module examines all pair-wise combinations of individuals and adds an extra score to the variant if two cases are homozygous for the non-reference allele for a recessive trait, or if two controls are homozygous with the reference allele for a dominant/complex trait. When the genotypes of a case and a control disagree, an extra score is also added to a dominant/complex trait. The values from the pair-wise comparisons are summed, generating

a 'pattern score' for each variant. Here, I chose the dominant/complex mode for computing pattern scores.

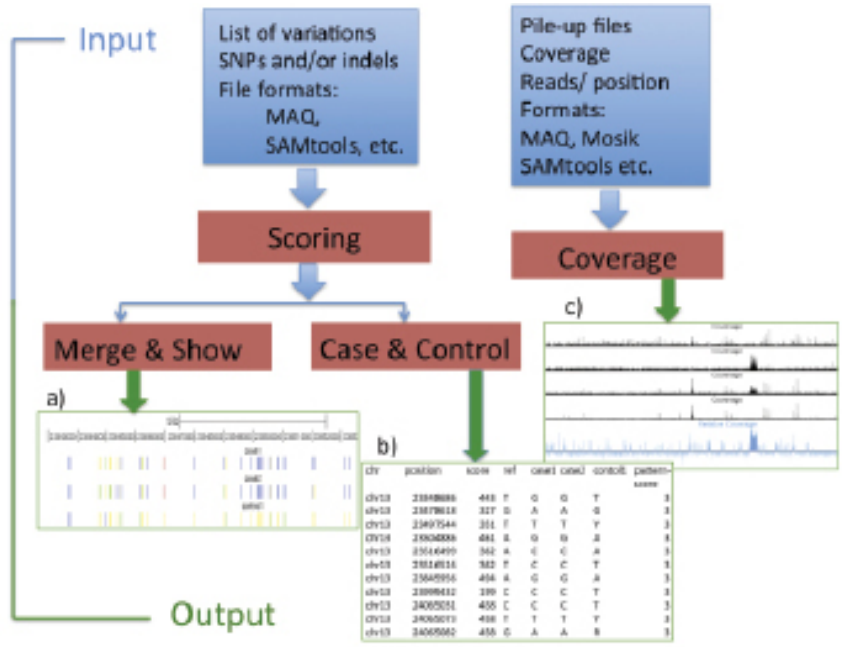


Figure 6.4: Overview of SEQscoring modules - SEQscoring takes lists of variants or coverage data in standard file formats such as MAQ (441) and SAMtools (442). First, the Scoring module scores variants (SNPs/indels) based on evolutionary conservation status of the genomic position for each variant. Then, (a) the Merge & Show module colour-codes these variants according to genotype and conservation status of the variants, which may facilitate visulisation of haplotypes and functional sequences. (b) The Case & Control module scores each variant or region based on its concordance with phenotype. (c) The Coverage module calculates differences in the depth of coverage between cases and controls in an attempt to localise structural variants. Adapted from (443).

6.3.5.2 Haplotype Likelihood Ratio (HLR) test for regions

The HLR test is a multi-marker association test that allows one to search for regions rather than individual variants of high phenotypic concordance. The difference to a standard, χ^2-based multi-marker test, which looks for differences in haplotype frequencies between cases and controls is that the HLR test models the likelihood that a single haplotype is present dominantly in cases, while the likelihoods for all other haplotypes proportionally decreased. It produces a LOD (logarithm [base 10] of odds) score, which is the maximum likelihood estimate for haplotype frequencies observed in the affected

dog population ($O_1...O_k$), given the distribution in the control population ($f_1...f_k$):

$$LOD_{ML} = log_{10}\frac{P(0_1,...,0_k|\alpha e_j+(1-\alpha)(f_1,...,f_k))}{P(0_1,...,0_k|f_1,...,f_k)} \ ,$$

where j is the haplotype on which the mutation arose, $1-\alpha$ is the recombination rate, and $e_j = 1$ when i=j and $e_j = 0$ when $i \neq j$. The test produces maximum likelihood estimates for j (i.e. associated haplotype) and α. Thus, while a χ^2-based association test identifies significant differences in haplotype frequencies, the HLR test models a specific scenario that is common in rapid selective events that were likely to occur during dog breed creation (404).

For a closer view of the candidate haplotype regions, I used the Merge & Show module of SEQscoring (Figure 6.4). The Merge & Show module merges all variants and their conservation scores for each individual dog, and colour-codes the variants that can be displayed in the UCSC genome browser (see Table 6.2 for the specific colour-codes). Such colour-coding of variants based on their genotypes and conservation status immediately allowed an easy comparison and investigation of the haplotype structure across the multiple dogs.

Colour-code	Conservation status	Genotype
Red	Yes	NR/NR
Pink	Yes	R/NR
Blue	No	NR/NR
Green	No	R/NR
Yellow	No	R/R

Table 6.2: Colour-codes for detected variants based on their evolutionary conservation and genotype - Variants within or near (±5bp) constrained elements are marked with 'Yes' in the Conservation status column. R stands for reference allele, and NR for non-reference allele.

6.4 Results

Resequencing of genomic regions (∼5Mb in total) in 12 dogs with CCD and 12 healthy dogs matched by breed using a high-throughput sequencing technology (Illumina GAII) and Nimblegen's hybrid capture yielded (*i*) high-quality sequencing alignments of over half billion reads, and (*ii*) >25,000 SNPs, >7,800 micro-indels and >170 deleted CNVs. Of these, ∼300 variants were found to be potentially functional variants residing within coding sequences, untranslated regions, or splice sites. Over 400 variants overlapped evolutionarily conserved elements that may be an indication of a functional regulatory role for that genomic region. The detected variants were prioritised for CCD by various criteria such as phenotypic correlation, evolutionary conservation status and a known risk haplotype within the *CDH2* gene that previously showed genome-wide significant association with CCD in Doberman Pinschers (357). Publicly available bioinformatics data including promoter-like regions, enhancer-like regions, and transcription factor binding site motifs encompassing these top variants allowed me to generate experimentally tractable hypotheses for the potential roles of these top candidate variants for CCD. Finally, by examining a group of variants present only in cases I identified candidate genes, rather than discrete variants, for CCD, whose functions and contributing biological pathways may be shared with human OCD.

6.4.1 Targeted resequencing of 5Mb canine genomic regions yielded extensive high-quality alignments

I resequenced 5Mb of canine genomic regions for 12 dogs with CCD and 12 healthy controls that have previously been identified by a GWAS (Appendix Figure 6.22 and Appendix Table 6.16) or evidence of a selective sweep among the dog breeds that are predisposed to CCD. In order to identify common mutations for CCD across breeds, dogs were selected from multiple breeds, and control dogs were matched by breed (Appendix Table 6.15). *Illumina*'s high-throughput sequencing machine (*Genome AnalyzerII*) produced a total of >550 million reads, divided into 12-39 million reads per individual dog. The reads were subsequently mapped to the reference canine genome, *CanFam2.0*, at mapping quality of >Q20 (i.e. 1/100 chance of being an incorrect alignment; *PF HQ Aligned Reads* in Appendix Table 6.19). Of the genomic regions covered by high-quality alignments, >53 billion bases (1.2-3.3 billion bases per individual) were

confidently called at 99% or higher accuracy (*PF HQ Aligned Q20 Bases* in Appendix Table 6.19). Of the Q20 reads, a total of ∼200 million reads were uniquely mapped to the reference genome, while the number of unique reads for each dog varied widely (1.4-13 millions per individual; *PF Unique Reads Aligned* in Appendix Table 6.19). In turn, the total number of unique callable bases was also reduced to ∼20 billions (0.14-1.3 billion bases per individual; *PF Unique Bases Aligned* in Appendix Table 6.19).

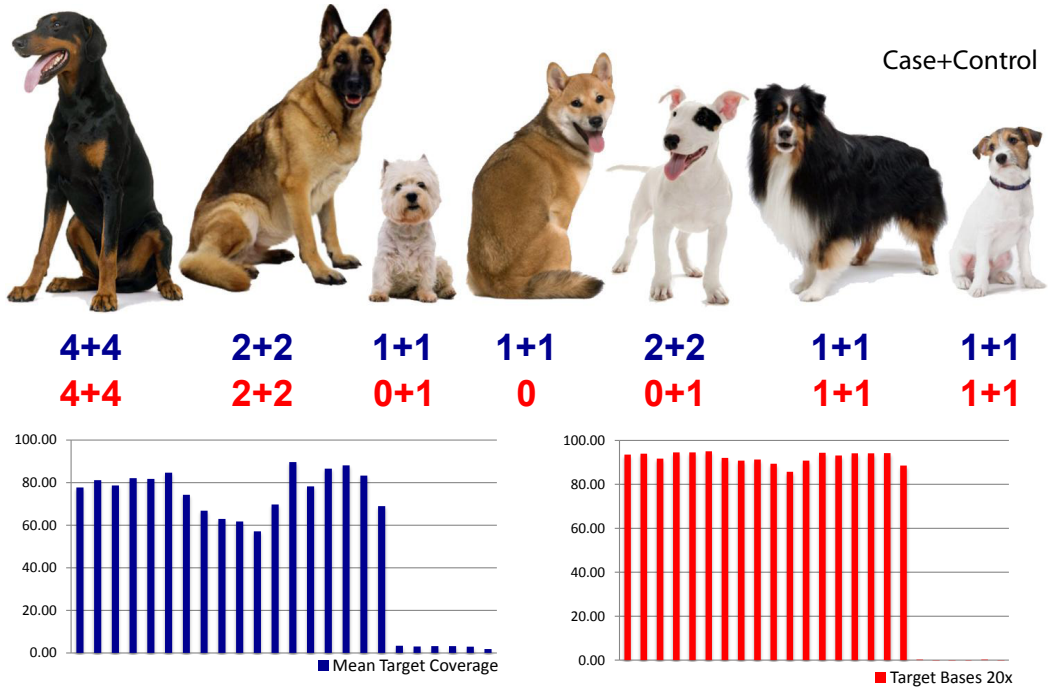


Figure 6.5: Sequenced samples and depth of coverage - The sequenced breeds include Doberman Pinscher, German Shepherd, West Highland White Terrier, Shiba Inu, Bull Terrier, Shetland Sheepdog, and Jack Russell Terrier (left to right). The numbers in blue indicate the number of dogs that were sequenced (# case dogs+# control dogs) and those in red indicate the number of dogs excluding the 6 samples with too low coverage (i.e. <4x). The mean target coverage for individual dogs is shown in blue bars and percentage of target bases that achieved >20x in red bars. For actual values, see Appendix Table 6.20.

Among the 24 dogs, 18 dogs achieved the mean target coverage of >76x (57-90x per dog; Figure 6.5 and *Mean Target Coverage* in Appendix Table 6.20), while read depth of the other 6 dogs were less than 4x (1.8-3.4x per dog; *Mean Target Coverage* in Appendix Table 6.20). In hybrid capture and sequencing step, experiments were conducted in 4 batches, each containing 6 samples. The duplication rate in the batch containing the 6 low-quality samples was particularly high, as indicated by the fact

that the numbers of aligned reads and called bases did not differ from those for the other 18 dogs with high coverage, while the numbers of uniquely mapping reads and bases were particularly small for these 6 samples (Appendix Table 6.19).

Among the 18 dogs with deep coverage, over 92% of the targeted region achieved target coverage of over 20x (86-95% per dog; *Target Bases 20x %* in Appendix Table 6.20), and 99% of all target bases had at least 2x per individual (99-99.64% per dog; *Target Bases 2x %* in Appendix Table 6.20). Almost all target bases were captured - 0.5% of the bases had no coverage (0.3-0.84% per dog; *Zero Coverage Targets %* in Appendix Table 6.20). Also, of these 18 dogs, 16 samples were valid case-control samples that were matched by breed (i.e. 4 pairs of Doberman Pinschers, 2 pairs of German Shepherds, 1 pair of Shetland Sheepdogs, and 1 pair of Jack Russell Terriers; Figure 6.5), and the two remaining dogs (West Highland White Terrier and Bull Terrier) were control samples.

6.4.2 A large number of SNPs and micro-indels including functional variants were detected via GATK

Using GATK (366)(367), I detected a total of 22,891 SNPs and 5,093 micro-indels (i.e. 2,284 insertions and 2,809 deletions that are <20bp) from the high-quality processed alignments (i.e. alignments after local realignment, duplicate marking, and base quality recalibration; see Materials and Methods) of the 16 case-control breed matched dogs: 7,981-11,148 SNPs, 1,072-1,408 micro-insertions, and 1,297-1,702 micro-deletions per dog (*Total* in Table 6.3, Appendix Figure 6.26, and Appendix Table 6.18).

Of these variants, 5,800 were known SNPs, leaving 75% (17,091/22,891) of the detected SNPs and all of the micro-indels being novel (Appendix Table 6.18 and Appendix Figure 6.27) (404). While a majority of the detected SNPs were intronic or intergenic, 173 SNPs (64 missense and 109 silent) were found to be coding mutations based on *CanFam2* Ensembl annotation (*Functional Types* in Table 6.3, Appendix Figure 6.28, and Appendix Table 6.18). 18 and 70 SNPs were found within 3'- or 5'-UTRs or intronic splice sites (i.e. <10bp of exon boundaries), respectively (*Functional types* in Table 6.3). For micro-indels, four and ten were found within UTRs and splice sites, respectively. Three micro-indels were predicted to cause a frameshift that would alter the open reading frame, resulting in the translation to different amino acids from the reference (*Functional types* in Table 6.3).

	SNP	INDEL
Total	22,891	5,093
Functional types		
Missense	64	NA
Nonsense	0	NA
Silent	109	NA
UTR	18	4
<10bp from splice site	70	10
frameshift	NA	3
Conservation score		
>7	217	65
>8	83	21
>9	16	0

Table 6.3: Counts for detected variants in terms of mutation types and evolutionary conservation - *Total* row shows the total count for detected SNPs and micro-indels (<20bp) among 16 case-control breed matched dogs. The *Functional types* table shows the number of variants for each functional category. The *Conservation score* table shows the number of variants within evolutionarily conserved regions and its associated conservation score (the higher the score, the more conserved the regions) (3).

In addition, over 400 variants (including both SNPs and micro-indels) were found to reside within evolutionarily conserved elements with conservation scores higher than 7 (*Conservation score* in Table 6.3), which indicates strong conservation of sequence throughout evolution according to the 29 mammals genome project (3). Also, the detected number of variants in the strongly conserved regions well exceeded the background distribution of conserved elements within the 5Mb target region (Figure 6.6). The excess of the variants in these functional sequences might reflect many, mild effect disease-contributing mutations that occurred among the affected dogs.

6.4.3 Replication of variant calls and the Ti/Tv ratio

In order to confirm the validity of variant calls from the resequencing data, the identical samples were genotyped using an independent technology (Illumina's CanineHD BeadChip). SNP calls for 370-394 positions of each dog from the sequencing and the array experiments were compared, and on average 99% (93.7-99.8%; # All positions column in Appendix Table 6.21) of SNP calls from the two experiments were identical. More specifically, calls for homozygous reference positions were the most accurate at an average of 99.8% (98.6-100%; # HomRef positions in Appendix Table 6.21). Homozygous alternative positions had the second highest accuracy with an average 99.5%

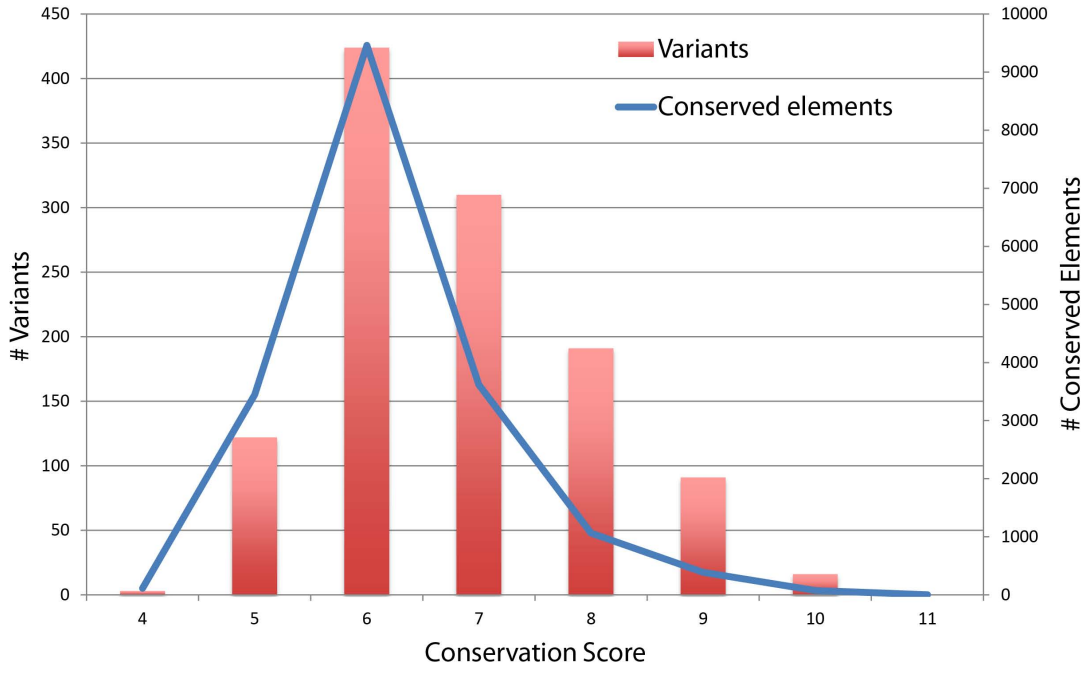


Figure 6.6: Histogram of detected variants and conserved elements in target regions - The red bar shows the number of detected SNPs and micro-indels among all the 18 sequenced dogs (y-axis on the left) for each bin of the conservation score (x-axis) calculated from the 29 mammals genome project (3). The blue line shows the number of conserved elements (y-axis on the right) for each conservation score bin (x-axis) within the ∼5Mb target region, showing the background distribution of conserved elements within the targeted genomic regions.

(98.3-100%; # HomAlt positions in Appendix Table 6.21). Heterozygous positions had the lowest accuracy with an average of 97.2% (79.8-100%; # Het positions in Appendix Table 6.21).

Another widely used way to confirm the validity of SNP discovery in sequencing experiments is to consider the Ti/Tv ratio. The Ti/Tv ratio is the ratio of the rates of transitions (i.e. point mutation that changes a purine nucleotide to another purine [A↔G] or a pyrimidine nucleotide to another pyrimidine [C↔T]; abbreviated to Ti) to transversions (i.e. substitution of a purine for a pyrimidine or *vice versa*; abbreviated to Tv). As there are twice as many possible transversions as transitions (transition - 4 possible changes; transversion - 8 possible changes), the expected Ti/Tv by random chance is 0.5. However, for spontaneous neutral mutations, transitions appear to be more favoured in the genome than are transversions (444), making the expected Ti/Tv ratio approximately 2-3 (445). Indeed, the Ti/Tv ratio of the detected SNPs was 2.76 (2.58-2.74 per dog, Appendix Table 6.22 and Appendix Figure 6.29), suggesting that the variant calls are likely of high accuracy.

6.4.4 Detected variants were prioritised by case-control comparison and using a known risk haplotype

Due to the small sample size and the diverse genetic differences across multiple breeds employed here, association tests were not applicable. Therefore, given that most of the target regions were shown to have strong association with CCD (Appendix Table 6.16 and Appendix Figure 6.22), I looked for obvious differences in the presence of detected variants between cases and controls and prioritised those variants that may contribute to CCD, using several functional criteria. Note that in this section I only examined variant calls from the 16 dogs that were breed matched case-control samples, to allow appropriate case-control comparison.

6.4.4.1 More functional variants were exclusively observed among cases

First, I examined whether there were variants that were found in any case dogs, but never in the control dogs, based on the assumption that those that are unique to affected dogs are more likely to be contributing to CCD. In total, 2,291 variants (2,002 SNPs and 289 micro-indels) were found only among case dogs (the intersection of the *All Variants* row and the **ALL** column of the sub-table **Case Unique** in Tables 6.4-6.6). Of these,

1,355/2,291 (59%) variants (i.e. 1,176/2,002 [58.7%] SNPs and 179/289 [61.9%] micro-indels) were genic (the intersection of the *All Variants in Genes* row and the **ALL** column of the **Case Unique** sub-table in Tables 6.4-6.6). 119/2,291 (5.2%) variants (i.e. 108/2,002 [5.4%] SNPs and 11/289 [3.8%] micro-indels) were within conserved elements (the intersection of the *Conserved Variants* row and the **ALL** column of the **Case Unique** sub-table in Tables 6.4-6.6), and 90/2,291 (3.9%) variants (i.e. 82/2,002 [4.1%] SNPs and 8/289 [2.8%] micro-indels) were within genic, conserved regions (the intersection of the *Conserved Variants in Genes* row and the **ALL** column of the **Case Unique** sub-table in Tables 6.4-6.6).

Altogether, there was no significant bias in the number of case-only variants compared to control-only variants (paired one-sided Wilcoxon signed rank test; $p = 0.63$ for all case-control pairs and $p = 0.41$ for the case-control pairs excluding the lower-quality Shetland Sheepdog pairs; see Table 6.7). On the other hand, when considering the case- and control-only variants within conserved elements, the case dogs appeared to have a greater number of exclusive variants (6-47 variants; *Conserved Uni Var #* column in **CASE** sub-table in Table 6.7) than control dogs (1-42 variants; *Conserved Uni Var #* column in **CONTROL** sub-table in Table 6.7). This greater burden of case-unique variants in case dogs became significant once lower-quality samples were excluded from the comparison (paired one-sided Wilcoxon signed rank test; $p = 0.075$ for all case-control pairs and $p = 0.018$ for the case-control pairs excluding lower-quality Shetland Sheepdog pairs). This suggests that there may be more functionally disrupting variants [that are suggested by evolutionary conservation status] among the affected dogs, and these functional variants may be contributing to CCD.

SNPs+INDELs

	ALL		NO CDH2		CDH2 ONLY	
	Variant #	% Total	Variant #	% Total	Variant #	% Total
Case Unique						
All Variants	2291		2061		230	
All Variants in Genes	1355	59.14	1125	54.59	230	100.00
Conserved Variants	119	5.19	103	5.00	16	6.96
Conserved variants in Genes	90	3.93	74	3.59	16	6.96
Control Unique						
All Variants	3116		3065		51	
All Variants in Genes	2032	65.21	1981	64.63	51	100.00
Conserved Variants	91	2.92	90	2.94	1	1.96
Conserved variants in Genes	55	1.77	54	1.76	1	1.96

Table 6.4: Case unique variants - The **Case Unique** sub-table shows the counts for variants that were found in any case dogs but not in any control dogs, while the **Control Unique** sub-table shows variants that were found in any control dogs but not in any case dogs. The **ALL** column shows counts for the entire target region, while the **NO CDH2** column is for the target region excluding the *CDH2* gene and the **CDH2 ONLY** column is for the *CDH2* genic region only. The *All Variants* row indicates the counts of all variant for each corresponding column, while the *All Variants in Genes* are for those that are within genic regions. The *Conserved Variants* row indicates the number of variants that are within evolutionarily conserved regions, while the *Conserved variants in Genes* are for those in conserved, genic regions. The *Variant #* and the *% Total* sub-columns show respectively the number of variants and percentage of the corresponding variants to the *All Variants* for each column.

SNPs

	ALL		NO CDH2		CDH2 ONLY	
	SNP #	% Total	SNP #	% Total	SNP #	% Total
Case Unique						
All Variants	2002		1808		194	
All Variants in Genes	1176	58.74	982	54.31	194	100.00
Conserved Variants	108	5.39	93	5.14	15	7.73
Conserved variants in Genes	82	4.10	67	3.71	15	7.73
Control Unique						
All Variants	2698		2657		41	
All Variants in Genes	1744	64.64	1703	64.09	41	100.00
Conserved Variants	79	2.93	79	2.97	0	0.00
Conserved variants in Genes	48	1.78	48	1.81	0	0.00

Table 6.5: Case unique SNPs - see Table 6.4

INDELs

	ALL		NO CDH2		CDH2 ONLY	
	INDEL #	% Total	INDEL #	% Total	INDEL #	% Total
Case Unique						
All Variants	289		253		36	
All Variants in Genes	179	61.94	143	56.52	36	100.00
Conserved Variants	11	3.81	10	3.95	1	2.78
Conserved variants in Genes	8	2.77	7	2.77	1	2.78
Control Unique						
All Variants	418		408		10	
All Variants in Genes	288	68.90	278	68.14	10	100.00
Conserved Variants	12	2.87	11	2.70	1	10.00
Conserved variants in Genes	7	1.67	6	1.47	1	10.00

Table 6.6: Case unique micro-indels - see Table 6.4

	All Uni Var #	Conserved Uni Var #	Cons/All (%)	Concordance (%)
CASE				
Db_het_15237	233	12	5.15	98.99
Db_het_15259	80	6	7.50	99.75
Db_homo_15255	426	19	4.46	99.75
Db_homo_15260	290	15	5.17	98.99
Shetland_spinner_5991	437	27	6.18	99.49
JR_spinner_3094	966	47	4.87	99.24
GS_spinner_5913	383	14	3.66	99.75
GS_spinner_5990	422	26	6.16	99.49
CONTROL				
Db_normal_15227	156	3	1.92	99.75
Db_normal_15253	335	6	1.79	99.24
Db_normal_15254	207	4	1.93	99.49
Db_normal_15262	275	6	2.18	99.49
Shetland_normal_1737	1476	41	2.78	93.67
JR_normal_205	1204	42	3.49	97.22
GS_normal_2722	130	4	3.08	99.75
GS_normal_5989	204	1	0.49	99.49

Table 6.7: The number of case- or control-only variants found in each dog - Each of the **CASE** and **CONTROL** sub-tables shows the 8 case and the 8 control dogs that are matched by breed, respectively. The **All Uni Var #** column lists the counts of variants that are present in any cases but not in controls for the **CASE** sub-table and *vice versa* for the **CONTROL** sub-table. The **Conserved Uni Var #** column indicates those from the **All Uni Var #** column that are within evolutionarily conserved elements. The **Cons/All** shows the percentage of the case-/control-only variants within conserved elements to all the case-/control-only variants. The **Concordance(%)** column shows the concordance rate between the detected variants from the two independent experiments, sequencing and SNP array (Appendix Table 6.21). The prefixes *Db, Shetland, JR* and *GS* indicate Doberman Pinscher, Shetland Sheepdog, Jack Russell Terrier and German Shepherds, respectively.

6.4.4.2 The number of case-only variants within breed is consistent with the breed genetic structure

Subsequently, I employed a more stringent filter to ensure that all case dogs possess the variants, which are required to be absent from all control dogs. This was based on the assumption that case-only variants that are commonly present among the affected dogs are more likely to be disease contributing. As there was no variant that met these criteria across all four breeds, I sought to select such variants for each breed. This approach may also allow an identification of breed-specific CCD predisposing variants that might be neglected in the approach employed above. Specifically, I first examined a case-control pair in each breed. The number of variants that are present in all cases but not in controls widely varies between breeds, from 184 (Doberman Pinschers) to 5,425 (Jack Russell Terriers) (Table 6.8). This may be due to the varying number of case-control pairs in each breed, as well as the breeds' heterozygosities (Doberman Pinscher, 0.38; German Shepherd, 0.40; Shetland Sheepdog, 0.43; Jack Russell Terrier, 0.73) (446)(447) (Table 6.8).

Breed	Heterozygosity	# CC pairs	# Case only variants
Doberman Pinscher	0.38	4	184
German Shepherd	0.4	2	881
Shetland Sheepdog	0.43	1	3,081
Jack Russell Terrier	0.73	1	5,425

Table 6.8: Heterozygosity of each breed, number of case-control pairs, and number of breed-specific case unique variants - The Breed column shows the name of each breed where the case-unique variants are counted. The Heterozygosity column shows each breed's heterozygosity calculated from (446)(447). The CC pairs column shows the number of dog pairs of case and control for each breed. The # Case only variants column shows the number of variants that are present in all case dogs but not in any control dogs for each breed.

I then looked for variants that are present in all cases but not in any controls across multiple breeds, in order to identify variants that are more commonly shared among affected dogs in more than one breed. As mentioned above, there were no variants present in all cases but no controls across all four breeds where matched case-control samples were available. Thus, to further examine the multi-breed case-only variants, I grouped two or three breeds in every possible combination and compared the cases and controls within each group.

In the combinations of 2 breeds, I identified 8-672 case-only variants present in all cases. Similar to the single breed case unique variants that were described above, the case-control comparison in the three pair-combinations of Jack Russell Terrier, i.e. the most heterozygous breed, (DP+JRT, GS+JRT and SS+JRT in Table 6.9) yielded many more case unique variants (81, 53 and 672 for DP+JRT, GS+JRT and SS+JRT, respectively), compared to the other combinations (19, 8, and 31 for DP+GS, DP+SS, and GS+SS, respectively; Table 6.9).

The highest number of case-only variants among the Shetland Sheepdogs and the Jack Russell Terriers (672) might reflect their smallest sample size (2 cases and 2 controls) compared to the other two combinations (DP+JRT, 5 cases and 5 controls; GS+JRT, 3 cases and 3 controls), considering that some variants may appear as case-only variants, but in fact are private variants especially when comparing a small number of case-control pairs. Nevertheless, the unusually large number of case-only variants detected in the SS+JRT group (672 variants) compared to the other two groups (53 for GS+JRT and 81 variants for DP+JRT) might be due to the relationship of the genetic structures among dog breeds (446) (see Discussion).

Breed	# CC pairs	# Case unique variants
DP+GS	6	19
DP+SS	5	8
DP+JRT	5	81
GS+SS	3	31
GS+JRT	3	53
SS+JRT	2	672

Table 6.9: Number of case-control pairs and case-only variants across 2 breeds - The Breed column shows each pair of breeds, from which the variants are compared. DP, GS, SS, and JRT stand for Doberman Pinscher, German Shepherd, Shetland Sheepdog, and Jack Russell Terrier, respectively. The CC pairs column shows the number of case-control pairs for the corresponding breed groups. The # Case unique variants column shows the number of variants that are present in all case dogs but not in control dogs within the corresponding breed groups.

Subsequently, I repeated the same analysis on groups of three breeds. When comparing the variants across three breeds, there were only four variants that were present in all cases but none of the controls. Of these, the two variants that were each detected in the DP+SS+JR and the DP+JR+GS groups were previously known SNPs (**SNP**

ID in Table 6.10). The other two variants were identified in the SS+JR+GS group: a micro-insertion (C→CT) at Chr34:38993178 and a micro-deletion (GAGAA→G) at Chr18:57224297. Although the candidacy of these variants remained uncertain as there were several control dogs possessing these variants in other breeds (calculated as the difference between the **# Individual** and the **# Case** columns in Table 6.10), these variants still have a potential to be variants that predispose certain breed groups to CCD.

Breeds	#Var.	Position	REF/ALT	SNP ID	AC	#Ind.	#Case
DP+SS+JR	1	chr7:61646990	G/A	BICF2G630561659	14	11	6
DP+SS+GS	0	N/A	N/A	N/A	N/A	N/A	7
DP+JR+GS	1	chr18:48998920	C/T	BICF2P1021589	11	10	6
SS+JR+GS	2	chr34:38993178	C/CT	N/A	13	13	4
		chr18:57224297	GAGAA/G	N/A	16	11	4

Table 6.10: Case only variants across three breeds - The **Breeds** column shows the combinations of three breeds, from which the variants are compared. DP, GS, SS, and JRT stand for Doberman Pinscher, German Shepherd, Shetland Sheepdog, and Jack Russell Terrier, respectively. The **#Var.** column shows the number of variants that are present in all case dogs but none of the control dogs in the three breeds shown in the **Breeds** column. The **Position** column indicates the genomic position of the case-only variants. The **REF/ALT** column shows the reference and alternative allele, respectively. The **SNP ID** column shows the SNP ID, if the case-only variant is known. The **AC** column indicates the allele counts within the 18 dogs for the case-only variants. The **#Ind.** column shows the number of dogs possessing this variant among the 18 dogs with deep sequence data. The **#Case** column shows the number of affected dogs possessing this variant.

6.4.4.3 Many variants showed phenotypic correlations

The approach employed in the previous sections to identify case-only variants is an immediate way to find variants with an extreme pattern of case-control differences, where the controls are homozygous wild type while only the cases have mutant alleles. However, the inheritance mode of a complex disorder such as CCD is much more complicated than this. Furthermore, this approach may neglect a plausible scenario for CCD, where the controls could be heterozygous with one susceptibility allele and one wild type allele, while the cases are mutant homozygous. Therefore, I employed a phenotypic correlation scoring scheme that takes into account both the wild type homozygosities for the controls (as considered in case-only variant approach) and the complex patterns of genotypic differences between cases and controls, using SEQscoring

(443). The dominant/complex mode of SEQscoring examines every pair-wise combination of individuals, and adds +1 to the score when the two controls are wild type homozygotes and when the genotypes of a case and a control differ (Figure 6.7; see Materials and Methods).

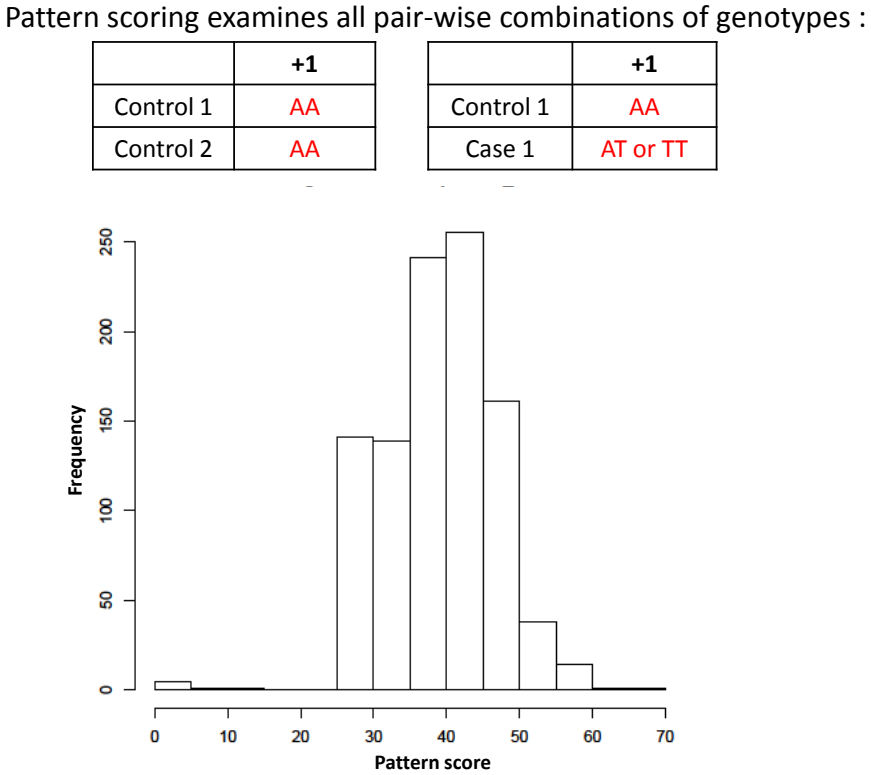


Figure 6.7: Phenotypic correlation scoring scheme for dominant/complex traits and histogram of phenotype scores for the detected variants - The pattern score of each single nucleotide variants within conserved regions was calculated using the dominant/complex inheritance mode of SEQscoring (443). The dominant/complex mode considers homozygosity of controls for the wild type allele. Therefore, it simply examines all pair-wise combinations of individuals, adding +1 to the score if the two controls are homozygous and in agreement with a dominant/complex trait, and +1 when a case and a control disagree in genotype. For each variant, the values are summed for all pair-wise comparisons, making a total score for each variant.

As a result, I scored 999 single nucleotide variants that reside within conserved regions (i.e. any regions that have SiPhy scores calculated from the 29 mammals project (3)), based on their phenotypic correlation, ranging from 0 to 70 with the median score of 39 (Figure 6.7). There was no evidence of correlation between the

variants' pattern score and their conservation score (Appendix Figure 6.31). I ranked the variants by their phenotypic correlation, based on the assumption that all the variants examined here are potentially functional, given evolutionary conservation at their genomic position, despite the varying degrees of conservation.

Of the scored 999 variants, all of the top 10 variants with the highest phenotype correlation scores (pattern score ≥60; Table 6.11) were found within genomic regions that showed strong associations with CCD (the top two variants in Chr7:61599000-61916800, $p = 1.58 \times 10^{-5}$ and the 3rd-10th variants in Chr7:63734000-63968500, $p = 2.06 \times 10^{-5}$; Appendix Tables 6.16 and 6.17). In fact, it was expected that several high pattern scoring variants in Chr7:63734000-63968500 would be observed, because Doberman Pinschers, the dog breed with the largest sample size, were selected for their possession of the phenotypically concordant risk alleles in this region (Materials and Methods; for a detailed analysis of this known risk haplotype, see section 6.4.4.4). Although the top two variants were found to be in a gene-desert region, the regulatory marks in their neighbouring regions may suggest a functional change in gene regulation by the variants in this region (see Discussion).

Chromosome	Position	Pattern score	SiPhy_29mammals
chr7	61632805	70	6
chr7	61834316	61	5
chr7	63779775	60	6
chr7	63796857	60	6
chr7	63796858	60	6
chr7	63802530	60	6
chr7	63806661	60	5
chr7	63814172	60	6
chr7	63814306	60	7
chr7	63814541	60	6

Table 6.11: Top 16 variants with high pattern scores (≥ 60) - Chromosome and **Position** columns show the genomic position of each top variant. **Pattern score** and **SiPhy_29mammals** show the phenotype correlation score and the conservation [SiPhy] score calculated from the 29 mammals project (3), respectively.

6.4.4.4 Resequencing allowed the pinpointing of candidate variants within *CDH2*, the only gene with a genome-wide significant association

Previously, a genome-wide association study (GWAS) for CCD using 92 rigorously phenotyped affected Doberman Pinschers and 68 healthy controls identified a genome-wide significant association of a canine Chr7 region with CCD (357). Dodman *et al.* also showed that fine-mapping of this region revealed a shorter region of moderate association spanning roughly 400kb (at Chr7:63.8-64.2Mb), likely reflecting the long haplotypes found within dog breeds. Of multiple SNPs within this haplotype, the most significantly associated SNP was located within *CDH2* whose protein mediates synaptic activity-regulated neuronal adhesion (448). Dogs showing multiple compulsive behaviours had a higher frequency of the risk allele than dogs with a less severe phenotype (60% and 43%, respectively, compared with 22% in unaffected dogs), suggesting that this region contain a risk haplotype for CCD.

Accordingly, I paid special attention to the *CDH2* region, in order to pinpoint candidate variants within the known risk haplotype. From deep sequence data from the 18 dogs, 621 SNPs and 321 micro-indels were identified within *CDH2*. Of these, four were found to change coding sequence, four to reside within UTRs, and 21 to overlap with evolutionarily conserved elements (Appendix Figure 6.32). However, these 8 functionally annotated variants (i.e. 4 coding changes and 4 UTR changes) did not show apparent case-control differences in their allele frequencies (Appendix Table 6.33). The 8 most phenotypically correlated variants showed phenotype scores $\geq$60 (Appendix Figure 6.32; see section 6.4.4.3), likely reflecting the risk haplotype that was suggested from the previous GWAS (357).

In order to narrow down the $\sim$400kb GWA-interval from Doberman Pinschers into a shorter risk haplotype, I employed the Haplotype Likelihood Ratio (HLR) test (404) that calculates the maximum likelihood estimate for the associated haplotype frequencies in the affected dog populations (LOD score; see Materials and Methods).

A visualisation of the LOD scores indicated regions of high phenotype-genotype concordance within the *CDH2* gene (Figure 6.8). As expected, the region next to the GWA-SNP position (Chr7:63867472) (357) showed a cluster of high LOD scores, narrowing down the $\sim$400kb haplotype into an $\sim$10kb region. Zooming into this $\sim$10kb genomic region of high phenotypic concordance, I examined the homo-/heterozygosities of individuals, where two cases (15237 and 15259) were expected to be heterozygotes

and the others (15255 and 15260) were expected to be homozygotes for the risk allele (357) (Figure 6.9; Materials and Methods). 31 variants were identified within this region, and of these, the genotypes of the 21 variants matched the known haplotype information (357), suggesting that these variants may be good candidate risk alleles for CCD. Of these, the genotypes of 9 variants were found to be non-reference alleles among the affected dogs, and four of these were novel variants that may be associated with a higher chance to be a risk alleles for CCD (Table 6.12).

Filter	**# Variant**
Haplotype Region	31
Matching GWAS haplotype	21
Non-reference in cases	9
Not known variant	4
Not in Doberman controls	0
Conserved elements	0

Table 6.12: Variants in a haplotype within *CDH2* among Doberman Pinschers - The Haplotype region shown in the **Filter** column was determined by considering the first genotypically discordant flanking SNPs from the outermost genotypically concordant SNPs. The Matching GWAS haplotype row shows the number of variants whose genotypes among the affected dogs are concordant with their known haplotype information from the previous GWAS. Of these 21, the number of variants where a non-reference allele was present in the affected dogs (9) are shown in the row, labelled Non-reference in cases. Of these 9 variants, four variants were novel (shown in the subsequent row, labelled Not known variant). Also, there were no non-reference alleles that were exclusive to cases among these 9 variants, hence shown as 0 in the row, labelled Not in Doberman controls. None of these 9 variants were found inside of evolutionarily conserved elements (the row, labelled Conserved elements).

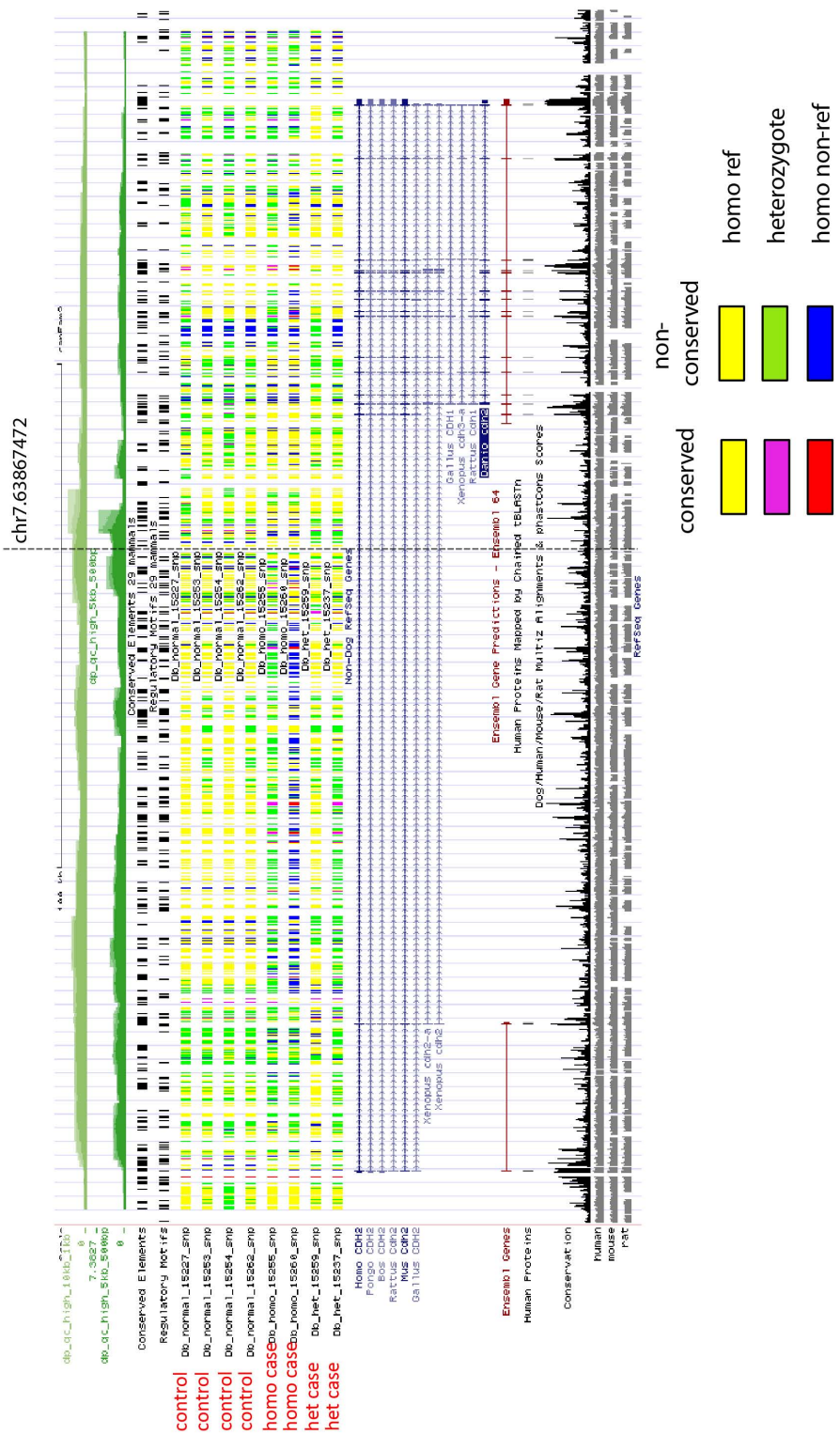


Figure 6.8: Genotypes and HLR scores within *CDH2* among Doberman Pinschers - The top two tracks show the maximum likelihood estimate for the associated haplotype frequencies among the affected Doberman Pinschers with different sliding windows and step sizes, light green for the 10kb window with a 1kb step and dark green for the 5kb window with a 500bp step. The 3rd and 4th tracks show evolutionarily conserved elements and regulatory elements inferred from 29 mammals project (3). The 5th-12th tracks show the genotypes of the identified SNPs for the Doberman Pinschers (5th-8th for controls, 9th-10th for affected homozygotes, and 11th-12th for affected heterozygotes). The colour codes for different genotypes are shown as in the within-figure legend. Lastly, *CDH2* transcripts and the conservation status of the region with various organisms are shown. The vertical line shows the SNP position showing genome-wide significant association.

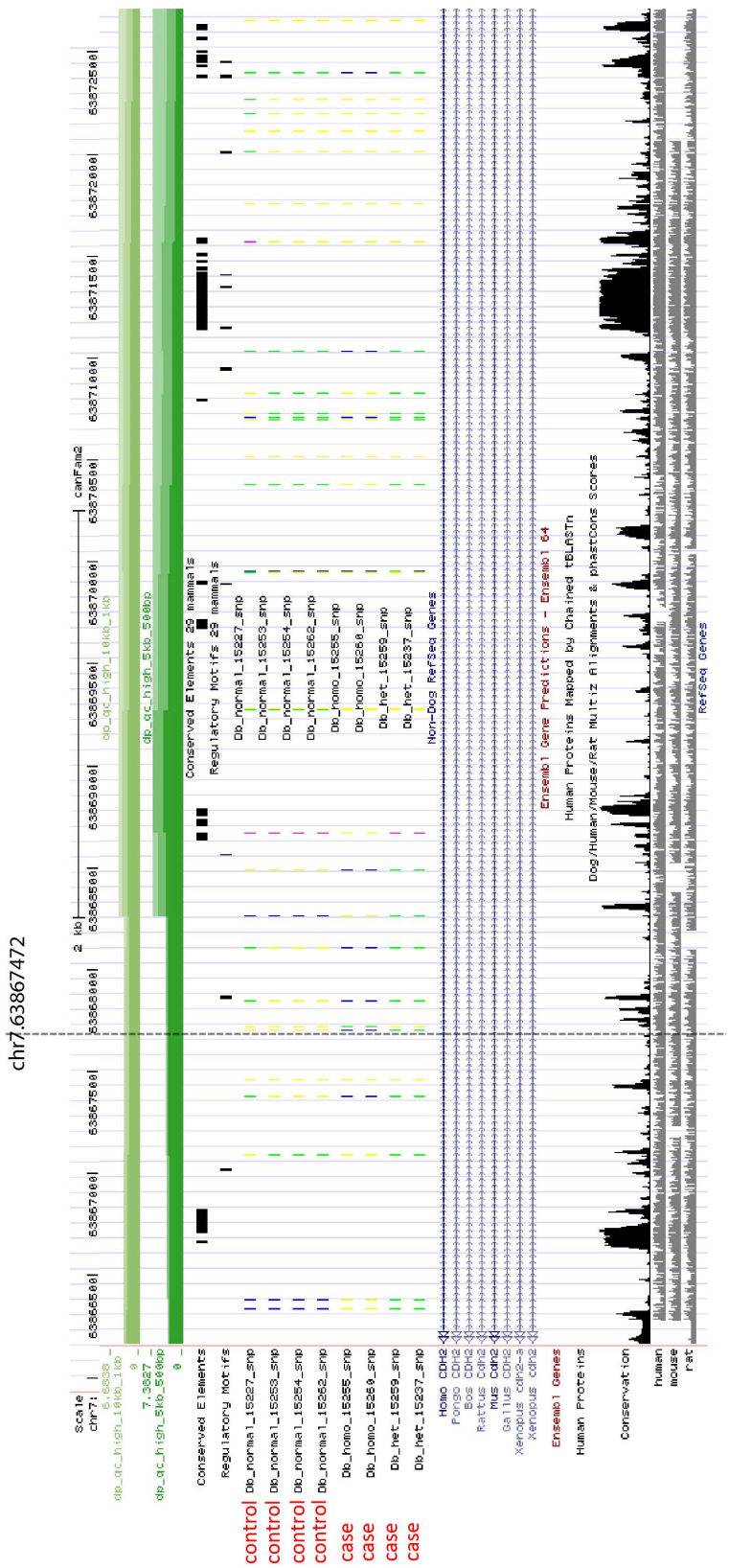


Figure 6.9: Genotypes and HLR scores within a risk haplotype in *CDH2* - See Figure 6.8

6.4.5 GenomeSTRiP reveals a number of CNV deletions

In addition to single nucleotide variants and micro-indels reported in section 6.4.2, I also detected CNVs from the sequence data using GenomeSTRiP (431), focused specifically on deletions (see Materials and Methods). I identified 173 deletions with a median length of 930bp from the 18 dogs with deep sequence coverage (Figure 6.10 and Appendix Figure 6.30). Over three quarters of the identified deletions were shorter than 2kb (Figure 6.10).

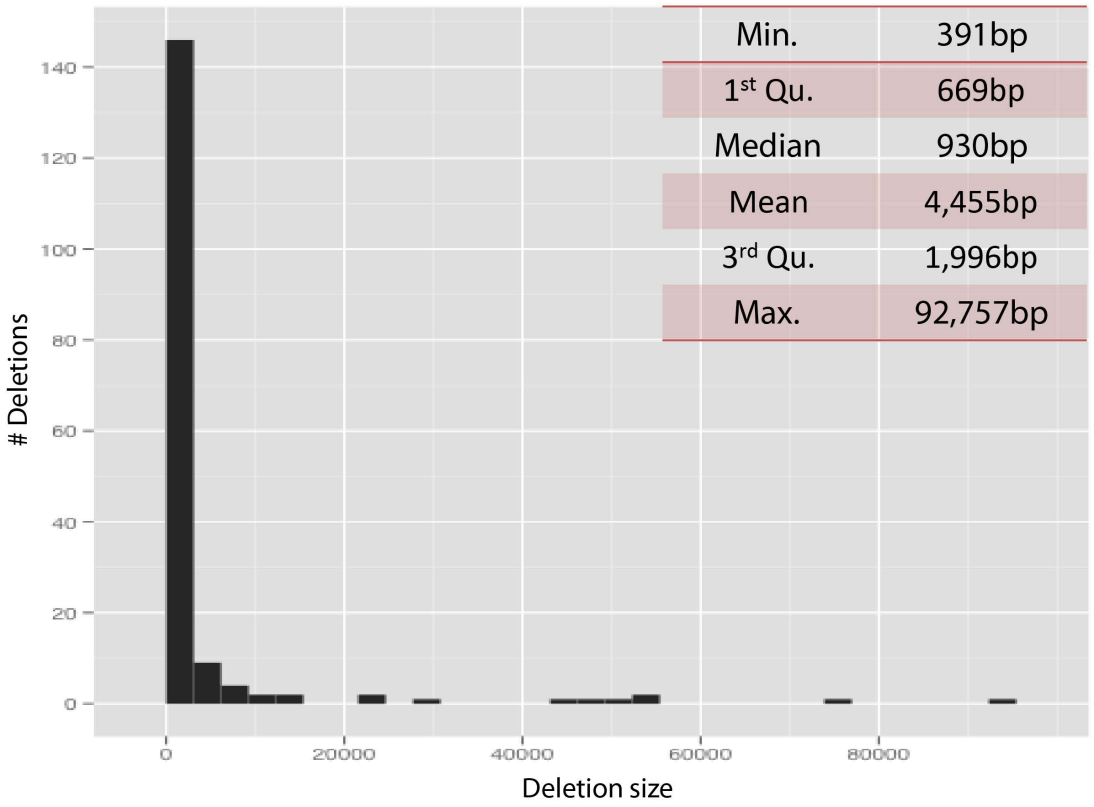


Figure 6.10: Size distribution of deletions identified among sequenced dogs - The histogram shows the number of detected deletions (y-axis) across their length in bp (x-axis). For the number of deletions detected for each dog, see Appendix Figure 6.30.

In order to identify candidate deletions that may predispose individuals to CCD, I inferred the detected deletions' copy number based on normalised sequence coverage depth using GenomeSTRiP (431), and investigated whether there were any differences in copy numbers between cases and controls. Unfortunately, the design of the sequenc-

ing experiment that was conducted in 4 batches made it difficult to distinguish whether the difference in estimated copy number between individuals come from the true copy number difference or simply reflect a strong batch effect (Appendix Figures 6.34 and 6.35). Due to this potential batch effect, I excluded deletions from candidates when the copy number difference segregates concordantly with batch membership, even if they were more common in cases than in controls. After excluding such deletions, there was only one 1.2kb deletion remaining (Chr29:44178339-44179516) that showed a case-control difference in its estimated copy number (deleted in the Jack Russell Terrier case but not in any controls; Figure 6.11).

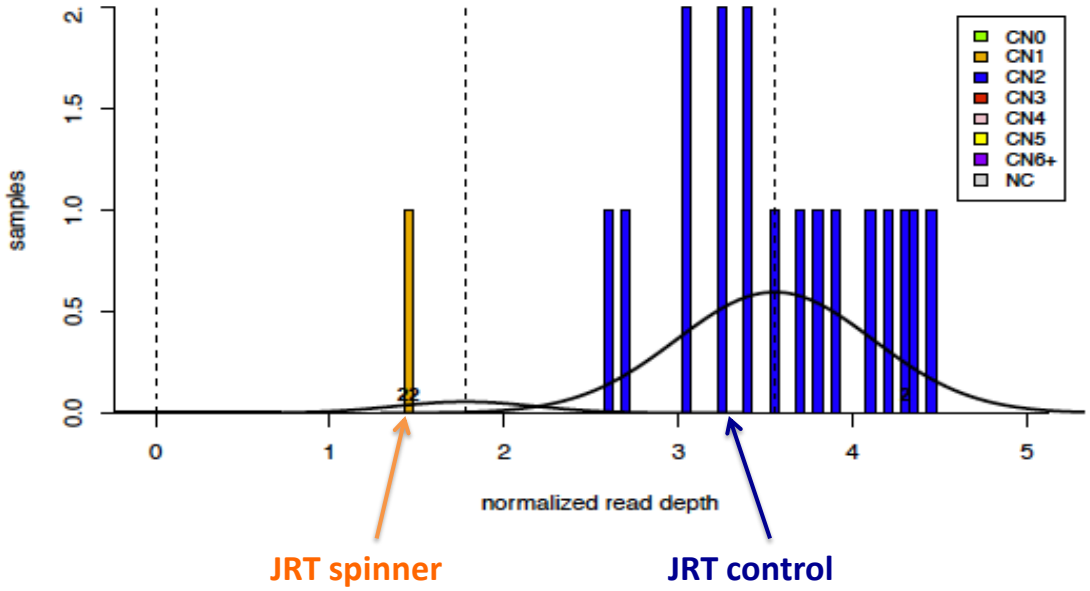


Figure 6.11: Copy number estimation of a deletion at Chr29:44178339-44179516 - Each bar represents individual dogs employed here. The x-coordinate of the bar is determined by the normalised read depth, inferred as estimated copy number of the detected deletion for each individual. The Y-axis shows the number of individual. JRT stands for Jack Russell Terrier. JRT case (with spinning phenotype, JRT spinner) and control (JRT control) dogs are indicated with orange and blue arrows, respectively. Normalized measurements of read depth (across the deleted segment) from all samples were fitted to a Gaussian mixture model (431).

In order to confirm the automatically detected deletion, I manually examined the raw alignments of reads from all 18 dogs within the genomic region surrounding the deletion. As a result, I observed a clear reduction in read depth in the 1.2kb deletion compared to the flanking regions of the affected Jack Russell Terrier's genome (Figure

6.12). The coverage plot indicates a deletion because of many abnormal reads with insert sizes that are larger than expected (Figure 6.12; abnormal reads shown in red; see Appendix Figures 6.23 and 6.24 for the relationship between insert size and deletion call). Taken together, the automatic and manual examination of the raw alignments confirmed that the 1.2kb deletion at Chr29 is present in the affected Jack Russell Terrier (Figure 6.12), but not in other control dogs.

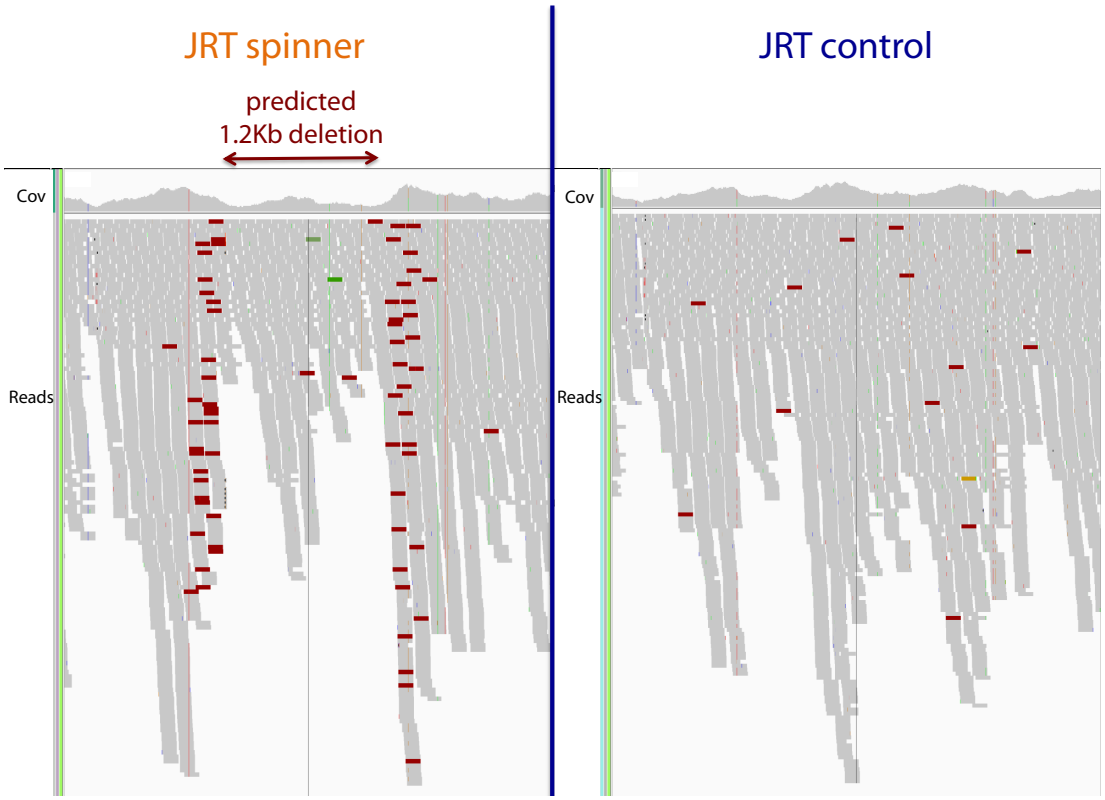


Figure 6.12: Pileup data in the region of 1.2kb deletion at Chr29 - Pileup data for the affected Jack Russell Terrier is shown on the left, while the healthy breed-matched control is shown on the right. The top track (Cov) shows the coverage of the genomic region, while the lower track (Reads) shows the individual reads that are aligned to this region. Abnormal reads that may be indicative of deletion [i.e. reads with an insert size larger than expected] are shown in red.

Importantly, the Chr29 deletion was found to wholly overlap the third exon of the *PGCP* (plasma glutamate carboxypeptidase precursor) gene, whose protein plays an important role in the hydrolysis of circulating peptides (449) (Figure 6.13). This 1.2kb length deletion, therefore, may disrupt the protein product of *PGCP*. A potential role

of this exon deletion for CCD is discussed later in this chapter (see Discussion).

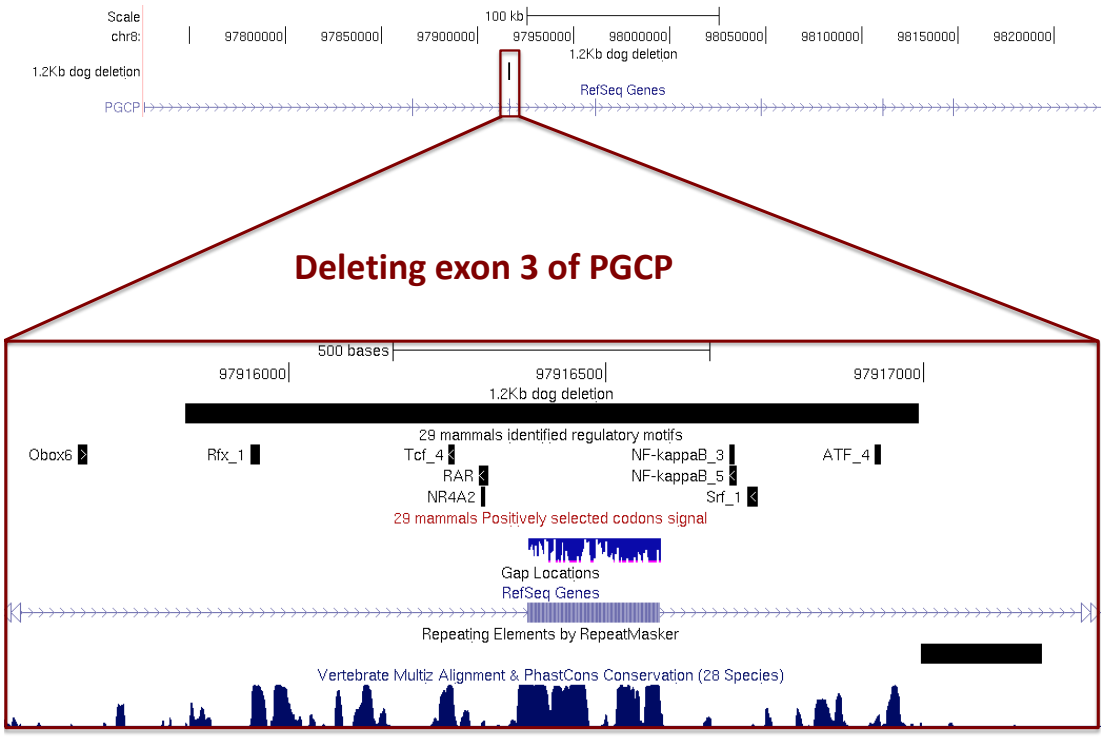


Figure 6.13: Exon deleting CNV in *PGCP* - The 1.2kb deletion at Chr29 was found in an affected Jack Russell Terrier deleting an exon of the Plasma Glutamate Carboxypeptidase Precursor (*PGCP*) gene. The canine region shown here is lifted over to hg19.

6.4.6 Canine genes with excessive case-only variants are orthologous or functionally related to human candidate genes for OCD

Besides identifying discrete candidate variants for CCD, I also sought to identify candidate genes for CCD, by considering multiple variants per gene (see section 5.4.5.2 for a similar approach in human OCD). First, I grouped case-only variants for each gene (see section 6.4.4.1). I also grouped control-only variants to control for the background number of unique variants per gene. I then counted genes based on the number of case-only (y-axis) and control-only variants (x-axis), where the identity line represents the situation when a gene does not attribute to CCD (Figure 6.14). Genes that are present above the identity line are potential candidate genes for CCD, as indicated by a relative excess of case-only variants compared to control-only variants. In contrast, genes that

are present below the identity line may be those with potentially protective variants for CCD, as indicated by a relative excess of control-only variants compared to case-only variants.

When considering all detected variants per gene, *CDH2, FNDC3B* (fibronectin type III domain containing 3B), *CTNNA2* (catenin [cadherin associated protein], alpha 2), and *PLD1* (phospholipase D1) genes were shown to have the greatest excesses of case-only variants (Figure 6.14). When considering variants within conserved elements, which may be an indication of functional genomic sequence, *CDH2* and *CTNNA2* still showed the greatest surfeit of case-only variants, while the other two were replaced by *PGCP* (plasma glutamate carboxypeptidase) and *ATXN1* (ataxin-1) (Figure 6.15). Even when corrected for gene size (by focussing on the ratio between the number of case- and control-only variants rather than the absolute difference in numbers of variants), these four genes ranked among the top genes (*CDH2*, 1st; *CTNNA2*, 2nd; *ATXN1*, 4th; *PGCP*, 10th; Appendix Table 6.23). As the affected Doberman Pinschers were selected for their risk haplotype within *CDH2*, I performed the same analysis excluding Doberman Pinschers, in order to examine whether the result that *CDH2* was the top hit was biased by that single breed. Even after excluding Doberman Pinschers, *CDH2* still remained one of the top hits in both cases of all genic variants and genic variants present in conserved region (Appendix Figure 6.36 and 6.37). Again when adjusted for gene size after excluding Doberman Pinschers, *CDH2* remained as the top gene for excessive case-only variants (Appendix Table 6.23). For the top genes for each breed, see Appendix Figures 6.38-6.45.

Subsequently, I sought to identify genes that were found to be interesting candidates for both OCD and CCD. I compared the canine genes presented with a surfeit of case-only variants (i.e. genes above the identity line in Figures 6.14 and 6.15, and Appendix Figures 6.36-6.45) with the 123 investigative genes that were identified as being associated with human OCD in section 5.4.5.2. *CTNNA2, FER* and *c5orf13* orthologues were identified in both humans (p_{perm} = 0.04, 0.03, and 0.025, respectively) and dogs (Figures 6.14 and 6.15, and Appendix Figures 6.36-6.45). Moreover, *CDH13*, a member of cadherin family, that also includes a CCD GWA-gene *CDH2*, was associated with human cases (p_{perm} = 0.001). *CTTNBP2*, whose protein product binds to cortactin (cortical actin binding protein), is a cell adhesion molecule similar to cadherin and was also associated with human OCD (p_{perm} = 0.018) (see Discussion).

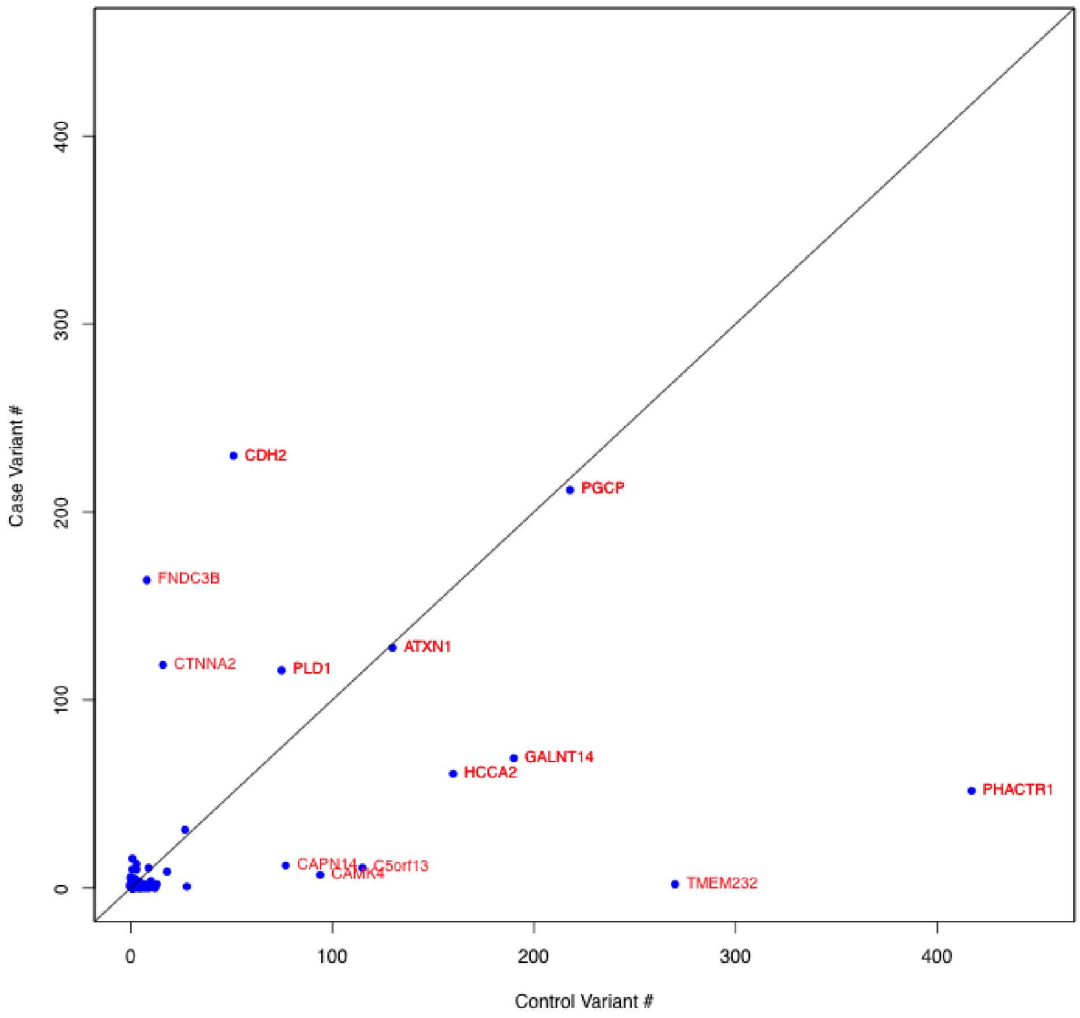


Figure 6.14: Gene plot based on the number of genic case- and control-only variants across all four breeds - The x-axis shows the number of control-only variants and the y-axis shows the number of case-only variants. Each blue dot represents a resequenced. Blue dots above the x=y line indicates genes that possess more case-only than control-only variants.

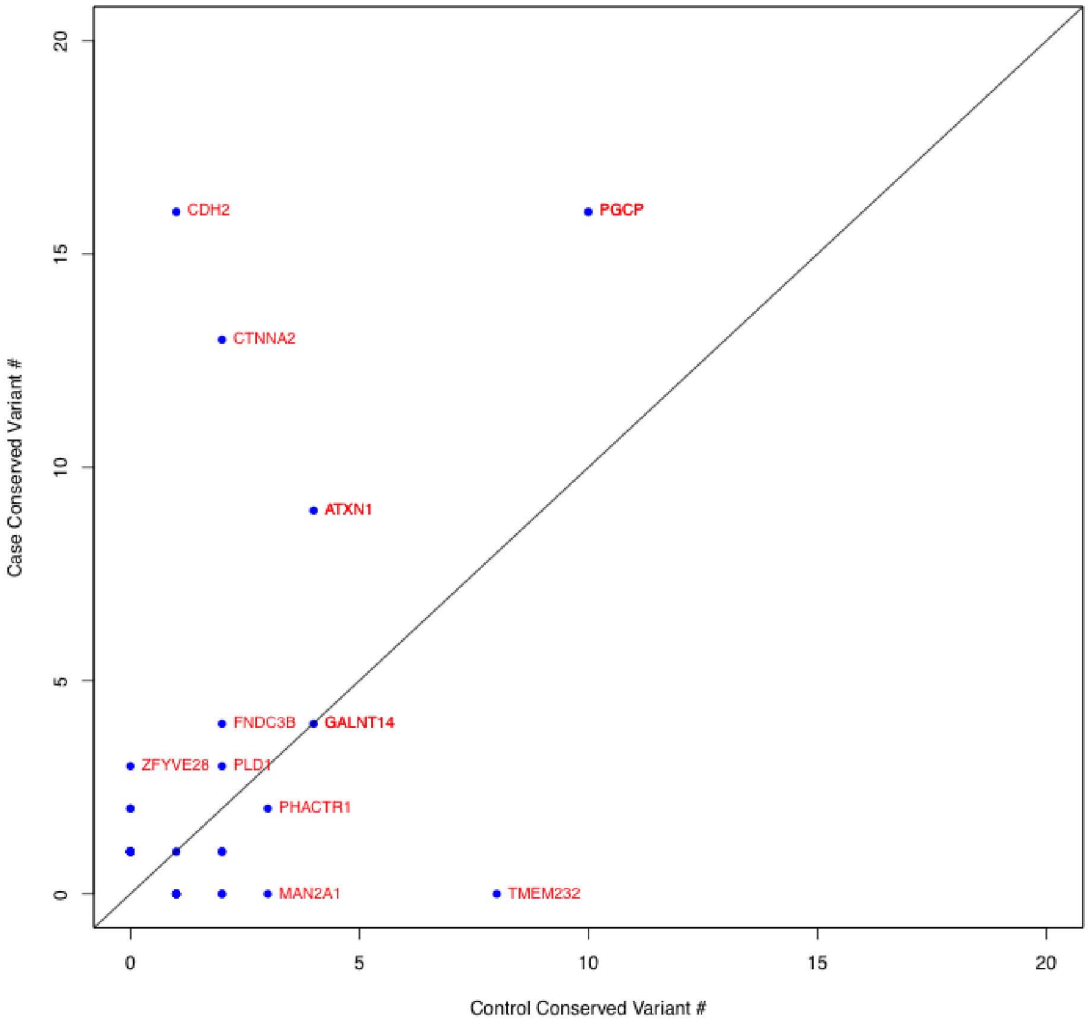


Figure 6.15: Gene plot based on the number of genic case- and control-only variants within conserved regions - See Figure 6.14

6.5 Discussion

6.5.1 Genetic structure of breeds may have affected number of identified variants

42% (966/2,291) of case-only variants were found in the Jack Russell Terrier case (JR_spinner_3094 in Table 6.7; see section 6.4.4.1), while other case dogs possess many fewer case-only variants for each dog (80-437; All Uni Var # in Table 6.7). This excess of unique variants in the Jack Russell Terrier case is likely to reflect many private variants that contribute to this breed's high genetic heterogeneity (0.73 compared to 0.38-0.43 among the other three breeds; Heterozygosity in Table 6.8). The genetic heterogeneity of the Jack Russell Terrier is widely accepted - the breed is considered to be among the least inbred dogs for those entered into the Kennel Club Registry (450). Close inbreeding among Jack Russell Terriers has been prohibited for its mental and physical protection, thus alleles of the domestic dog population are likely to introgress this breed more than for other modern breeds. The notion that case-only variants from the Jack Russell Terrier case include many private variants is supported by the fact that the control Jack Russell Terrier dog was also found to possess a large proportion (39% [1,204/3,116]) of all control-only variants.

For other breeds, affected Doberman Pinschers have 80-426 case-only variants, while control dogs have 156-335 control-only variants (row headings pre-fixed with *Db_* in Table 6.7); German Shepherd cases have 383-422 case-only variants per individual while healthy German Shepherds have 130-204 control-only variants (row headings pre-fixed with *GS_* in Table 6.7); and the Shetland Sheepdog case has 437 case-only variants while the control dog has 1,476 control-only variants. The unusually high number of control-only variants in the Shetland Sheepdog is probably due to the relatively low-quality variant calls rather than true variations, considering that the concordance of variant calls between the sequencing data and the SNP array data was lower in the control Shetland Sheepdog (93.7%; *Shetland_normal_1737* row of *Concordance (%)* column in Table 6.7) than it was in other dogs (97.2-99.8%; all other rows of *Concordance (%)* column in Table 6.7), particularly for heterozygous positions (79.84% compared to 91-100% in other dogs; #Het positions column in Appendix Table 6.21).

In addition to heterozygosity within a breed, the genetic relationship between the breeds appears to influence the multi-breed group analysis for case-only variants. The

unusually large number of case-only variants detected in the SS+JRT group (672 variants), compared to the other two groups with Jack Russell Terrier (53 and 81 variants for GS+JRT and DP+JRT, respectively) might be influenced by the genetic relationship between the dog breeds (446) (for an interpretation in terms of within breed's heterozygosity, see section 6.4.4.2).

Parker *et al.* identified four genetic clusters among 85 domestic dog breeds and these predominantly contained breeds with similar geographic origin, morphology, or role in human activities (446). According to their clustering analysis, each Doberman Pinscher and German Shepherd showed the highest membership coefficient for two different clusters of dog breeds (0.686 in the Red cluster for Doberman Pinschers and 0.858 in the Blue cluster for German Shepherds; Table 6.13), in which Terriers also showed the strongest genetic clustering coefficients (0.921 in Blue and 0.844 in Red cluster; Table 6.13). On the other hand, Shetland Sheepdog, the breed whose combination with Jack Russell terrier yielded an unusually large number of case-only variants, was dominantly clustered in Green (clustering coefficient 0.684; Table 6.13), in which Terriers showed little association (coefficients 0.02-0.303; Table 6.13).

Breed	Yellow	Blue	Green	Red
DP	0.015	0.103	0.194	0.688
GS	0.006	0.858	0.090	0.046
SS	0.017	0.105	0.684	0.193
Terriers	0.009-0.368	0.079-0.921	0.02-0.303	0.047-0.844

Table 6.13: Average membership coefficient of breeds to genetic clusters inferred from 85 domestic dog breeds - Extracted from supplementary Table 2 of (446). Yellow, blue, green, and red indicate the four genetic clusters inferred from the microsatellite genotypes of 85 domestic dog breeds. As the clustering coefficient of Jack Russell Terrier was not reported, the range of membership coefficients from 12 terriers examined in this study are shown in this table. The 12 terriers include Tibetan Terrier, Manchester Terrier (Toy), Australian Terrier, West Highland White Terrier, Airedale Terrier, American Hairless Rat Terrier, Cairn Terrier, Bedlington Terrier, Kerry Blue Terrier, Irish Terrier, Soft Coated Wheaten Terrier, and Miniature Bull Terrier.

In my analysis, the case-only variants counted for the combinations of breeds (Tables 6.8 and 6.9) are probably the mixture of breed-specific variants, as well as individual-specific and true case-specific variants. When the genetic structures among multiple breeds are similar, there would be fewer breed-specific variants than when the breeds are genetically distant. Accordingly, it is likely that when the genetic structures are similar

between a pair of breeds examined here, the reduced number of breed-specific variants was reflected in the observed case-only variants, hence resulting in fewer case-only variants in the groups of genetically close breeds (81 and 53 variants in the DP+JRT and the GS+JRT, respectively) than in the genetically distinct breed group (672 variants in the SS+JRT).

This notion of an influence of the genetic similarities/differences between breeds in the number of case-only variants also holds for other breed combinations. For instance, considering that the DP+GS group has a larger sample size (6 cases and 6 controls) than the DP+SS group (5 cases and 5 controls), one would expect that case-control comparison of the DP+GS group may yield fewer case-only variants, because more breed-specific variants should be filtered out as more breed-matched case-control pairs are included in the comparison (see above). Nevertheless, 19 case-only variants were identified in the larger DP+GS group while only 8 case-only variants were seen in the smaller DP+SS group. This might again be explained by the fact that the genetic structures of both Doberman Pinscher and Shetland Sheepdog can be mostly accounted for by Green and Red clusters (DP, 88.2%; SS, 87.7%; sum of Green and Red columns for each breed in Table 6.13), while the genetic structure of German Shepherd can be accounted for only 13.6% by those two clusters (Table 6.13).

6.5.2 Chr7:61693835 and its flanking region may have regulatory role for CCD

Within a target region determined from the GWAS for CCD on Doberman Pinschers (the best SNP at Chr7:61865715, $p = 1.58 \times 10^{-5}$; Appendix Table 6.16), I identified two SNPs from resequencing that have notable case-control differences in Doberman Pinschers: Chr7:61693835 $T \rightarrow C$ variant was detected in 75% (3/4) of the affected Doberman Pinschers, while a variant that is 20bp apart, at Chr7:61693855 $A \rightarrow T$, was detected in the other affected Doberman Pinscher (Figure 6.16). These two variants were not present in any of the control dogs used here.

Multiple lines of bioinformatical evidence suggests a strong regulatory activity in the flanking region of these two variants. Strong histone marks are present in this region, indicating its potential role for promoter- or enhancer-like activity (*Layered H3K4Me1* and *Layered H3K27Ac* tracks in Figure 6.16). The chromatin state that was computationally inferred from ChIP-seq data also suggested a strong enhancer

activity for this region (the bottom track in orange in Figure 6.16). Furthermore, this region is evolutionarily extremely conserved. Of 43 vertebrates with alignment information in this region, 43/43 of these vertebrates possess the reference alleles at the positions of both variants, showing the maximum possible conservation of these two genomic positions (T and A alleles for Chr7:61693835 and Chr7:61693855, respectively; positions marked with red arrows in Figure 6.16).

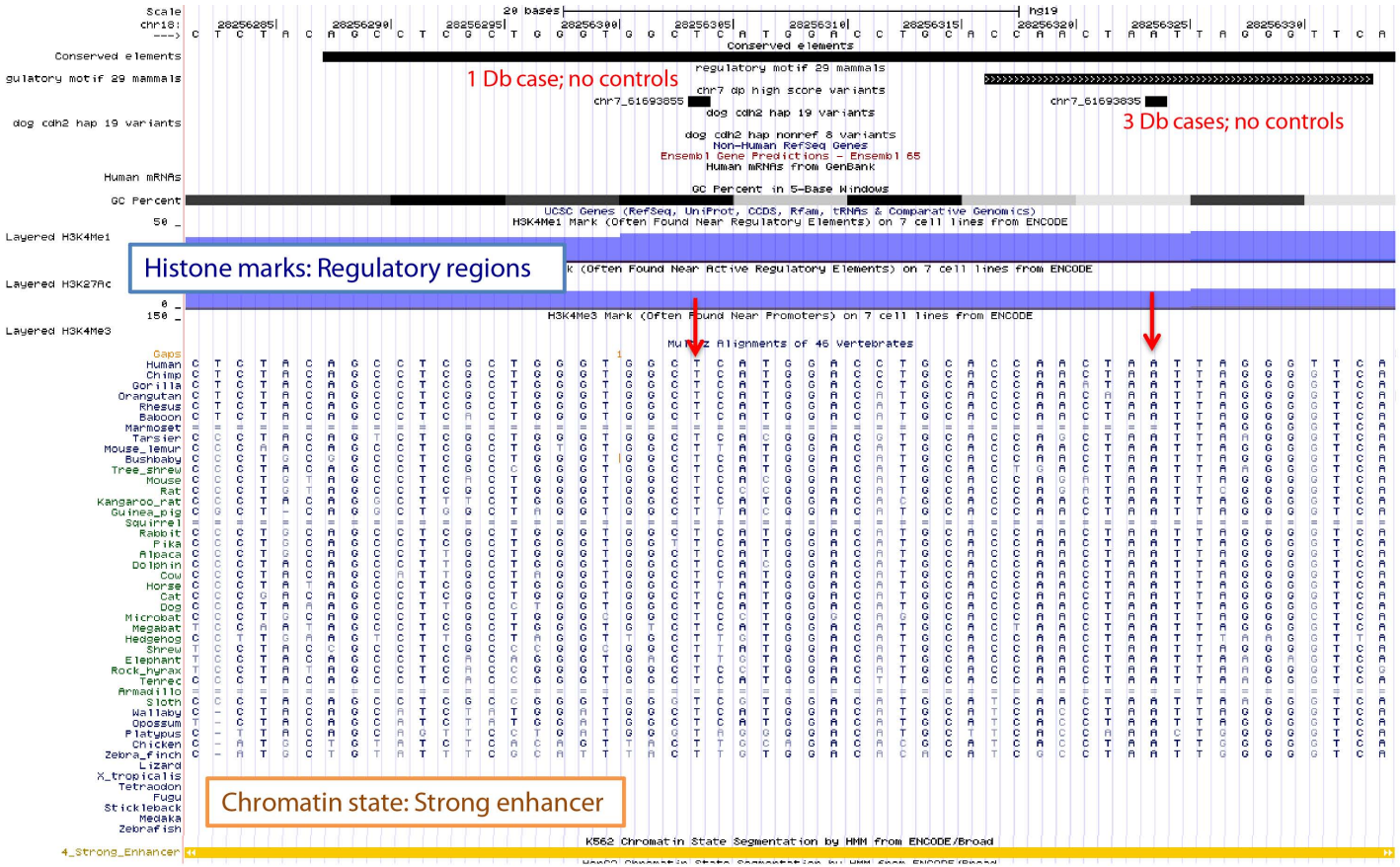


Figure 6.16: Variants *Chr7:61693835* $T \rightarrow C$ and *Chr7:61693855* $A \rightarrow T$, and their neighbouring regions lifted on to the human genome hg19 - The top two tracks show conserved elements and the regulatory motifs inferred from 29 mammals (3). The third track shows the two variants at Chr7:61693835 $T \rightarrow C$ and Chr7:61693855 $A \rightarrow T$. Note that the original canine sequence was inverted when lifted on to the human genome. The tracks coloured with purple show the experimentally proven histone marks that may be an indication of regulatory regions (5). Next, the alignments of 46 vertebrates are shown for this region, with the two variants' positions indicated with red arrows (397)(398). Note that the two identified variants were reverse complementary to the alignments shown here. The bottom track with orange colour shows the chromatin state segmentation for the human cell type K562 (blood). A common set of the states across the cell types were learned by computationally integrating ChIP-seq data using a hidden Markov Model (HMM). In total, fifteen states were used to segment the genome, and these states were then grouped and coloured to highlight the predicted functional elements (451)(452).

Given such strong bioinformatical evidence for regulatory activity of this region (Figure 6.16), I also examined whether the regions' sequence would be mapped to any known transcription factor binding sites. Using TRANSFAC® (129), I identified three protein-binding DNA sequence motifs that are highly similar to the flanking (±10bp) sequence of the variant *Chr7:61693855* $A \rightarrow T$ (matrix similarity>0.86 and core similarity>0.96). The identified motifs are, namely, binding sites for Paired mesoderm homeobox protein 2 (PRRX2), POU domain class 2 transcription factor 1 (POU2F1), and newborn ovary homeobox protein (NOBOX) (Figure 6.17). In addition, the consensus sequence of these three motifs assigns the highest positional weight (see Appendix text 6.6.7) to the reference A allele at position Chr7:61693855 (indicated with red arrow in Figure 6.17), suggesting that this position is likely to serve as an important base for transcription factor-DNA binding.

All three transcription factors appear to have the potential to play relevant roles in neurodevelopment: PRRX2 is a putative RNA polymerase II transcription factor that plays a role in embryonic development; POU2F1 also mediates embryonic development and has been linked to Alzheimer's disease, for which OCD had previously been suggested as a potential risk factor (453); and NOBOX may be involved in nervous system development (129).

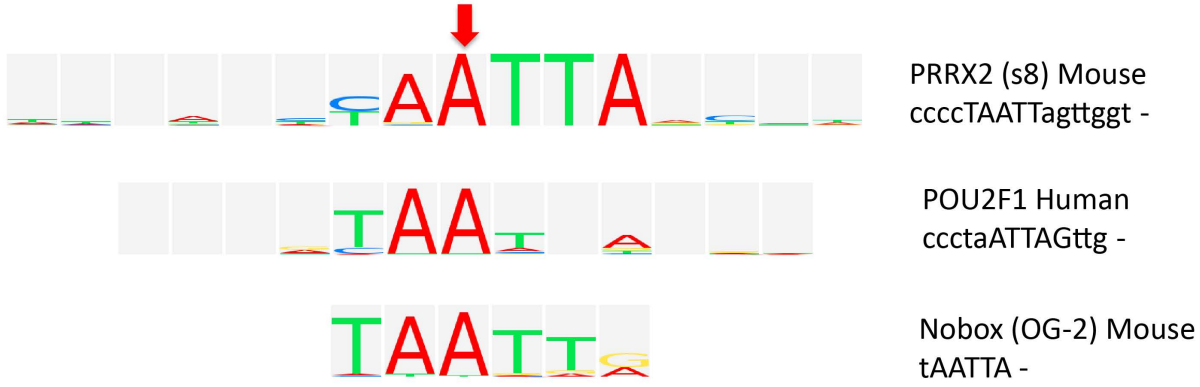


Figure 6.17: Transcription factor binding sites mapped to Chr7:61693835 region by TRANSFAC® - The variant's position is indicated with a red arrow. On the right, the transcription factors that potentially bind to the sequence and the organism where the motif was found are shown. Below this, the consensus sequence are displayed in full, with a core sequence shown in capitals. For the methodology used to find the motifs, see Appendix text 6.6.7.

Considering the strong regulatory marks (Figure 6.16) and highly matched protein-DNA binding motifs in this region (Figure 6.17), Chr7:61693855 and its neighbouring DNA sequence may act as an enhancer (i.e. a short stretch of DNA that can be bound by trans-acting factors such as transcription factors) to regulate a gene's transcription (454).

An enhancer may be *cis*-acting, located upstream or downstream of the gene it regulates. However, an enhancer does not need to be in proximity to the regulated gene in regard to the number of nucleotides. In fact, enhancers can be located several hundred thousand base pairs away from the transcription start site, or even on a different chromosome (455).

Chr7:61693855 and its flanking sequence lie in a 'gene desert' region, from which the transcription start site of the nearest gene, *DSC3*, is ~292kb away (Figure 6.18). *DSC3* gene encodes Desmocollin-3, a member of the cadherin superfamily (456), of which the protein product of *CDH2*, the only gene shown to have genome-wide association with CCD, is also a member. Desmocollin-3 is found primarily in epithelial cells where it constitutes adhesive proteins of the desmosome cell-cell junction together with other desmosomal family members and desmogleins, playing an essential role in cell adhesion and desmosome formation (456). Desmosomal proteins are an integral part of the maintenance of tissue architecture and the loss of these components leads to a lack of adhesion and a gain of cellular mobility (457).

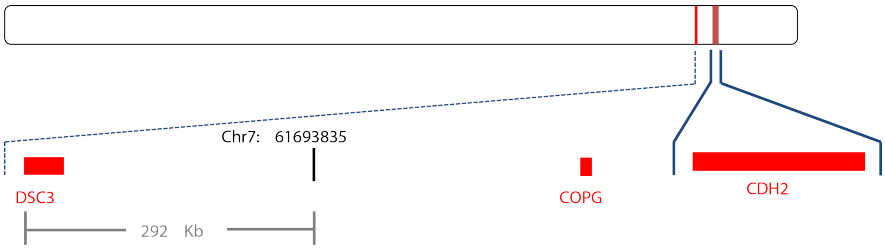


Figure 6.18: Two CCD candidate regions in canine chromosome 7 - Two regions in canine chromosome 7 were targeted for resequencing based on the previous GWAS (Appendix Table 6.16). In the Chr7:61.6Mb region, the Chr7:61693855 $A \rightarrow T$ variant is distantly flanked by *DSC3* and a pseudogene *COPG*. In the Chr7:63.7Mb GWA-interval, there is only one protein-coding gene, *CDH2* (Appendix Figure 6.22 and Appendix Table 6.16).

Taken together, one may hypothesise that any [or any combinations] of the transcription factors PRRX2, POU2F1, and NOBOX have a potential to play a role

in a mechanism related to the development of CCD. More specifically, the variant Chr7:61693855 $A \rightarrow T$ within the binding sites for these transcription factors may affect transcriptional activities, possibly of the neighbouring gene, a cadherin *DSC3*, among CCD affecteds. This hypothesis is experimentally tractable. The DNA-protein binding of this region can be confirmed by an EMSA. Then the DNA-interacting transcription factors may be identified by *supershift*, using an antibody that recognizes PRRX2, POU2F1, or NOBOX specifically. Expression changes of the genes that are affected by this variant may then be investigated using standard mRNA quantification methods (e.g. RT-qPCR). In this case, the closest and functionally relevant *DSC3* gene may be tested. However, it should be noted that such a gene regulation is likely to be cell-/tissue-type specific, thus testing on a relevant cell-type is important (see section 5.5.2.4 for details of these experiments).

6.5.3 A 1.2kb deletion of an exon in *PGCP* may play a role in CCD

As the 1.2kb deletion that was found only in the affected Jack Russell Terrier wholly overlaps with exon 3 of *PGCP*, this deletion may affect the protein product of *PGCP* (Figure 6.13), although there is a possibility that exon 3 is spliced out in some transcripts due to alternative splicing. While the deletion was private to a single affected dog, the GWAS for CCD with almost a hundred affected Doberman Pinschers and over 70 matched controls showed a strong association of SNPs in proximity to this deletion (Figure 6.19). This suggests that a disruption of the genomic region of the *PGCP* exon 3 may be commonly underlying CCD.

The protein product of *PGCP* is largely uncharacterised, but it is suggested to play an important role in the hydrolysis of circulating peptides (449), particularly glutamate, the principal excitatory neurotransmitter in the vertebrate nervous system (458). Glutamatargic synapses are the key component of the CSTC circuit of the brain that are responsible for repetitive, compulsive behaviours (see section 5.2.1.2). Indeed, PGCP is highly expressed in the thalamus, a compartment of the CSTC brain circuit (Figure 6.20). Considering the relevance of PGCP's carboxypeptidase activity and its expression profile in brain tissues, it may be worthwhile to examine a role of PGCP in CCD, particularly the disruption of the gene's exon 3.

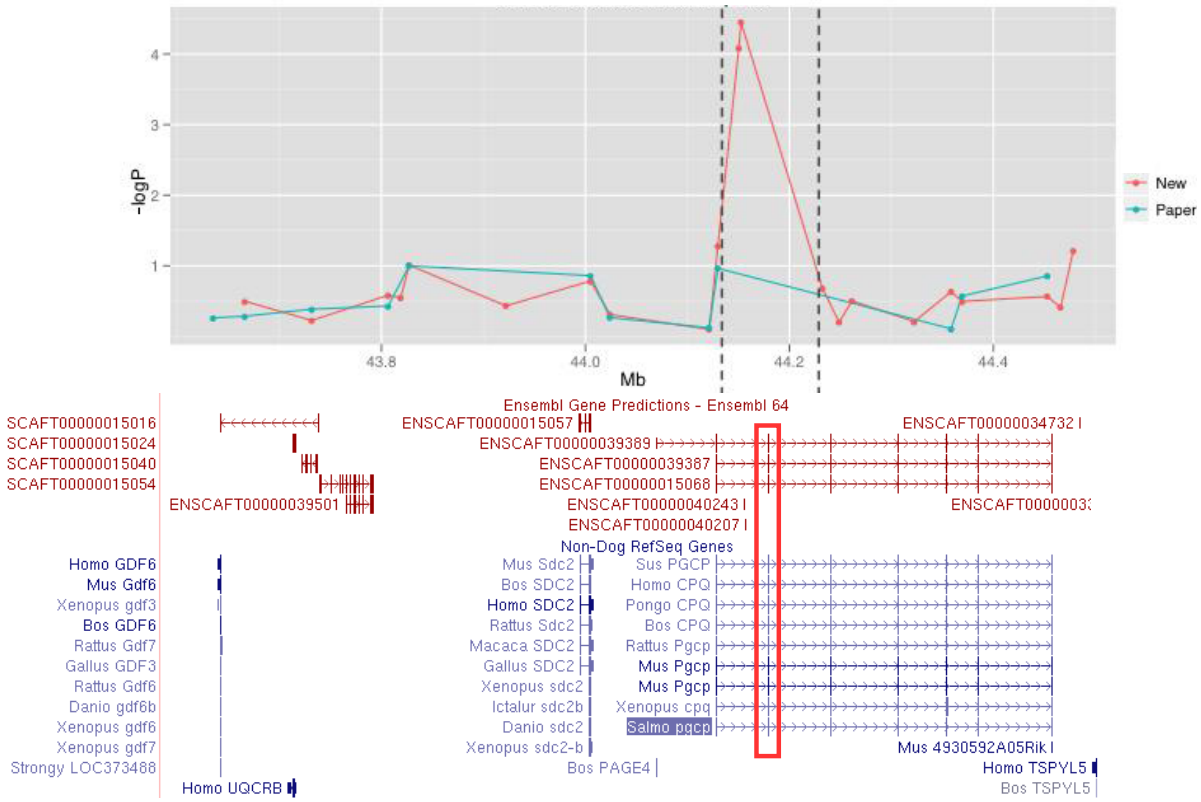


Figure 6.19: Association signal with CCD in *PGCP* region - The line graphs show -logP values for each SNP position. The red and green lines represent the CCD GWA-signals from the reference (357). Green (shown as Paper) represents the data as published in (357), and red (shown as New) represents the data in (357) that were reanalysed (see Appendix Figure 6.22 and Appendix Table 6.16). An associated genomic interval is indicated with dashed line. The lower tracks show gene transcripts including *PGCP* in this genomic region. The exon affected by a 1.2kb deletion is indicated with red box.

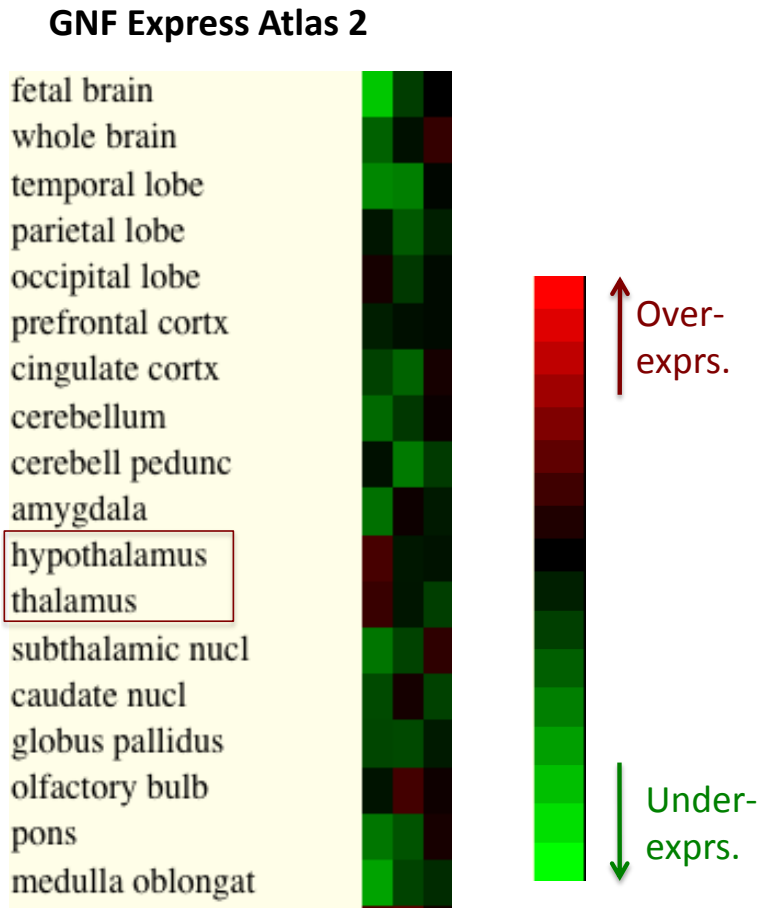


Figure 6.20: Expression profile of PGCP in the brain - This graph shows the expression level of PGCP in human brain tissues from the GNF Gene Expression Atlas 2 (459). Red indicates overexpression in the tissue, and green indicates underexpression.

6.5.4 Candidate genes may play a role in synaptic formation, and may function in concert

As shown in section 6.4.6, *PGCP, CDH2, CTNNA2*, and *ATXN1* possessed a surfeit of case unique variants within evolutionarily conserved elements (Figure 6.15), and hence were prioritised as candidate genes for CCD. Here, I will discuss the population or breed patterns of these genes' case-only variants in more detail, and their functions in the context of CCD.

PGCP is a carboxypeptidase, in which an exon deletion was found in a Jack Russell Terrier case (see section 6.5.3). *PGCP* was also shown to possess excessive case-only variants in conserved elements across all four breeds collectively (see section 6.4.6; Figure 6.15), as well as in each German Shepherd, Jack Russell Terrier, and Shetland Sheepdog breed separately (Appendix Figures 6.41, 6.43, and 6.45). In addition, *PGCP*, together with *CDH2*, are the genes that have most case-only variants in absolute number (16 variants) that are within conserved regions. This may suggest that not only the private *PGCP* exon deletion, but also other functional mutations within the gene may be commonly affected among CCD individuals. Given the PGCP's role in hydrolysing glutamate, *PGCP* appears an excellent candidate gene for CCD (see section 6.5.3).

CDH2 was previously associated with CCD in Doberman Pinschers (357). Importantly, even when I excluded Doberman Pinschers, *CDH2* still showed a substantial surfeit of case unique variants across the remaining breeds collectively (i.e. German Shepherd, Shetland Sheepdog, and Jack Russell Terrier; Appendix Figure 6.37), as well as in each German Shepherd and Jack Russell Terrier separately (Appendix Figures 6.41 and 6.43). This again suggests that a disruption of *CDH2* is a common mechanism of mutation involved in CCD across multiple breeds, rather than a breed-specific CCD predisposing variant.

CTNNA2 encodes a cadherin-associated protein, α-N-catenin. An excess of case-only variants within conserved regions was found for *CTNNA2* across all four breeds collectively (see section 6.4.6; Figure 6.15), and in German Shepherds, and Jack Russell Terrier independently (Appendix Figures 6.41 and 6.43). Among the case-only variants for this gene, 13 were found to be within conserved regions, which may have a potential to exert functional consequence (Figure 6.15).

ATXN1 has been associated with spinocerebellar ataxia type 1 (460). Case-only variants in conserved regions were again excessively present in *ATXN1* across all the

breeds employed here collectively (see section 6.4.6; Figure 6.15), and in German Shepherds (Appendix Figure 6.41) or Jack Russell Terriers (Appendix Figure 6.43). Among the case-only variants for *ATXN1*, 9 were found to lie within conserved regions, which also have a potential to have functional outcome.

Notably, the protein products of *CDH2* (N-cadherin), *CTNNA2* (α-N-catenin), and *ATXN1* (ataxin-1) are known to form a protein complex that function in concert: A role of cadherin-catenin complexes as key regulators for cell adhesion has been well documented (461). In particular, recent work has revealed an essential role for these complexes in synaptic formation and function (461). N-cadherin, together with other type I cadherins such as E-cadherin, mediate calcium-dependent adhesion in cells (462)(463). They adhere the cells expressed with cadherins to other cells that also expressed with cadherins via homophilic or heterophilic cadherin-based interactions (464). N-/E-cadherin directly bind to β-catenin, which is indirectly linked to the actin cytoskeleton via α-N-catenin at synapses (465) (Figure 6.21).

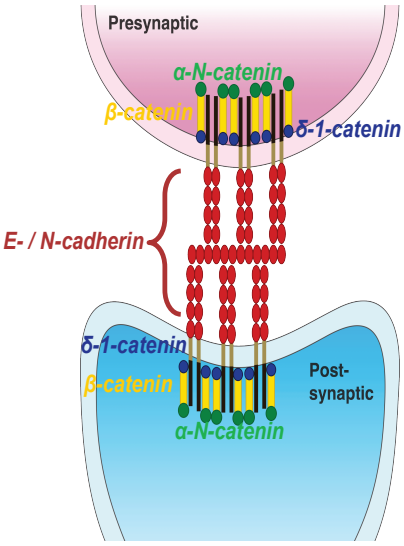


Figure 6.21: A schematic cartoon of the cadherin-catenin complex in synapse assembly - δ-1-catenin increases the time spent by cadherin complexes at the membrane. β-catenin provides the link between N- or E-cadherin and α-N-catenin that affects the pre-synaptic assembly as well as dendritic spine formation and motility at the post-synaptic density.

In addition to adhesion, N-cadherin may play an important role in the assembly of diverse synapses. N-cadherin is expressed in all synaptic contacts in the early steps of

synapse formation, while it specifically accumulates at excitatory synapses, as synapses mature (466). Furthermore, several studies have suggested that the link of the N-cadherin complex to the actin cytoskeleton via α-N-catenin is crucial in dendritic spines, and perturbations in the spine morphology affect the assembly of the pre- and post-synaptic junctions (467)(468)(469).

While the function of ataxin-1 is largely unknown, it has been demonstrated that ataxin-1 occupies the promoter region of E-cadherin in vivo and that ataxin-1 activates the promoter in a CtBP2 (C-terminal-binding protein 2; a transcriptional corepressor)-mediated transcriptional regulation manner (470). Also, it has been shown that the expression of E-cadherin and α-N-catenin is strictly regulated spatiotemporally, specifically being expressed in the fore-brain region of mouse embryonic brain (471). Taken together, mutated ataxin-1 may dysregulate the level of E-cadherin, which in turn, affect synaptic formation and development of the fore-brain region. Lastly, in addition to direct cadherin-mediated adhesion (Figure 6.21), β-catenin could also function via *Wnt* signalling pathway, which regulates pre-synaptic differentiation (472).

In summary, scientific literatures have shown that the top CCD candidate genes (*CDH2, CTNNA2*, and *AXTN1*) play important roles in synaptic formation, particularly related to cell adhesions and scaffolding. This may suggest that such biological processes are commonly affected among individuals with CCD.

6.5.5 Perturbed synaptic formation may be a common mechanism for compulsive disorder in both humans and dogs

Despite the less stringent criterion (i.e. simply having a surfeit of case-only variants) for candidate gene selection in CCD study (see section 6.4.6), the separate identifications of orthologues as being candidate genes for each OCD and CCD (i.e. *CTNNA2, FER*, and *c5orf13*; Table 6.14) may be worthwhile to be discussed. These orthologues might be commonly disrupted among affecteds, which in turn contribute to compulsive phenotypes even in different species. In fact, these OCD-/CCD-associated orthologues appear functionally convergent. In addition to the well known role of the cadherin-catenin complex [encoded by *CDH2* and *CTNNA2*] in synaptic formation (Figure 6.21; see section 6.5.4), FER, a member of the FPS/FES family of nontransmembrane receptor tyrosine kinases, regulates cell-cell adhesion mediated by cadherin (473), by interacting with δ-catenin to enhance the lifetime of the cadherin complexes at the membrane (474). In

terms of *c5orf3*, although its function is currently not well understood, it may promote axonal regeneration and regulate neuronal differentiation (475).

Human	Dog
CTNNA2, CTTNBP2	*CTNNA2*
FER	*FER*
c5orf13	*c5orf13*
CDH13	*CDH2*

Table 6.14: Candidate genes identified for compulsive disorder in both human and dog - Candidate genes that are either orthologous or functionally related are shown in the same row.

While several orthologues with multiple variants were identified in each OCD and CCD, given the genetic complexity of the disorder, it is likely that separate mutations that are involved in common biological functions exert disease symptoms. Thus, if shared genetic aetiologies between OCD and CCD were to be identified, genes with functional similarities or relationships should be considered, as well as their orthologues. Subsequently, I found that functions of human OCD candidates *CTTNBP2* and *CDH13* (see section 5.4.5.2) are highly related to those of two canine candidate genes, *CTNNA2* and *CDH2*, respectively (Table 6.14; see section 6.4.6).

The protein product of *CTTNBP2* (i.e. cortactin-binding protein 2) interacts with cortactin, which is a central regulator of the actin cytoskeleton (476). Cortactin is known to interact with δ-catenin, a binding partner of N-cadherin (477) (Figure 6.21). A role for δ-catenin in neuronal morphogenesis has been reported previously (478), and given that N-cadherin plays a crucial role in neuronal morphology, it is likely that the protein product of *CTTNBP2* acts on neuronal morphogenesis via δ-catenin and N-cadherin. As discussed in section 6.5.4, *CTNNA2* encodes α-N-catenin that constitutes the synapse-forming cadherin-catenin complex with N-cadherin and δ-catenin (Figure 6.21), through which CTTNBP2 may act.

CDH2 and *CDH13* were identified as candidate genes in CCD and OCD, respectively, and both are members of the cadherin family and are involved in calcium-dependent cell-cell adhesion (479). T-cadherin (encoded by *CDH13*) is highly expressed in the central nervous system, and in a known inhibitor of neurite outgrowth (480). In addition, *CDH13* is associated with neurological symptoms that are frequently co-morbid with compulsive disorder in humans, such as autism and ADHD (481).

The identification of these candidate orthologues from OCD and CCD studies were largely independent. Target regions for CCD were solely determined by either CCD-GWAS or selective sweep signals (see section 6.3.2), while human OCD target regions were selected mostly based on the CSTC brain circuit (see section 5.2.1.2). Specifically, human candidate genes *CTNNA2*, *CTTNBP2*, and *CDH13* were targeted based on the fact that their protein products are highly expressed in the post-synaptic density. Athough the candidate orthologues *FER* and *c5orf13* were included in both OCD and CCD studies based on the previous CCD-GWAS or canine selective sweep signals, simply resequencing these orthologues in two separate studies do not lead to their associations with each case cohort, given that hundreds of genes were targeted and tested for each OCD and CCD study.

In summary, the orthologues (i.e. *CTNNA2, FER*, and *c5orf13*) and functionally related genes (*CDH2, CDH13*, and *CTTNBP2*) identified from the OCD and CCD studies herein, appear to function in concert, to regulate synaptic formation and neuronal morphology. Importantly, similar biological processes were indeed identified via functional enrichment analysis of human OCD sequence data, such as regulation of neurogenesis, neuron projection morphogenesis, and cell adhesion (see section 5.5.6).

6.5.6 Conclusions

Resequencing and analysis of canine regions showing evidence for associations with CCD or selective sweep signals for CCD-predisposed breeds among 12 affected dogs and 12 breed-matched healthy controls allowed the identification of several candidate variants and genes for CCD. These candidate variants and genes will need to be validated in a larger data set and followed-up by molecular and/or animal model experiments. Notably, the top candidate genes from the CCD study functionally overlap with the genes that were associated with human OCD (Chapter 5). Furthermore, the orthologues and related genes identified in both species are known to function together in synaptic formation and neuronal morphogenesis. Indeed, similar biological processes such as neurogenesis, synaptic formation, and cell adhesion were identified as being highly enriched among the human genes associated with OCD. Such functional convergence of the candidate genes/pathways between OCD and CCD suggest that these common biological processes may underlie compulsive disorder in both humans and dogs.

6.6 Appendices

6.6.1 Study design

Table 6.15: Sample information

No.	BankID	Breed	Phenotype	Sequencing ID
1	2722	German shepherd	control	pool 1.1
2	6119	West Highland White terrier	control	pool 1.2
3	2724	Bull terrier	control	pool 1.3
4	5989	German shepherd	control	pool 1.4
5	1737	Shetland sheepdog	control	pool 1.5
6	5991	Shetland sheepdog	spinner	pool 1.6
7	3094	Jack Russell terrier	spinner	pool 1.7
8	5990	German shepherd dog	spinner	pool 1.8
9	5913	German shepherd dog	spinner	pool 1.9
10	205	Jack Russell terrier	control	pool 1.10
11	15260	Doberman Pinscher	Affected homozygote	pool 1.11
12	15237	Doberman Pinscher	Affected heterozygote	pool 1.12
13	15227	Doberman Pinscher	Unaffected non risk allele	pool 2.1
14	15253	Doberman Pinscher	Unaffected non risk allele	pool 2.2
15	15259	Doberman Pinscher	Affected heterozygote	pool 2.3
16	15254	Doberman Pinscher	Unaffected non risk allele	pool 2.4
17	15255	Doberman Pinscher	Affected homozygote	pool 2.5
19	15262	Doberman Pinscher	Unaffected non risk allele	pool 2.6
20	N/A	Bull terrier	chase light	pool 2.7
21	105502	SHIBA	N/A	pool 2.8
22	misc265	West Highland White terrier	Case	pool 2.9
23	misc259	Bull terrier, Chloe	case	pool 2.10
24	N/A	SHIBA Alex	N/A	pool 2.11
25	N/A	Bull terrier Tarzeys	control	pool 2.12

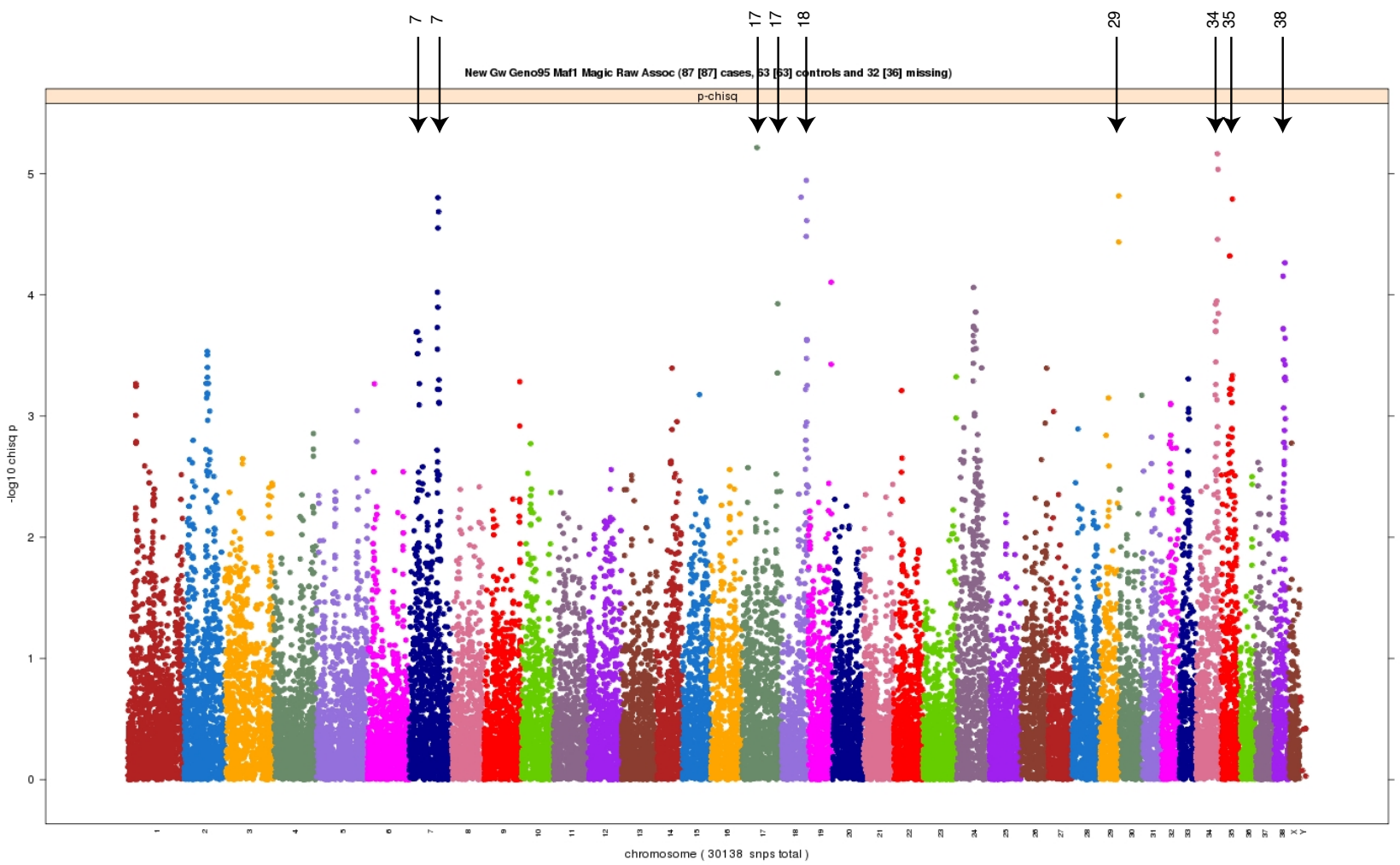


Figure 6.22: GWAS for CCD in Doberman Pinschers - The plot shows the GWAS results in which the data from (357) had been reanalysed with a different algorithm, MaGIC (482). Unpublished.

Table 6.16: P-values for the top GWA-intervals that were recalculated from (357). See Appendix Figure 6.22. Unpublished.

Chr	Start	End	Size (kb)	Comment	SNP position	P-value
7	61599000	61916800	317.8	OCD Association	chr7.61865715	1.58×10^{-5}
7	63734000	63968500	234.5	OCD Association	chr7.63867472	2.06×10^{-5}
17	27631900	27926000	294.1	OCD Association	chr17.27736094	6.09×10^{-6}
18	48719000	49010600	291.6	OCD Association	chr18.48909564	1.56×10^{-5}
18	57119000	57451000	332	OCD Association	chr18.57375179	2.44×10^{-5}
29	44099300	44611000	511.7	OCD Association	chr29.44152594	1.53×10^{-5}
34	38914700	39816800	902.1	OCD Association	chr34.38965782 chr34.39694895	6.85×10^{-6} 1.91×10^{-7}
35	15192000	15797300	605.3	OCD Association	chr35.15696211	1.83×10^{-5}
35	18464100	18881000	416.9	OCD Association	chr35.18565131	1.62×10^{-5}

Chr	Start	End	Size	Comment	Cumulative (Mb)
3	3624489	3665093	41	Homoz in Doberman, German Shepherd, Pit Bull Terrier	0.04
3	3780109	3798474	19	Homoz in Doberman, German Shepherd, Pit Bull Terrier	0.06
3	3875619	4029146	154	Homoz in Doberman, German Shepherd, Pit Bull Terrier	0.21
3	4593671	4695416	102	Homoz in Doberman, German Shepherd, Pit Bull Terrier	0.31
3	4903560	5084816	182	Homoz in Doberman, German Shepherd, Pit Bull Terrier	0.50
3	5942967	6250178	308	Homoz in Doberman, German Shepherd	0.80
7	61600000	61915715	316	OCD Association	1.12
7	63735000	63967472	233	OCD Association	1.35
17	27632930	27925000	293	OCD Association	1.64
17	46107426	46110040	3	Conserved	1.65
17	46270032	46270549	1	Conserved	1.65
17	46277750	46278220	1	Conserved	1.65
17	46448380	46449011	1	Conserved	1.65
17	46457814	46478790	21	Homoz in Doberman, Pit Bull Terrier	1.67
17	46538591	46539733	2	Conserved	1.67
17	46607478	46608034	1	Conserved	1.67
17	46660254	46762847	103	Homoz in Doberman, German Shepherd, Belg Terv, Irish Wolfhound	1.77
18	48720000	49009564	290	OCD Association	2.06
18	57120000	57450000	330	OCD Association	2.39
19	12899000	12906000	7	Homoz in Doberman, German Shepherd	2.40
19	13514903	13656613	142	Homoz in Doberman, German Shepherd	2.54
19	13692000	13695000	3	Homoz in Doberman, German Shepherd	2.54
29	44100382	44610000	510	OCD Association	3.05
34	38915782	39815702	900	OCD Association	3.95
35	15193000	15796211	604	OCD Association	4.56
35	18465131	18880000	415	OCD Association	4.97

Table 6.17: Target regions for resequencing - The regions include the top associated regions (Appendix Figure 6.22 and Table 6.16), the top fixed regions that are overlapping (homozygous) between the breeds predisposed with CCD (Doberman Pinschers, German Shepherds, Bull Terriers among others), and conserved regions around the homozygote blocks.

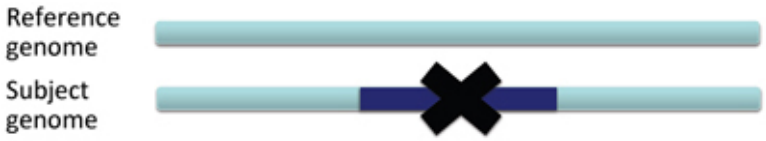


Figure 6.23: A schematic cartoon illustrating a genome with deletion - Adapted from (440).

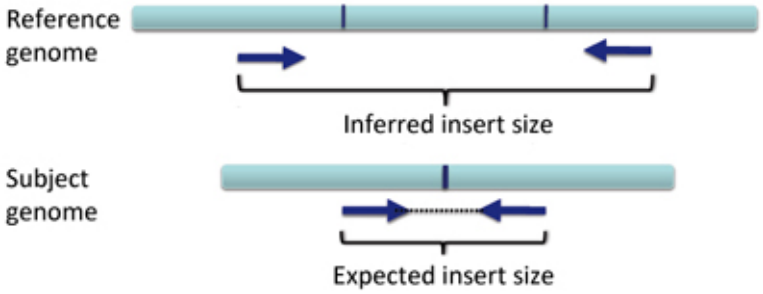


Figure 6.24: The difference occurred between inferred insert size and expected insert size due to a deletion in a subject genome - The "expected" insert size is obtained from the distribution of the distance between the paired-end reads from the sequencing of the subject genome. The "inferred" insert size is the insert size that would result when the same pair of reads is mapped to the reference genome. Thus, in the case of a deletion, the inferred insert size is greater than the expected insert size. When there are many such aberrant reads that have larger than expected inferred insert sizes within a certain region, this indicates a deletion. Adapted from (440).

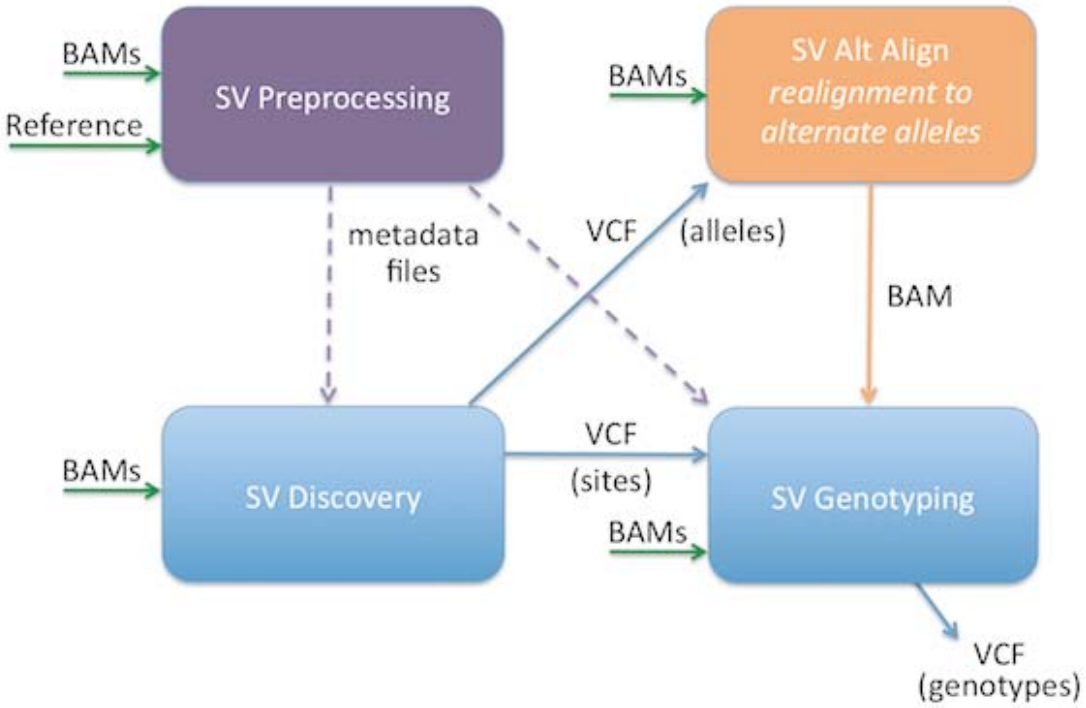


Figure 6.25: Workflow of modules in GenomeSTRiP - GenomeSTRiP (431) consists of four modules. Required input and output file formats for each module are indicated with arrows.

6.6.2 Sequencing statistics

Table 6.18: Functional variant counts per sample

FunctionalClass	Novelty	Sample	nSNPs	nInsertions	nDeletions
all	all	15227v2	10004	1284	1589
all	all	15237v2	9463	1222	1461
all	all	15253v2	10032	1283	1605
all	all	15254v2	8312	1121	1465
all	all	15255v2	10784	1331	1702
all	all	15259v2	9619	1235	1525
all	all	15260v2	7981	1072	1297
all	all	15262v2	9543	1225	1548
all	all	1737v2	10147	1397	1697
all	all	205v2	10346	1319	1605
all	all	2722v2	9763	1247	1571
all	all	2724v2	10153	1275	1565
all	all	3094v2	11148	1408	1672
all	all	5913v2	10964	1382	1639
all	all	5989v2	9231	1201	1540
all	all	5990v2	10019	1265	1580
all	all	5991v2	8832	1195	1550
all	all	6119v2	9136	1233	1565
all	all	all	22891	2284	2809
all	known	15227v2	3830	0	0
all	known	15237v2	3531	0	0
all	known	15253v2	3764	0	0
all	known	15254v2	3309	0	0
all	known	15255v2	3948	0	0
all	known	15259v2	3604	0	0
all	known	15260v2	2935	0	0
all	known	15262v2	3599	0	0
all	known	1737v2	3550	0	0
all	known	205v2	3689	0	0
all	known	2722v2	3608	0	0
all	known	2724v2	3330	0	0
all	known	3094v2	3976	0	0
all	known	5913v2	4016	0	0
all	known	5989v2	3505	0	0
all	known	5990v2	3590	0	0
all	known	5991v2	3324	0	0
all	known	6119v2	3256	0	0
all	known	all	5800	0	0
all	novel	15227v2	6174	1284	1589
all	novel	15237v2	5932	1222	1461
all	novel	15253v2	6268	1283	1605
all	novel	15254v2	5003	1121	1465
all	novel	15255v2	6836	1331	1702

Continued on next page

FunctionalClass	Novelty	Sample	nSNPs	nInsertions	nDeletions
all	novel	15259v2	6015	1235	1525
all	novel	15260v2	5046	1072	1297
all	novel	15262v2	5944	1225	1548
all	novel	1737v2	6597	1397	1697
all	novel	205v2	6657	1319	1605
all	novel	2722v2	6155	1247	1571
all	novel	2724v2	6823	1275	1565
all	novel	3094v2	7172	1408	1672
all	novel	5913v2	6948	1382	1639
all	novel	5989v2	5726	1201	1540
all	novel	5990v2	6429	1265	1580
all	novel	5991v2	5508	1195	1550
all	novel	6119v2	5880	1233	1565
all	novel	all	17091	2284	2809
missense	all	15227v2	26	0	0
missense	all	15237v2	22	0	0
missense	all	15253v2	23	0	0
missense	all	15254v2	31	0	0
missense	all	15255v2	34	0	0
missense	all	15259v2	29	0	0
missense	all	15260v2	22	0	0
missense	all	15262v2	29	0	0
missense	all	1737v2	27	0	0
missense	all	205v2	26	0	0
missense	all	2722v2	27	0	0
missense	all	2724v2	31	0	0
missense	all	3094v2	30	0	0
missense	all	5913v2	24	0	0
missense	all	5989v2	22	0	0
missense	all	5990v2	25	0	0
missense	all	5991v2	28	0	0
missense	all	6119v2	23	0	0
missense	all	all	64	0	0
missense	known	15227v2	9	0	0
missense	known	15237v2	11	0	0
missense	known	15253v2	10	0	0
missense	known	15254v2	12	0	0
missense	known	15255v2	11	0	0
missense	known	15259v2	11	0	0
missense	known	15260v2	10	0	0
missense	known	15262v2	10	0	0
missense	known	1737v2	10	0	0
missense	known	205v2	12	0	0
missense	known	2722v2	7	0	0
missense	known	2724v2	8	0	0
missense	known	3094v2	13	0	0
missense	known	5913v2	9	0	0

Continued on next page

FunctionalClass	Novelty	Sample	nSNPs	nInsertions	nDeletions
missense	known	5989v2	9	0	0
missense	known	5990v2	6	0	0
missense	known	5991v2	9	0	0
missense	known	6119v2	8	0	0
missense	known	all	16	0	0
missense	novel	15227v2	17	0	0
missense	novel	15237v2	11	0	0
missense	novel	15253v2	13	0	0
missense	novel	15254v2	19	0	0
missense	novel	15255v2	23	0	0
missense	novel	15259v2	18	0	0
missense	novel	15260v2	12	0	0
missense	novel	15262v2	19	0	0
missense	novel	1737v2	17	0	0
missense	novel	205v2	14	0	0
missense	novel	2722v2	20	0	0
missense	novel	2724v2	23	0	0
missense	novel	3094v2	17	0	0
missense	novel	5913v2	15	0	0
missense	novel	5989v2	13	0	0
missense	novel	5990v2	19	0	0
missense	novel	5991v2	19	0	0
missense	novel	6119v2	15	0	0
missense	novel	all	48	0	0
nonsense	all	15227v2	0	0	0
nonsense	all	15237v2	0	0	0
nonsense	all	15253v2	0	0	0
nonsense	all	15254v2	0	0	0
nonsense	all	15255v2	0	0	0
nonsense	all	15259v2	0	0	0
nonsense	all	15260v2	0	0	0
nonsense	all	15262v2	0	0	0
nonsense	all	1737v2	0	0	0
nonsense	all	205v2	0	0	0
nonsense	all	2722v2	0	0	0
nonsense	all	2724v2	0	0	0
nonsense	all	3094v2	0	0	0
nonsense	all	5913v2	0	0	0
nonsense	all	5989v2	0	0	0
nonsense	all	5990v2	0	0	0
nonsense	all	5991v2	0	0	0
nonsense	all	6119v2	0	0	0
nonsense	all	all	0	0	0
nonsense	known	15227v2	0	0	0
nonsense	known	15237v2	0	0	0
nonsense	known	15253v2	0	0	0
nonsense	known	15254v2	0	0	0

Continued on next page

FunctionalClass	Novelty	Sample	nSNPs	nInsertions	nDeletions
nonsense	known	15255v2	0	0	0
nonsense	known	15259v2	0	0	0
nonsense	known	15260v2	0	0	0
nonsense	known	15262v2	0	0	0
nonsense	known	1737v2	0	0	0
nonsense	known	205v2	0	0	0
nonsense	known	2722v2	0	0	0
nonsense	known	2724v2	0	0	0
nonsense	known	3094v2	0	0	0
nonsense	known	5913v2	0	0	0
nonsense	known	5989v2	0	0	0
nonsense	known	5990v2	0	0	0
nonsense	known	5991v2	0	0	0
nonsense	known	6119v2	0	0	0
nonsense	known	all	0	0	0
nonsense	novel	15227v2	0	0	0
nonsense	novel	15237v2	0	0	0
nonsense	novel	15253v2	0	0	0
nonsense	novel	15254v2	0	0	0
nonsense	novel	15255v2	0	0	0
nonsense	novel	15259v2	0	0	0
nonsense	novel	15260v2	0	0	0
nonsense	novel	15262v2	0	0	0
nonsense	novel	1737v2	0	0	0
nonsense	novel	205v2	0	0	0
nonsense	novel	2722v2	0	0	0
nonsense	novel	2724v2	0	0	0
nonsense	novel	3094v2	0	0	0
nonsense	novel	5913v2	0	0	0
nonsense	novel	5989v2	0	0	0
nonsense	novel	5990v2	0	0	0
nonsense	novel	5991v2	0	0	0
nonsense	novel	6119v2	0	0	0
nonsense	novel	all	0	0	0
silent	all	15227v2	40	0	0
silent	all	15237v2	37	0	0
silent	all	15253v2	45	0	0
silent	all	15254v2	40	0	0
silent	all	15255v2	50	0	0
silent	all	15259v2	48	0	0
silent	all	15260v2	37	0	0
silent	all	15262v2	43	0	0
silent	all	1737v2	45	0	0
silent	all	205v2	55	0	0
silent	all	2722v2	38	0	0
silent	all	2724v2	40	0	0
silent	all	3094v2	51	0	0

Continued on next page

FunctionalClass	Novelty	Sample	nSNPs	nInsertions	nDeletions
silent	all	5913v2	42	0	0
silent	all	5989v2	42	0	0
silent	all	5990v2	45	0	0
silent	all	5991v2	55	0	0
silent	all	6119v2	53	0	0
silent	all	all	109	0	0
silent	known	15227v2	16	0	0
silent	known	15237v2	17	0	0
silent	known	15253v2	17	0	0
silent	known	15254v2	16	0	0
silent	known	15255v2	20	0	0
silent	known	15259v2	20	0	0
silent	known	15260v2	19	0	0
silent	known	15262v2	17	0	0
silent	known	1737v2	17	0	0
silent	known	205v2	22	0	0
silent	known	2722v2	13	0	0
silent	known	2724v2	16	0	0
silent	known	3094v2	23	0	0
silent	known	5913v2	15	0	0
silent	known	5989v2	17	0	0
silent	known	5990v2	15	0	0
silent	known	5991v2	20	0	0
silent	known	6119v2	20	0	0
silent	known	all	31	0	0
silent	novel	15227v2	24	0	0
silent	novel	15237v2	20	0	0
silent	novel	15253v2	28	0	0
silent	novel	15254v2	24	0	0
silent	novel	15255v2	30	0	0
silent	novel	15259v2	28	0	0
silent	novel	15260v2	18	0	0
silent	novel	15262v2	26	0	0
silent	novel	1737v2	28	0	0
silent	novel	205v2	33	0	0
silent	novel	2722v2	25	0	0
silent	novel	2724v2	24	0	0
silent	novel	3094v2	28	0	0
silent	novel	5913v2	27	0	0
silent	novel	5989v2	25	0	0
silent	novel	5990v2	30	0	0
silent	novel	5991v2	35	0	0
silent	novel	6119v2	33	0	0
silent	novel	all	78	0	0

Table 6.19: Sequencing statistics - numbers of aligned reads

External ID	Mean Read Length (bp)	PF HQ Aligned Reads	PF HQ Aligned Q20 Bases	PF Unique Reads Aligned	PF Unique Bases Aligned
2722v2	101	25,699,764	2,531,330,774	9,120,129	910,366,662
6119v2	101	25,352,785	2,494,346,249	8,806,429	878,489,074
2724v2	101	23,853,375	2,344,090,890	8,388,534	836,201,648
5989v2	101	27,909,911	2,747,206,720	9,501,019	948,552,788
1737v2	101	26,722,999	2,631,130,773	9,471,348	945,271,757
5991v2	101	28,547,479	2,813,906,870	9,758,321	974,962,641
3094v2	101	16,315,860	1,608,449,481	12,727,660	1,267,941,716
5990v2	101	14,829,971	1,462,658,087	11,670,180	1,162,734,835
5913v2	101	14,406,888	1,420,485,540	11,378,360	1,133,632,267
205v2	101	14,042,492	1,383,909,425	10,675,417	1,062,769,474
15260v2	101	12,305,136	1,206,526,318	9,590,819	951,139,666
15237v2	101	14,513,695	1,426,238,023	11,472,319	1,139,683,488
15227v2	101	33,219,069	3,235,871,996	11,603,539	1,154,846,994
15253v2	101	30,002,920	2,921,570,537	10,564,662	1,050,532,650
15259v2	101	32,439,262	3,156,547,803	11,166,279	1,110,107,291
15254v2	101	34,366,566	3,344,060,444	11,506,248	1,143,518,339
15255v2	101	31,927,819	3,112,133,336	10,911,564	1,086,228,716
15262v2	101	24,953,473	2,414,409,590	9,555,896	945,386,032
Bull terrierv2	101	16,304,078	1,586,974,504	1,763,851	174,947,174
105502v2	101	16,248,911	1,581,728,100	1,755,163	174,069,743
misc265v2	101	18,600,625	1,814,149,695	1,862,364	184,993,135
misc259v2	101	17,884,859	1,743,544,200	1,863,758	185,261,292
SHIBA Alexv2	101	13,637,829	1,327,624,109	1,393,687	138,011,629
Bull terrier tarzeysv2	101	38,873,612	3,260,755,460	3,943,548	333,859,729

Table 6.20: Sequencing statistics - read depth of coverage - The **Mean Target Coverage** column shows the read depth of coverage for each dog, while the other columns show percentage target region covered by a given read depth shown in each column heading.

External ID	Mean Target Coverage	Zero Coverage Targets %	Target Bases 2x %	Target Bases 20x %
2722v2	77.67	0.39	99.46	93.52
6119v2	81.16	0.31	99.64	93.97
2724v2	78.67	0.46	99.55	91.69
5989v2	82.03	0.45	99.49	94.51
1737v2	81.74	0.46	99.48	94.56
5991v2	84.65	0.47	99.50	95.03
3094v2	74.28	0.62	99.36	92.09
5990v2	66.78	0.75	99.22	90.83
5913v2	62.97	0.56	99.39	91.27
205v2	61.70	0.49	99.31	89.39
15260v2	57.07	0.80	99.19	85.77
15237v2	69.67	0.73	99.31	90.79
15227v2	89.60	0.47	99.35	94.38
15253v2	78.24	0.48	99.35	93.17
15259v2	86.50	0.30	99.54	94.14
15254v2	88.04	0.36	99.53	94.09
15255v2	83.24	0.62	99.37	94.23
15262v2	68.95	0.84	99.00	88.59
Bull terrierv2	3.38	13.03	67.45	0.18
105502v2	3.02	14.87	63.24	0.14
misc265v2	3.15	13.08	65.65	0.14
misc259v2	3.21	13.32	66.05	0.14
SHIBA Alexv2	2.93	17.11	58.23	0.18
Bull terrier tarzeysv2	1.78	28.99	34.09	0.05

Table 6.21: Concordance between sequencing and SNP array data

Sample ID	#All positions	% matching	#Het positions	% matching	#HomAlt positions	% matching	#HomRef positions	%matching
15227v2	394	99.75	98	100.00	136	100.00	160	99.38
15237v2	391	98.99	93	97.89	135	100.00	163	99.39
15253v2	392	99.24	96	97.96	134	99.26	162	100.00
15254v2	393	99.49	70	98.59	132	99.25	191	100.00
15255v2	394	99.75	153	100.00	109	100.00	132	99.25
15259v2	394	99.75	85	100.00	142	99.30	167	100.00
15260v2	391	98.99	20	90.91	168	99.41	203	100.00
15262v2	393	99.49	73	97.33	139	100.00	181	100.00
1737v2	370	93.67	99	79.84	110	100.00	161	100.00
205v2	384	97.22	135	93.75	105	100.00	144	98.63
2722v2	394	99.75	83	100.00	149	99.33	162	100.00
2724v2	392	99.24	117	99.15	116	98.31	159	100.00
3094v2	392	99.24	127	97.69	129	100.00	136	100.00
5913v2	394	99.75	117	100.00	142	99.30	135	100.00
5989v2	393	99.49	56	100.00	175	99.43	162	99.39
5990v2	393	99.49	68	100.00	159	98.76	166	100.00
5991v2	393	99.49	103	100.00	93	98.94	197	99.49
6119v2	392	99.24	60	96.77	152	99.35	180	100.00

Table 6.22: Ti/Tv ratio per individual dog

Sample	nTi	nTv	tiTvRatio
15227v2	7330	2674	2.74
15237v2	6868	2595	2.65
15253v2	7307	2725	2.68
15254v2	6032	2280	2.65
15255v2	7887	2897	2.72
15259v2	7047	2572	2.74
15260v2	5789	2192	2.64
15262v2	6948	2595	2.68
1737v2	7368	2779	2.65
205v2	7566	2780	2.72
2722v2	7098	2665	2.66
2724v2	7356	2797	2.63
3094v2	8126	3022	2.69
5913v2	7950	3014	2.64
5989v2	6704	2527	2.65
5990v2	7288	2731	2.67
5991v2	6415	2417	2.65
6119v2	6587	2549	2.58
all	16796	6095	2.76

6.6.3 Variant statistics

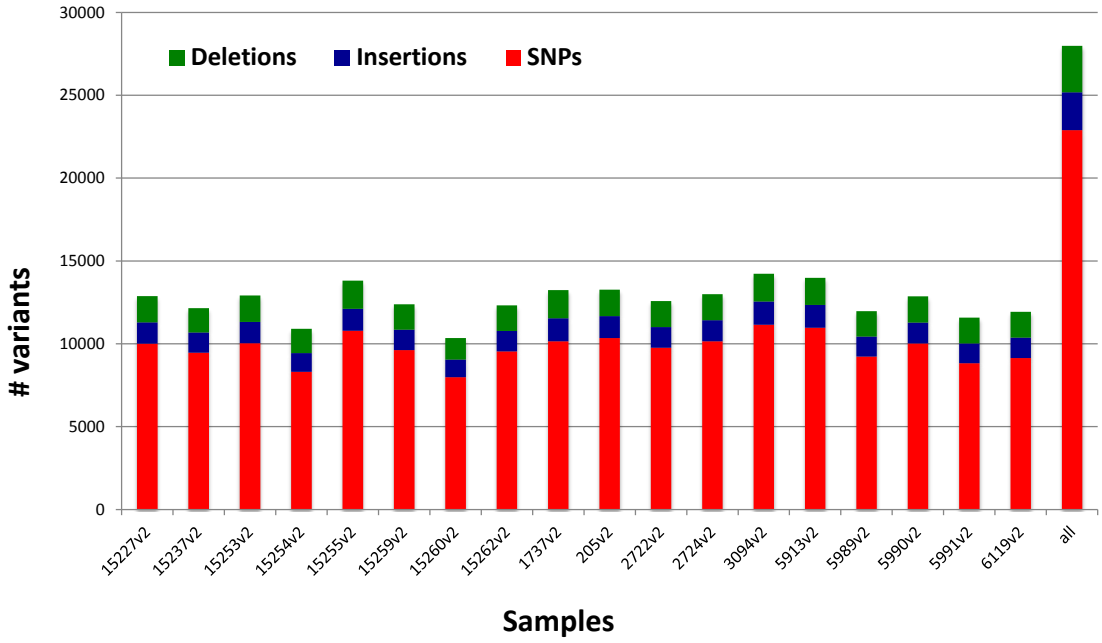


Figure 6.26: Variant types per pool - The graph shows the number of SNPs and micro-indels identified for each dog using GATK (366). See in-figure legend for the colour-key.

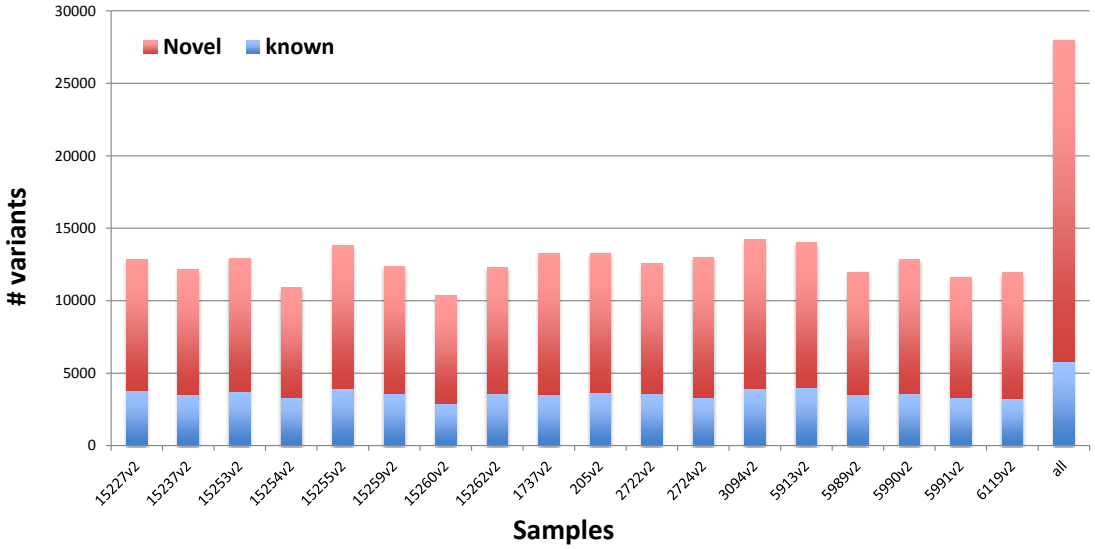


Figure 6.27: Variant novelty per pool - See in-figure legend for the colour-key.

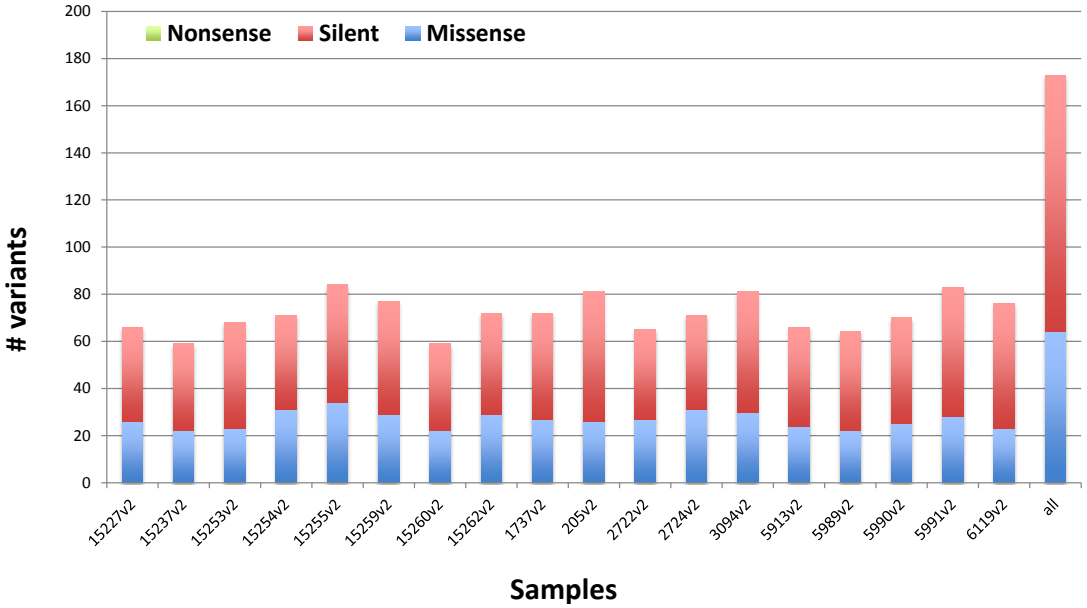


Figure 6.28: Functional variants per pool - See in-figure legend for the colour-key.

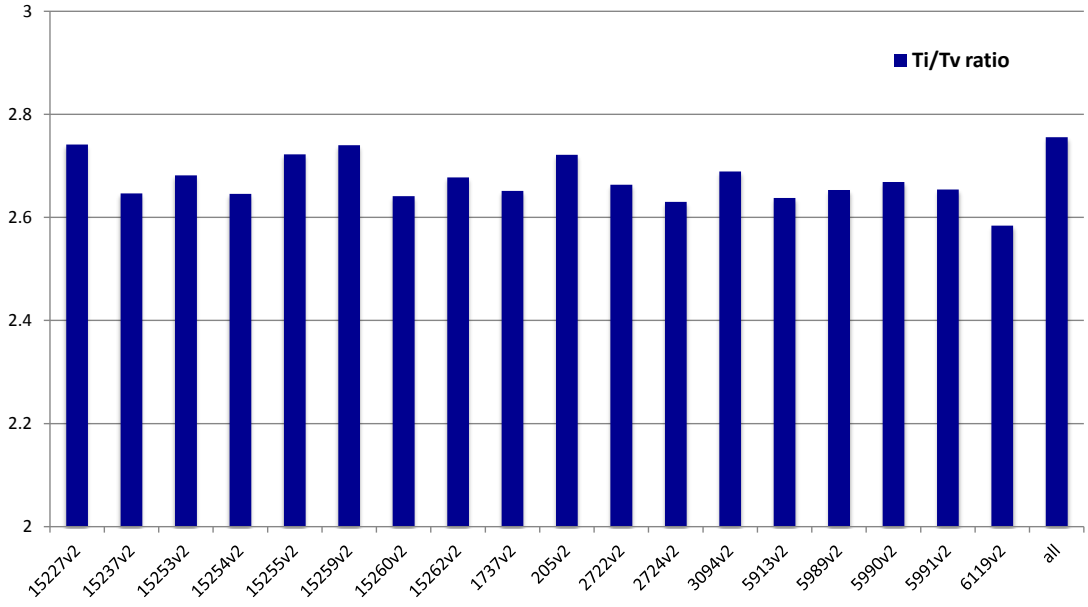


Figure 6.29: Ti/Tv ratio per pool -

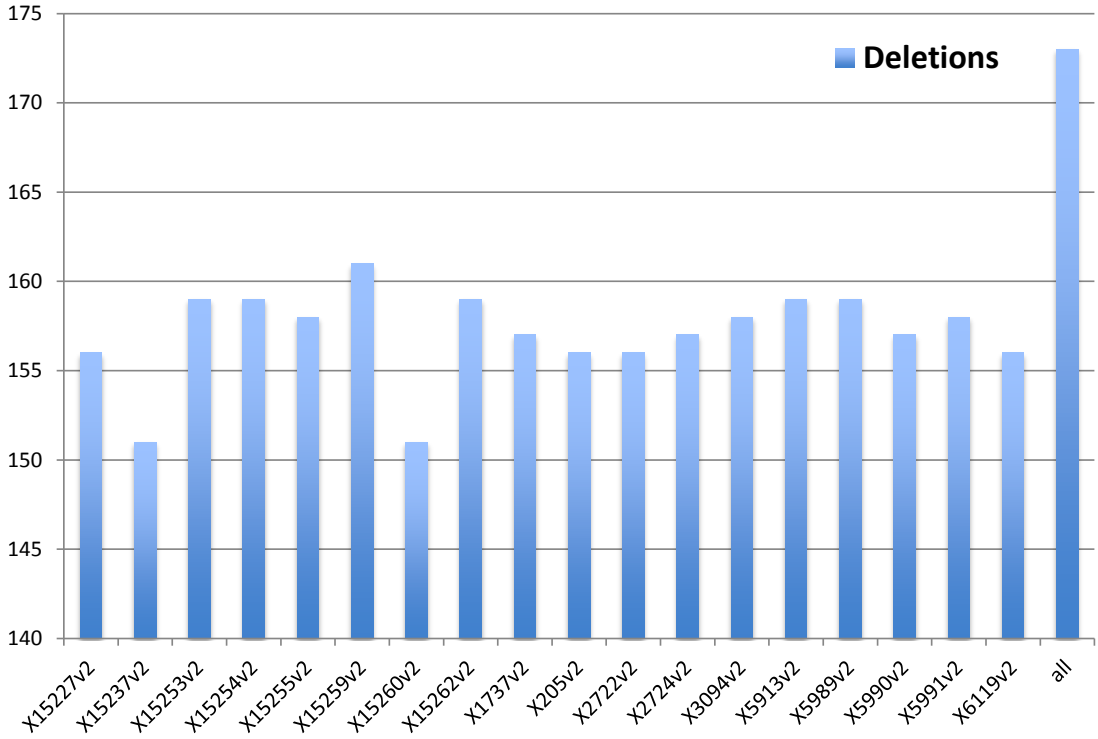


Figure 6.30: Copy number deletions per pool - Copy number deletions were called using GenomeSTRiP (431)

6.6.4 Prioritising variants

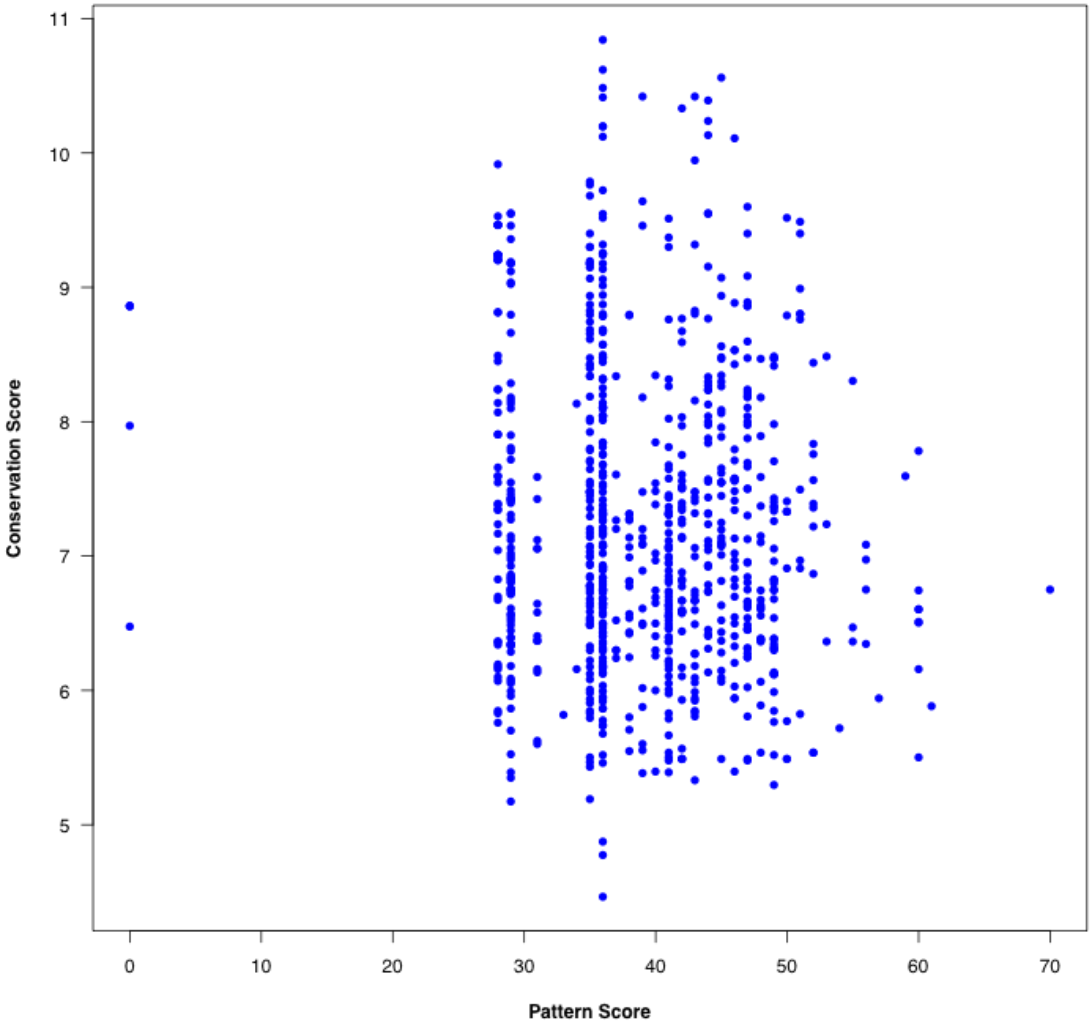


Figure 6.31: Variants plotted for conservation status and pattern score - SNPs that were analysed by SEQscoring (443) are shown.

# SNPs	621
# INDELs	321
# Coding change	4
# UTR change	4
# Conserved	21

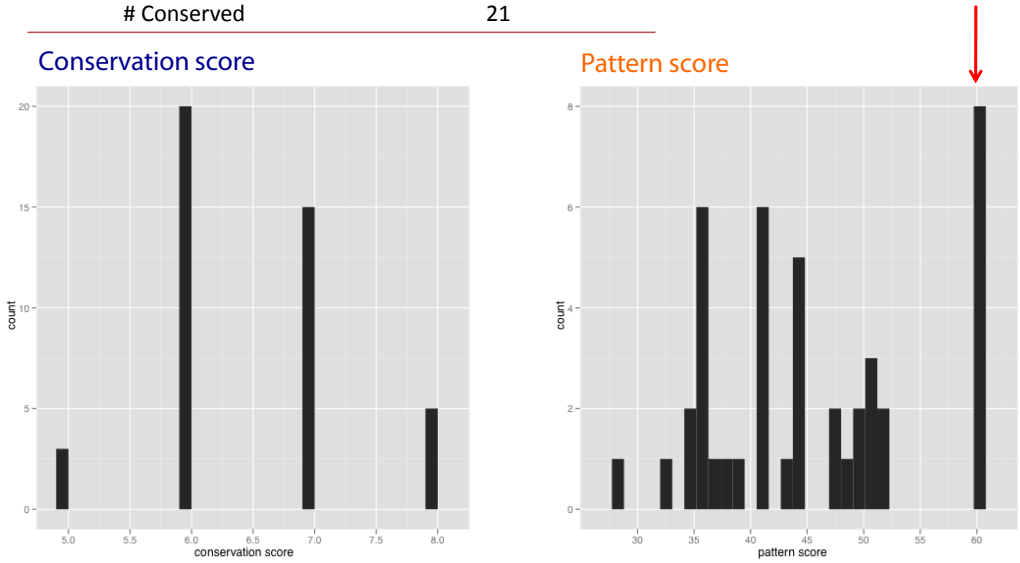


Figure 6.32: Variant statistics within *CDH2* - The number of SNPs and micro-indels within *CDH2* and functional variants are shown. The histograms are shown for the level of conservation of variant position (left) and for the pattern scores of variants (right). The segregation of variants with high pattern scores (indicated with red arrow) likely reflects the presence of a risk haplotype in *CDH2* among case Doberman Pinschers.

	Conservation score (>4)	Pattern score	Dogs with variant
Coding change 1	7	41	DP case (15255), JRT ctrl (205), GS spinner (5913), SS spinner (5991)
Coding change 2	8	51	All except DP case (15237), 3 DP ctrl (15253, 15254, 15262), SS ctrl (1737), SS spinner (5991)
Coding change 3	8	51	All except DP case (15237), 3 DP ctrl (15253, 15254, 15262), SS ctrl (1737), SS spinner (5991)
Coding change 4	8	43	DP case (15260), BT ctrl (2724), GS spinner (5913), GS ctrl (5989), WHWT ctrl (6119)
UTR change 1	NA (DEL)	NA	All
UTR change 2	NA (DEL)	NA	All except 2 DP case (15237, 15260), DP ctrl (15262), JRT ctrl (205), GS ctrl (5989)
UTR change 3	NA	NA	3 DP ctrl (15227, 15253, 15262)DP case (15259), SS ctrl (1737), JRT ctrl (205), BT ctrl (2724), GS ctrl (5989), GS spinner (5990), WGWT ctrl (6119)
UTR change 4	NA (IN)	NA	DP ctrl (15254)

Figure 6.33: Coding and UTR changes in *CDH2* region - 4 coding mutations and 4 UTR mutations within the *CDH2* region are shown for their conservation scores, pattern scores, and the dogs harbouring these mutations.

6.6.5 Batch effect in calling CNVs

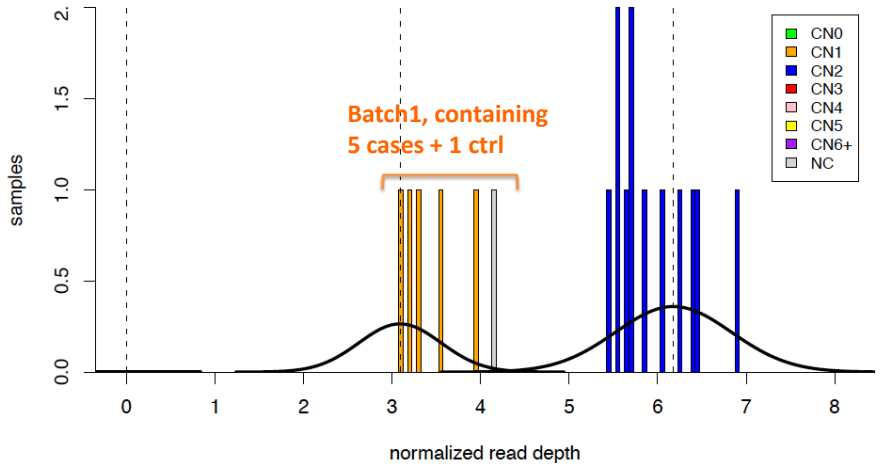


Figure 6.34: Batch effect in calling CNVs 1 - The plot shows population-scale distribution of read depth at a called deletion. Normalized measurements of read depth (across the deleted segment) from all samples were fitted to a Gaussian mixture model (431).

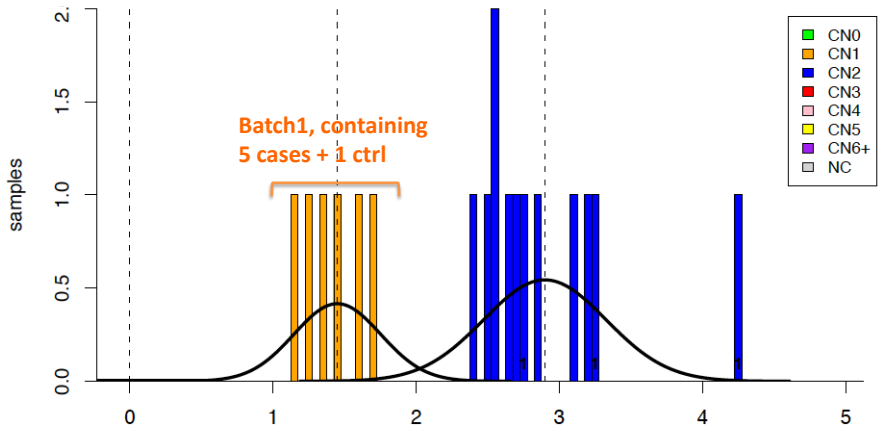


Figure 6.35: Batch effect in calling CNVs 2 - See Appendix Figure 6.34.

6.6.6 Case- and control-only variants within genes

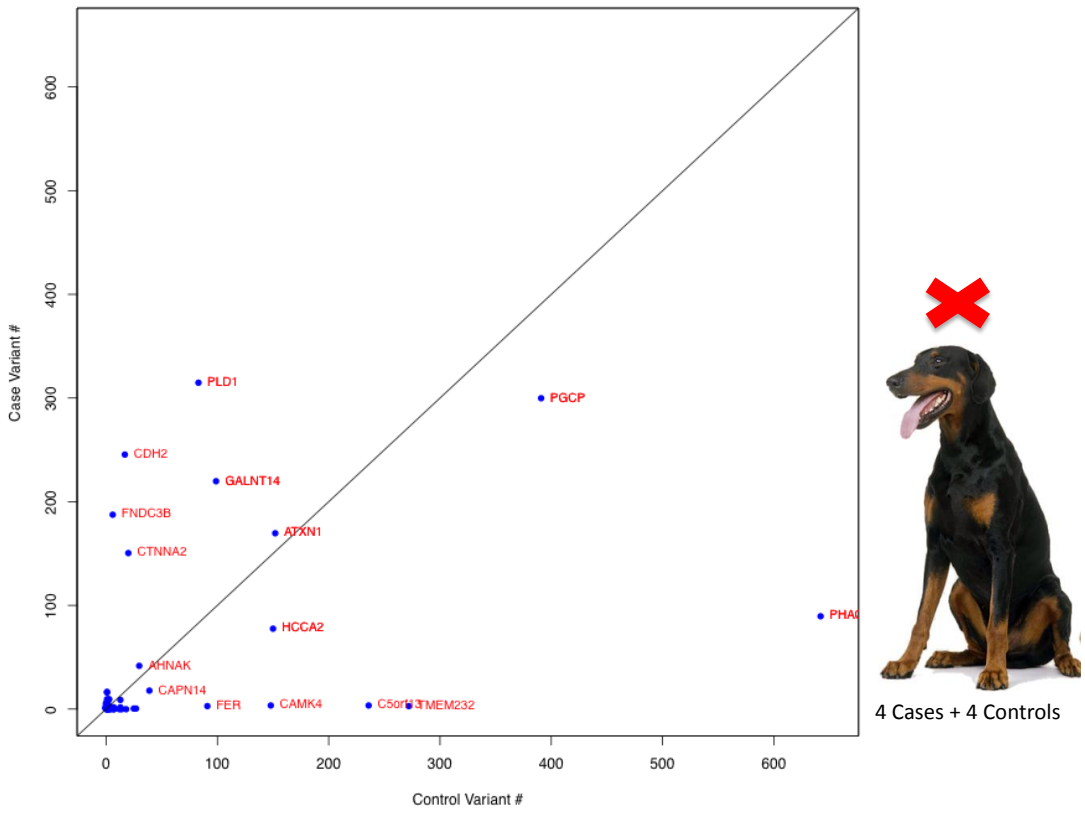


Figure 6.36: Case- and control-only variants in gene across all breeds excluding Doberman Pinschers - See Figure 6.14.

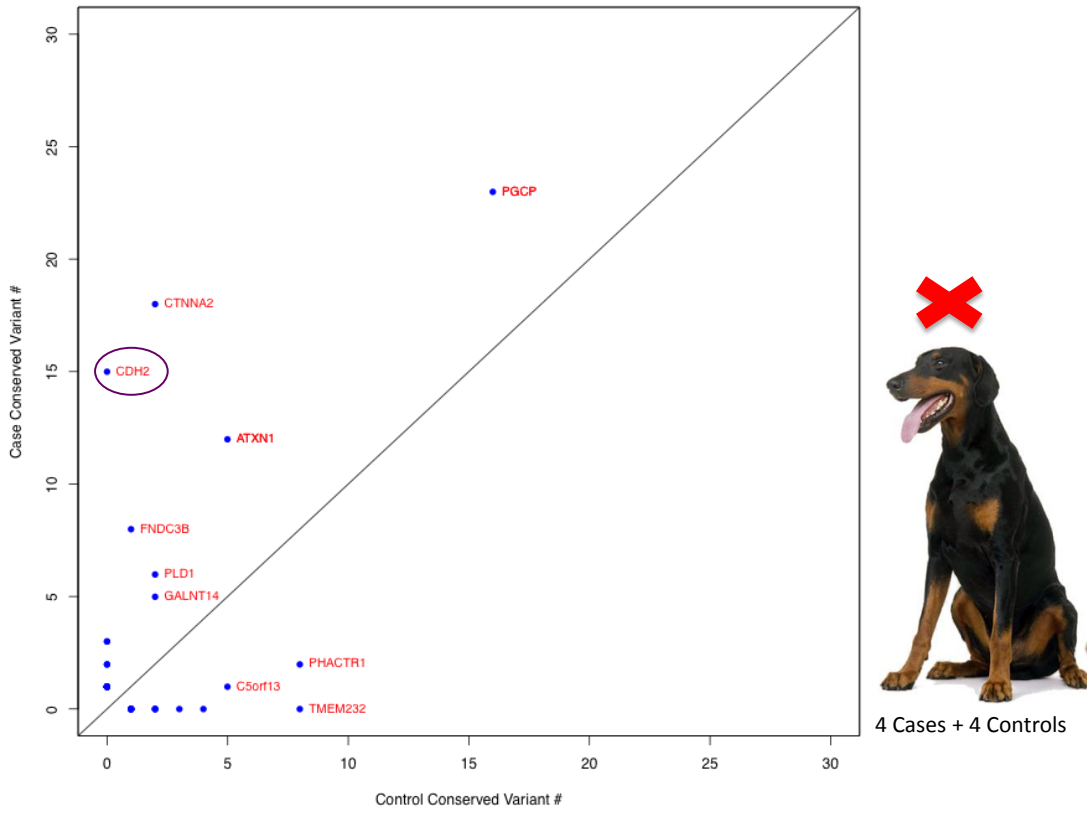


Figure 6.37: Case- and control-only variants in conserved regions of gene across all breeds excluding Doberman Pinschers - See Figure 6.14.

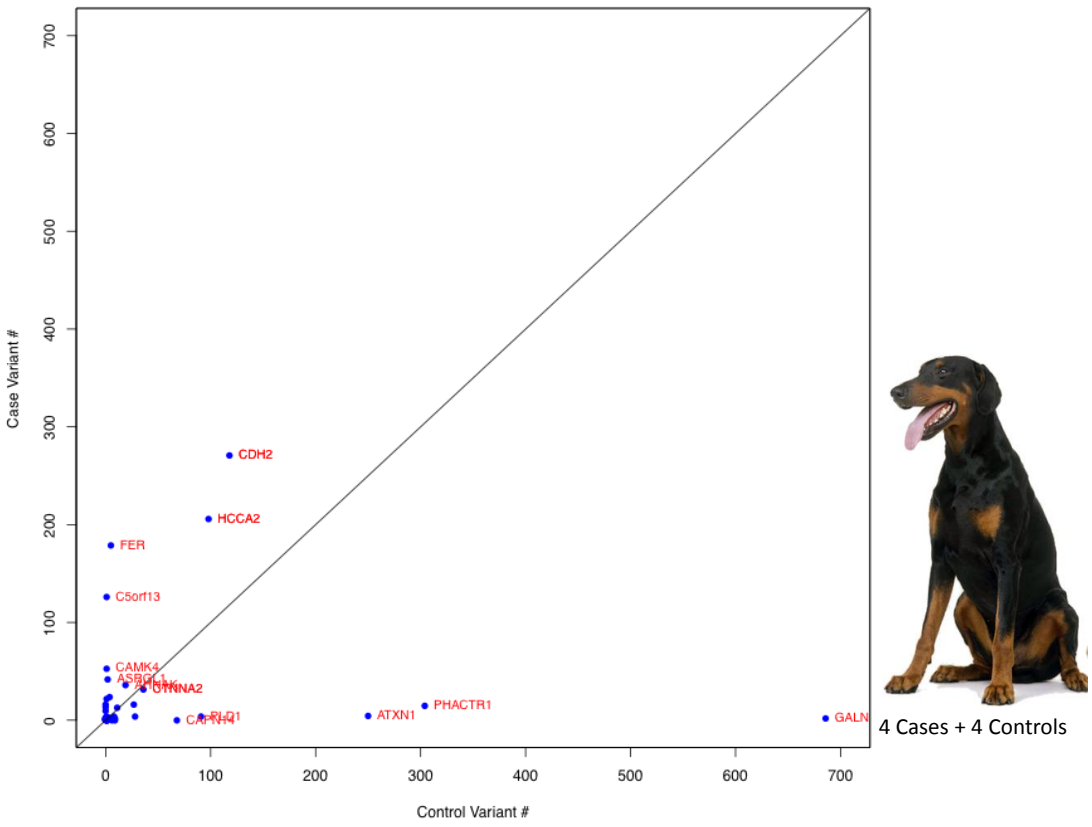


Figure 6.38: Case- and control-only variants in gene for Doberman Pinschers - See Figure 6.14.

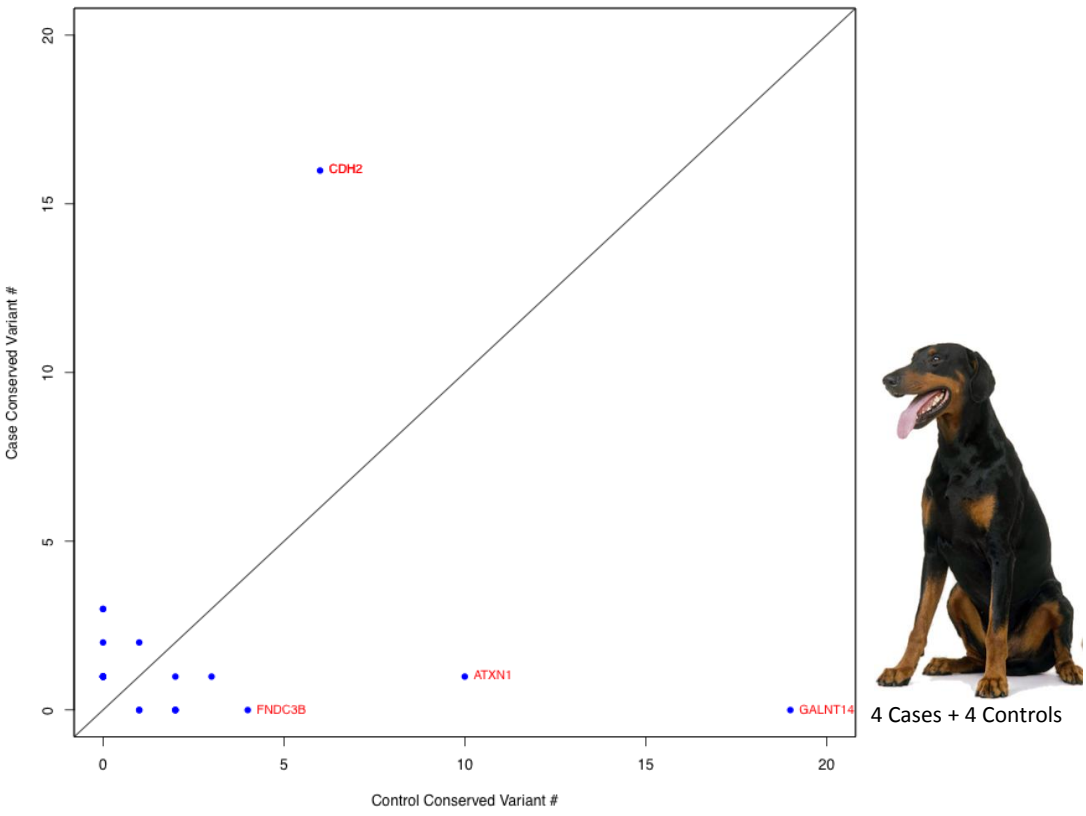


Figure 6.39: Case- and control-only variants in conserved regions of gene for Doberman Pinschers - See Figure 6.14.

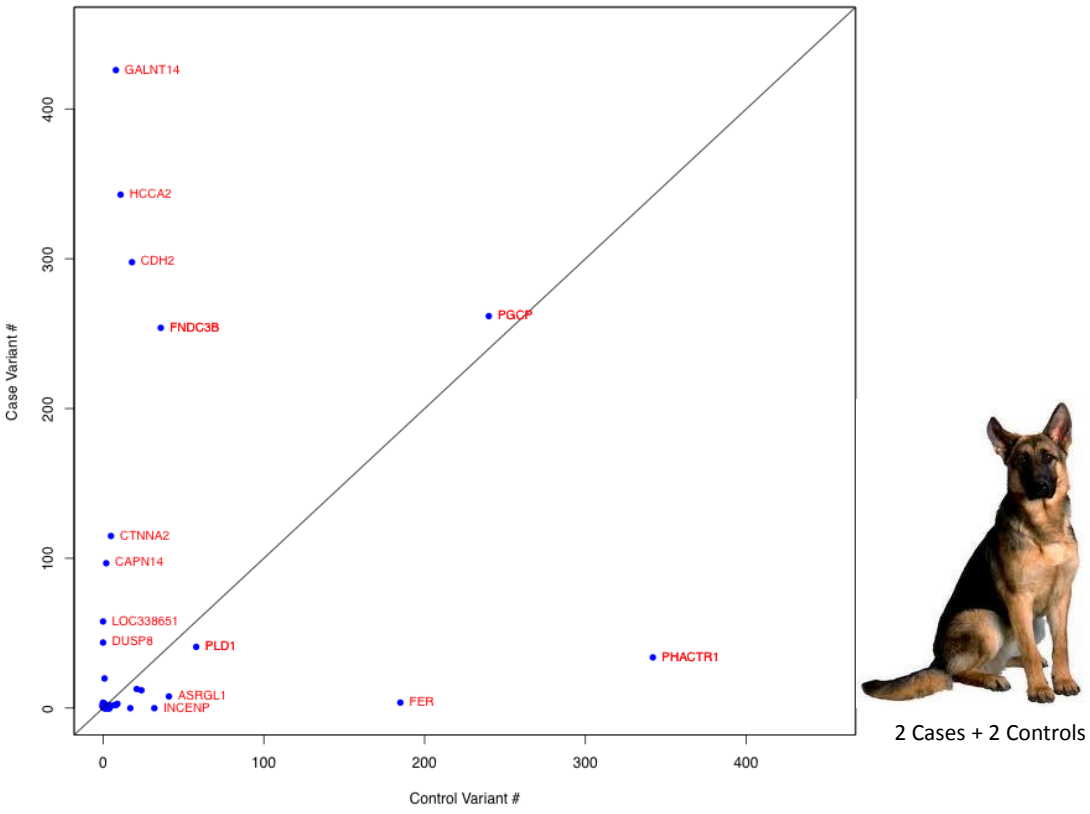


Figure 6.40: Case- and control-only variants in gene for German Shepherds - See Figure 6.14.

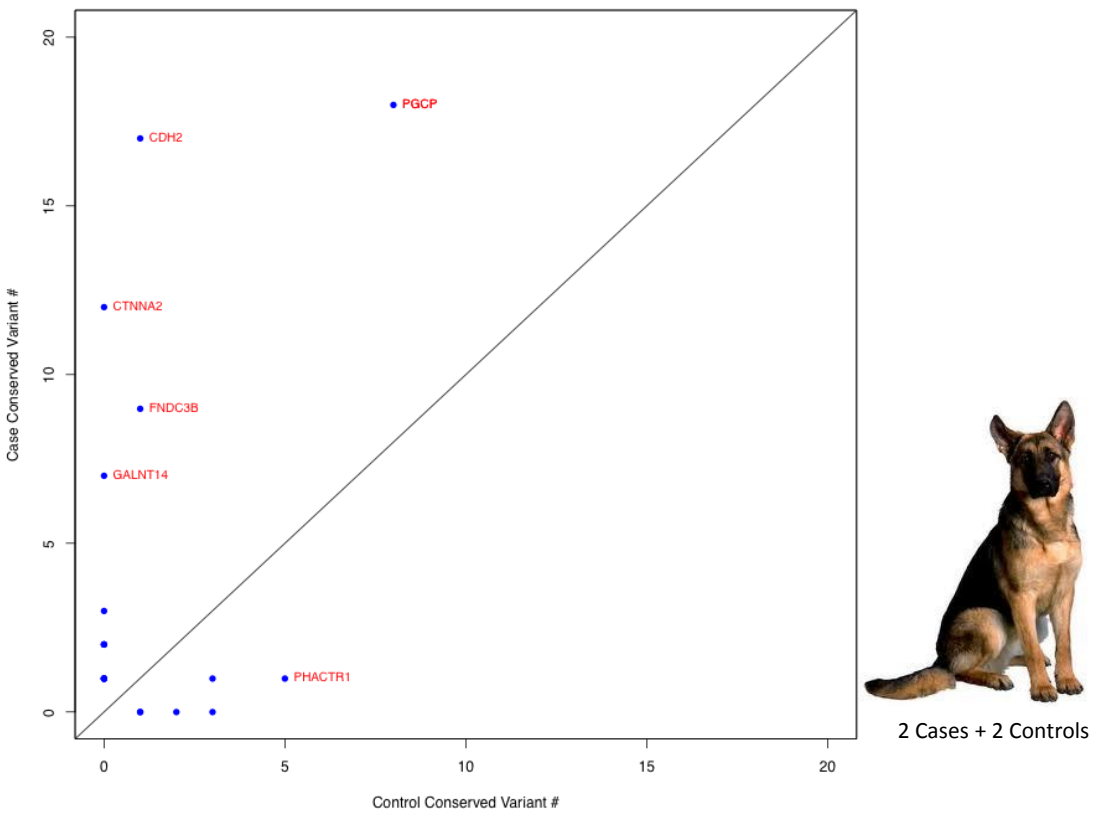


Figure 6.41: Case- and control-only variants in conserved regions of gene for German Shepherds - See Figure 6.14.

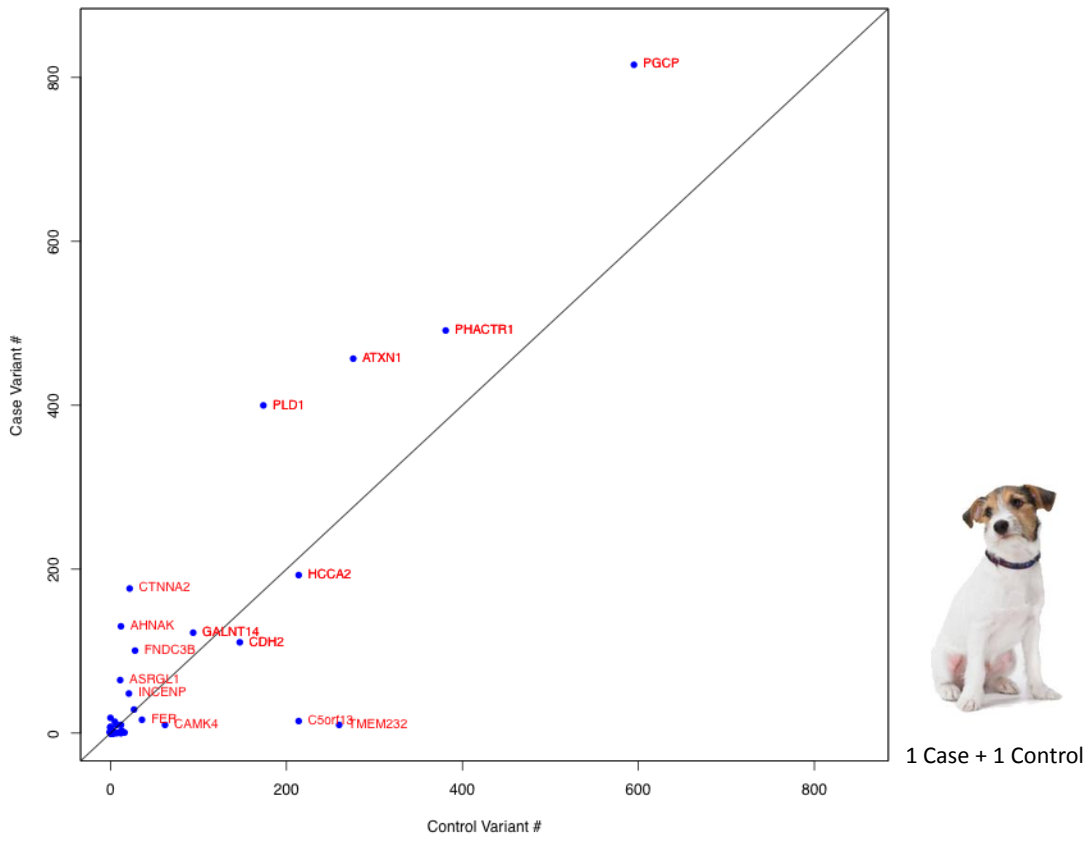


Figure 6.42: Case- and control-only variants in gene for Jack Russell Terriers - See Figure 6.14.

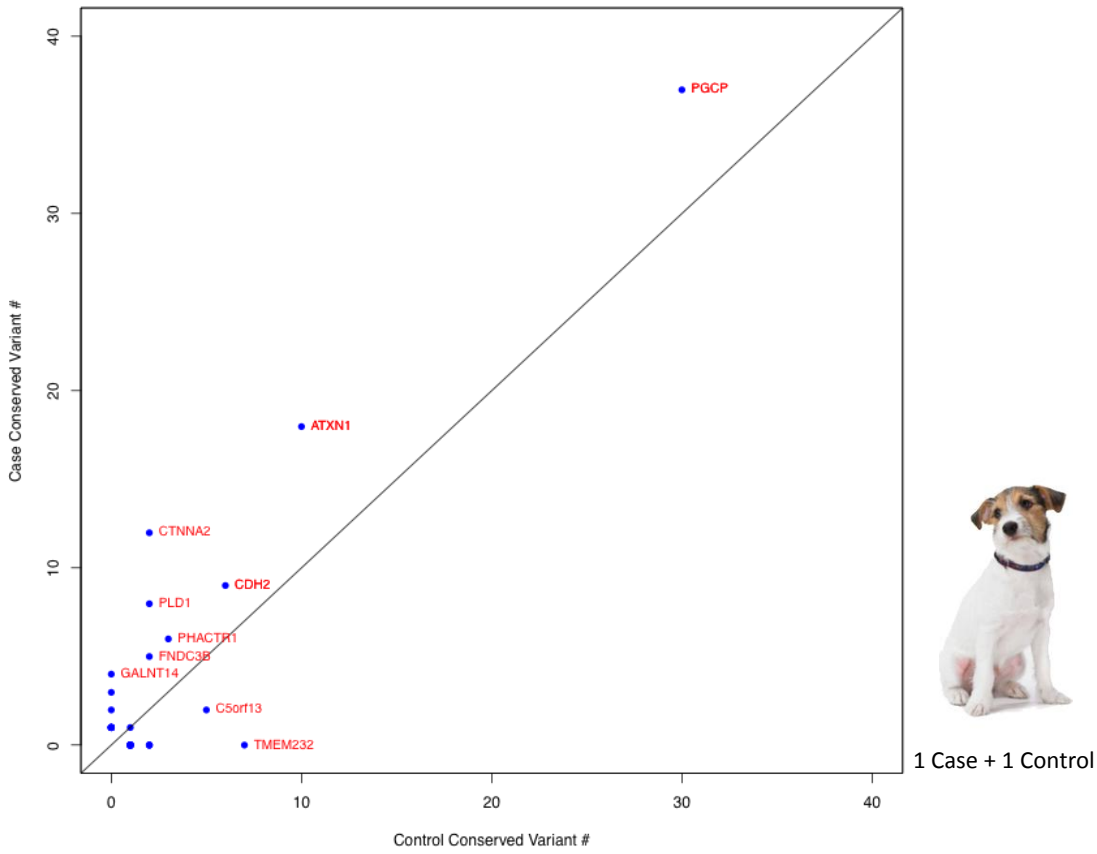


Figure 6.43: Case- and control-only variants in conserved regions of gene for Jack Russell Terriers - See Figure 6.14.

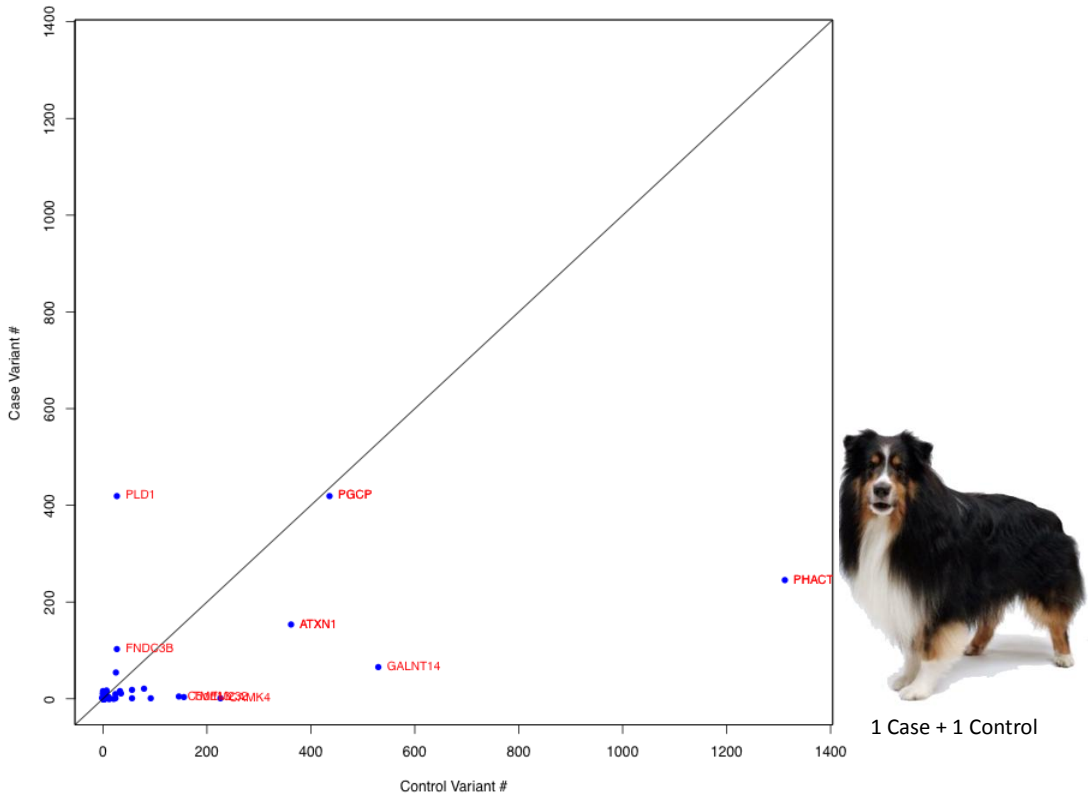


Figure 6.44: Case- and control-only variants in gene for Shetland Sheepdogs - See Figure 6.14.

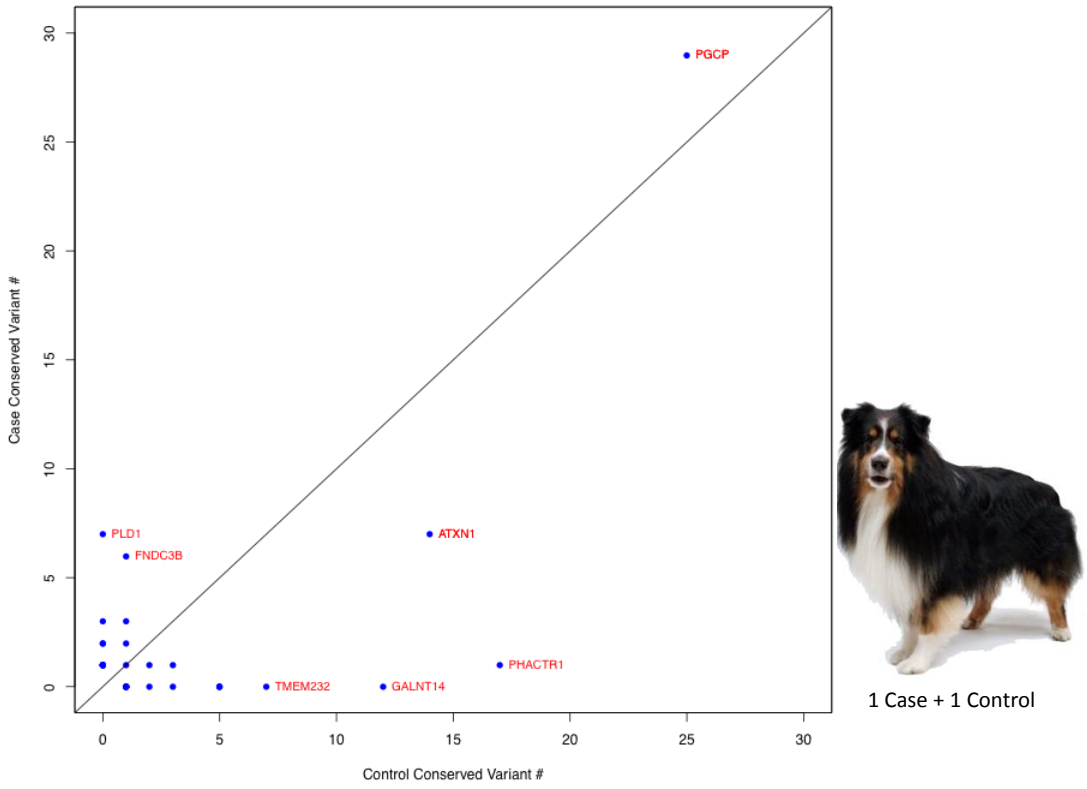


Figure 6.45: Case- and control-only variants in conserved regions of gene for Shetland Sheepdogs - See Figure 6.14.

Gene	ALL			DP			OTHERS		
	Case	Control	Ratio	Case	Control	Ratio	Case	Control	Ratio
CDH2	16	1	8	16	6	2.29	8	1	4
CTNNA2	12	2	4	0	2	0	5	2	1.67
ZFYVE28	3	0	3	NA	NA	NA	0	1	0
ATXN1	10	4	2	1	10	0.09	3	0	3
TMEM212	2	0	2	NA	NA	NA	0	2	0
CHRM1	2	0	2	NA	NA	NA	23	16	1.35
KIAA1530	2	0	2	NA	NA	NA	2	8	0.22
NOP14	2	0	2	NA	NA	NA	6	2	2
SLC22A8	2	0	2	2	0	2	1	0	1
PGCP	16	10	1.45	0	2	0	NA	NA	NA
FNDC3B	4	2	1.33	0	4	0	0	1	0
CAPN14	2	1	1	NA	NA	NA	1	5	0.17
PLD1	3	2	1	0	1	0	0	4	0
HAUS3	1	0	1	NA	NA	NA	1	0	1
MXD4	1	0	1	NA	NA	NA	18	2	6
SLC22A6	1	0	1	1	0	1	0	1	0
SORCS2	1	0	1	NA	NA	NA	0	1	0
TADA2B	1	0	1	NA	NA	NA	0	2	0
TBC1D14	1	0	1	NA	NA	NA	1	0	1
TNIP2	1	0	1	NA	NA	NA	1	0	1
WDR74	1	0	1	NA	NA	NA	1	0	1
GALNT14	6	6	0.86	0	19	0	0	1	0
PHACTR1	2	3	0.5	0	2	0	0	3	0
KIAA0232	1	1	0.5	NA	NA	NA	1	0	1
LRRTM1	1	1	0.5	NA	NA	NA	1	0	1
KRTAP5-8	1	2	0.33	1	2	0.33	1	0	1
C5orf13	1	2	0.33	3	0	3	0	2	0
CAMK4	1	2	0.33	1	0	1	1	0	1
AHNAK	0	1	0	1	0	1	2	0	2
DUSP8	0	1	0	2	1	1	1	0	1
MOB2	0	1	0	NA	NA	NA	1	0	1
TSPYL5	0	1	0	NA	NA	NA	1	0	1
FAM193A	0	1	0	NA	NA	NA	1	0	1
MAN2A1	0	5	0	NA	NA	NA	0	8	0
PJA2	0	2	0	1	0	1	1	0	1
STX5	0	1	0	1	3	0.25	0	1	0
TMEM232	0	8	0	NA	NA	NA	NA	NA	NA
WDR36	0	1	0	NA	NA	NA	3	0	3
HCCA2	NA	NA	NA	1	0	1	0	1	0
TNFSF10	NA	NA	NA	0	1	0	12	5	2
EPB41L4A	NA	NA	NA	1	0	1	2	0	2
FER	NA	NA	NA	3	0	3	15	0	15
MFSD10	NA	NA	NA	NA	NA	NA	NA	NA	NA

Table 6.23: Number of case- and control-only variants in conserved regions and their case to control ratio - ALL column shows for all breeds, DP column for Doberman Pinschers, OTHERS column for all breeds except for Doberman Pinschers; Ratio = (# case-only variants)/(# control-only variants+1); NA indicates the situation when no case- and control-only variants were detected for that gene.

6.6.7 Identifying transcription factor binding DNA motifs

Match^{TM} is designed to identify transcription factor binding sites in uncharacterised sequences by comparing them to a library of sequences that are linked to TRANSFAC® (129)(483). Firstly, a library of sequences with distribution matrices are derived from the DNA sites for which a binding of a specific transcription factor is known from the TRANSFAC® matrix table. The binding sites for a group of orthologous transcription factors are aligned and then, at each position, the frequency of the four nucleotides (A, C, G, and T) is counted, generating the consensus in a form of distribution matrices.

The sequence comparison between these consensus sequences and the uncharacterised sequence is then done in two steps: First, the 'core' of a matrix is compared to the input sequence, and then only where the core similarity is higher than the initially chosen threshold, the whole matrix is compared. The matrix-'core' used by Match^{TM} consists of the five consecutive nucleotide positions, which together yield the highest conservation value. The 5bp-core (capital letters in the result set) just serves to speed up the calculations, but it cannot define the whole binding matrix/sites sufficiently on its own. By default, matrix similarity cut-off=0.86 and core similarity cut-off=0.96 are used to identify a matching consensus sequence.

In actual calculation, the procedure may be described as Match^{TM} searches for subsequences x of an input sequence s, which are good matches to a matrix of TRANSFAC®. The similarity score of a subsequence x of sequence s with length L is calculated in the following way:

$$mat_sim = mat_sim(x) = mat_sim(b1, b2, ...bL) = \frac{Currentfreq - Minfreq}{Maxfreq - Minfreq},$$

where ($b1$,...,bL) are the nucleotides of x,
$f_{i,B}$ is the frequency of B to occur at position i,
$Currentfreq = \sum_{i=1}^{L} I(i) f_{i,bi}$, $Minfreq = \sum_{i=1}^{L} I(i) f_i^{min}$, $Maxfreq = \sum_{i=1}^{L} I(i) f_i^{max}$,
and information vector $I(i) = \sum_{B=1}^{4} f_{i,B} ln(f_{i,B})$.

6. RESEQUENCING OF GENOMIC REGIONS FOR COMPULSIVE DISORDER IN DOGS

7

Conclusions and future perspectives

The studies presented within this thesis employed novel comparative approaches to analysing genetic variants identified in patients with neuropsychiatric disorders. Despite the major role for model organisms in biomedical research, the genomic comparison of humans and model organisms for a specific disease has been less well exploited. Here, I demonstrated that comparative genomics approaches can be applied effectively to different ends in studying the genetics of complex human neuropsychiatric disorders. Through the comparative genomic analyses of human ASD and OCD with the mouse and the dog models, respectively, I identified a number of genetic candidates, analyses of which, using various functional genomics resources, provided important biological insights into disease aetiology.

7.1 ASD project

In Chapter 3, I exploited >5,000 Mouse Genome Informatics (MGI) phenotypes annotated with ∼6,000 human genes and objectively identified over 70 mouse model phenotypes that are significantly associated with ASD *dn* CNVs. Given the strikingly detailed concordance between the human ASD presentations including frequently co-morbid symptoms and associated mouse model phenotypes, if a comprehensive description of patients' symptoms becomes available, it would be interesting to investigate whether each patient with affected candidate genes exhibits symptoms that are comparable to the mouse model phenotypes that are annotated with their mouse orthologues. If that is the case, it would be a testament to the approach employed here that it not only was able to identify candidate genes for autism, but also classified these genes into subsets that are involved in the same biological process, whose disruption may lead to the specific symptoms of ASD.

Given that only a quarter of human genes are associated with the MGI phenotypes, not all human genes that contribute to the mouse phenotypic enrichments could be identified in Chapter 3. Thus, in Chapter 4, using ∼170,000 pair-wise interactions from ∼13,000 human proteins and their orthologues, I identified >100 additional ASD *dn* CNV genes whose protein products directly interact with the products of candidate genes from Chapter 3. The strong interconnectivity among these candidate genes implied that they function in concert. Their functional convergence was confirmed by over-representation of GO terms, mainly related to synaptic assemblies and functions.

Importantly, the distinct associations of regulatory GO terms between deleted genes and duplicated genes in ASD suggested a model for the mechanism of action between copy number deletions and copy number duplications found in ASD. The model suggests that deletion of positive regulators and duplication of negative regulators that act on the same biological processes lead to a similar outcome. Although this does not apply as a hard rule, much experimental evidence in the scientific literature supports this proposal. In addition, this model could be applied to interpret several ASD-relevant signalling pathways.

This mechanistic model of mechanism is the first proposal to be supported by statistically robust findings and that can explain how copy number duplications and deletions may exert similar neuropsychiatric symptoms. Although the model is based on genetic variants identified from ASD, it may also be applicable to other neuropsychiatric diseases such as schizophrenia that is thought to have a similar genetic architecture to ASD. Finally, as a complete genome-wide mouse knockout information will be available from the International Knockout Mouse Consortium (IKMC) (16) and a mouse phenotyping and other functional resources will become more comprehensive, regular revisiting of the approaches employed throughout Chapters 3 and 4 will perhaps be worthwhile.

7.2 OCD project

In Chapters 5 and 6, I took a different comparative approach, by employing domestic rather than laboratory animals as a model organism for human OCD. Dog is a good principal model for comparative analysis of human diseases for several reasons. Dogs have almost identical gene contents to humans, often display clinically very similar disorders to humans, and share their living spaces with humans to a large extent. Taking advantage of the extensive linkage disequilibrium in modern dog breeds, a previous GWAS identified canine genomic regions that are strongly associated with CCD in Doberman Pinschers. In order to identify disease genes and variants, candidate regions for CCD were resequenced in case and control dogs. In parallel, the corresponding regions in the human genome and other OCD candidate genes suggested from the neurobiology of OCD were also resequenced in human cases and controls. Case-control comparison of identified variants yielded a number of interesting candidate variants

and genes in both OCD and CCD studies. Comparative analyses of the top candidates between OCD and CCD revealed functional overlaps of the candidate genes, convergent in terms of the synaptic structure and assemblies. However, given the limited size of the sample cohorts, replication of these findings in a larger cohort will be necessary. While the replication study lies beyond the scope of this thesis, further resequencing of additional human case and control samples has been planned already. The new sequence dataset will be analysed with a similar scheme to that presented in Chapters 5 and 6. Lastly, the newly published cell-type specific transcription (484) and regulatory marks for the human genome (485)(486) from the ENCODE project will likely enhance the analyses to prioritise variants for OCD. Given that the activation of regulatory DNA varies in a cell-type specific manner (487), these regulatory maps on hundreds of different cell types including those from brain tissues will allow a more accurate identification of active regulatory sequence in a disease-relevant cell line.

References

[1] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. **Initial sequencing and analysis of the human genome**. *Nature*, **409**(6822):860–921, 2001. 2

[2] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, et al. **The sequence of the human genome**. *Science's STKE*, **291**(5507):1304, 2001. 2

[3] Kerstin Lindblad-Toh, Manuel Garber, Or Zuk, Michael F Lin, Brian J Parker, Stefan Washietl, Pouya Kheradpour, Jason Ernst, Gregory Jordan, Evan Mauceli, and et al. **A high-resolution map of human evolutionary constraint using 29 mammals**. *Nature*, **478**(7370):476–82, 2011. 2, 6, 7, 29, 211, 213, 328, 334, 335, 344, 345, 348, 361

[4] W. Miller, K.D. Makova, A. Nekrutenko, and R.C. Hardison. **Comparative genomics**. *Annu. Rev. Genomics Hum. Genet.*, **5**:15–56, 2004. 3

[5] S Kamholz, L Liefer, K Wetterstrand, Fs Collins, Tr Gingeras, D Kampa, Ea Sekinger, J Cheng, with Feingold Ea Hirsch, H, Pj Good, and Guyer. **The ENCODE (ENCyclopedia Of DNA Elements) Project.** *Science*, **306**(5696):636–640, 2004. 3, 310, 361

[6] K.R. Rosenbloom, T.R. Dreszer, J.C. Long, V.S. Malladi, C.A. Sloan, B.J. Raney, M.S. Cline, D. Karolchik, G.P. Barber, H. Clawson, et al. **ENCODE whole-genome data in the UCSC Genome Browser: update 2012**. *Nucleic acids research*, **40**(D1):D912–D917, 2012. 3, 29

[7] S.B. Hedges, J. Dudley, and S. Kumar. **TimeTree: a public knowledge-base of divergence times among organisms**. *Bioinformatics*, **22**(23):2971–2972, 2006. 4

[8] T.H. Jukes and M. Kimura. **Evolutionary constraints and the neutral theory**. *Journal of molecular evolution*, **21**(1):90–92, 1984. 4

[9] A. Ratnakumar, S. Mousset, S. Glémin, J. Berglund, N. Galtier, L. Duret, and M.T. Webster. **Detecting positive selection within genomes: the problem of biased gene conversion**. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**(1552):2571–2580, 2010. 4

[10] L. Duret and N. Galtier. **Biased gene conversion and the evolution of mammalian genomic landscapes**. *Annual review of genomics and human genetics*, **10**:285–311, 2009. 4

REFERENCES

[11] N. Galtier and L. Duret. **Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution**. *TRENDS in Genetics*, **23**(6):273–277, 2007. 4

[12] R.C. Hardison. **Comparative genomics**. *PLoS biology*, **1**(2):e58, 2003. 5

[13] A.T. Chinwalla, L.L. Cook, K.D. Delehaunty, G.A. Fewell, L.A. Fulton, R.S. Fulton, T.A. Graves, L.D.W. Hillier, E.R. Mardis, J.D. McPherson, et al. **Initial sequencing and comparative analysis of the mouse genome**. *Nature*, **420**(6915):520–562, 2002. 4, 6

[14] S. Meader, C.P. Ponting, and G. Lunter. **Massive turnover of functional sequence in human and other mammalian genomes**. *Genome research*, **20**(10):1335–1343, 2010. 6, 7

[15] W.C. Skarnes, B. Rosen, A.P. West, M. Koutsourakis, W. Bushell, V. Iyer, A.O. Mujica, M. Thomas, J. Harrow, T. Cox, et al. **A conditional knockout resource for the genome-wide study of mouse gene function**. *Nature*, **474**(7351):337–342, 2011. 6

[16] **International Knockout Mouse Consortium**, Sep 2012. 6, 409

[17] Carol J Bult, Janan T Eppig, James A Kadin, Joel E Richardson, and Judith A Blake. **The Mouse Genome Database (MGD): mouse biology and model systems**. *Nucleic Acids Research*, **36**(Database issue):D724–D728, 2008. 6, 28, 43, 60, 103

[18] D. Altshuler, M.J. Daly, and E.S. Lander. **Genetic mapping in human disease**. *science*, **322**(5903):881–888, 2008. 7, 8

[19] D.E. Reich and E.S. Lander. **On the allelic spectrum of human disease**. *TRENDS in Genetics*, **17**(9):502–510, 2001. 8

[20] N. Risch and K. Merikangas. **The future of genetic studies of complex human diseases**. *Science*, **273**(5281):1516, 1996. 8

[21] A. Chakravarti et al. **Population geneticsmaking sense out of sequence**. *Nature genetics*, **21**(Suppl 1):56–60, 1999. 8

[22] T.A. Manolio, F.S. Collins, N.J. Cox, D.B. Goldstein, L.A. Hindorff, D.J. Hunter, M.I. McCarthy, E.M. Ramos, L.R. Cardon, A. Chakravarti, et al. **Finding the missing heritability of complex diseases**. *Nature*, **461**(7265):747–753, 2009. 8

[23] P. Levitt et al. **The conundrums of understanding genetic risks for autism spectrum disorders**. *Nature Neuroscience*, 2011. 8, 20

[24] R. Uher. **The role of genetic variation in the causation of mental illness: an evolution-informed framework**. *Molecular psychiatry*, **14**(12):1072–1082, 2009. 8, 9, 17, 19

[25] P. Lichtenstein, B.H. Yip, C. Björk, Y. Pawitan, T.D. Cannon, P.F. Sullivan, and C.M. Hultman. **Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study**. *The Lancet*, **373**(9659):234–239, 2009. 8, 12, 20

[26] J. McClellan and M.C. King. **Genetic heterogeneity in human disease**. *Cell*, **141**(2):210–217, 2010. 8

[27] Christian R Marshall, Abdul Noor, John B Vincent, Anath C Lionel, Lars Feuk, Jennifer Skaug, Mary Shago, Rainald Moessner, Dalila Pinto, Yan Ren, and et al. **Structural Variation of Chromosomes in Autism Spectrum Disorder**. *The American Journal of Human Genetics*, **82**(2):477–488, 2008. 9, 26, 38, 42

[28] D. Pinto, Klei L. Pagnamenta, A.T., Merico D. Anney, R., Conroy J. Regan, R., Correia C. Magalhaes, T.R., and B.S. et al. Abrahams. **Functional impact of global rare copy number variation in autism spectrum disorders.** *Nature*, 2010. 9, 26, 42, 44, 52, 90

[29] T. Walsh, J.M. McClellan, S.E. McCarthy, A.M. Addington, S.B. Pierce, G.M. Cooper, A.S. Nord, M. Kusenda, D. Malhotra, A. Bhandari, et al. **Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia**. *Science's STKE*, **320**(5875):539, 2008. 9, 17

[30] R.C. Kessler, P. Berglund, O. Demler, R. Jin, K.R. Merikangas, and E.E. Walters. **Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication**. *Archives of general psychiatry*, **62**(6):593, 2005. 10

[31] M. Rutter. **Incidence of autism spectrum disorders: Changes over time and their meaning**. *Acta Paediatrica*, **94**(1):2–15, 2005. 10

[32] H.W. Hoek. **Incidence, prevalence and mortality of anorexia nervosa and other eating disorders**. *Current opinion in psychiatry*, **19**(4):389, 2006. 10

[33] J. McGrath, S. Saha, D. Chant, and J. Welham. **Schizophrenia: a concise overview of incidence, prevalence, and mortality**. *Epidemiologic reviews*, **30**(1):67–76, 2008. 10, 11

[34] P. Bebbington and R. Ramana. **The epidemiology of bipolar affective disorder**. *Social psychiatry and psychiatric epidemiology*, **30**(6):279–292, 1995. 10

[35] L. Andrade, J.J. Caraveo-anduaga, P. Berglund, R.V. Bijl, R.D. Graaf, W. Vollebergh, E. Dragomirecka, R. Kohn, M. Keller, R.C. Kessler, et al. **The epidemiology of major depressive episodes: results from the International Consortium of Psychiatric Epidemiology (ICPE) Surveys**. *International journal of methods in psychiatric research*, **12**(1):3–21, 2003. 10

[36] M.M. Weissman, R.C. Bland, G.J. Canino, C. Faravelli, S. Greenwald, H.G. Hwu, P.R. Joyce, E.G. Karam, C.K. Lee, J. Lellouch, et al. **Cross-national epidemiology of major depression and bipolar disorder**. *JAMA: the journal of the American Medical Association*, **276**(4):293–299, 1996. 10

REFERENCES

[37] N. Husain, F. Creed, and B. Tomenson. **Depression and social stress in Pakistan**. *Psycho@articleharris1998excess, title=Excess mortality of mental disorder., author=Harris, E.C. and Barraclough, B., journal=The British Journal of Psychiatry, volume=173, number=1, pages=11–53, year=1998, publisher=RCP logical medicine*, **30**(02):395–402, 2000. 10

[38] S. Lee, A. Tsang, YQ Huang, YL He, ZR Liu, MY Zhang, YC Shen, RC Kessler, et al. **The epidemiology of depression in metropolitan China**. *Psychological medicine*, **39**(5):735, 2010. 10

[39] B. Pavlova, R. Uher, E. Dragomirecka, and H. Papezova. **Trends in hospital admissions for eating disorders in a country undergoing a socio-cultural transition, the Czech Republic 1981–2005**. *Social psychiatry and psychiatric epidemiology*, **45**(5):541–550, 2010. 10

[40] J.B. Kirkbride, P. Fearon, C. Morgan, P. Dazzan, K. Morgan, J. Tarrant, T. Lloyd, J. Holloway, G. Hutchinson, J.P. Leff, et al. **Heterogeneity in incidence rates of schizophrenia and other psychotic syndromes: findings from the 3-center AeSOP study**. *Archives of General Psychiatry*, **63**(3):250, 2006. 10

[41] E.C. Harris and B. Barraclough. **Excess mortality of mental disorder.** *The British Journal of Psychiatry*, **173**(1):11–53, 1998. 11

[42] U. Hiroeh, L. Appleby, P.B. Mortensen, and G. Dunn. **Death by homicide, suicide, and other unnatural causes in people with mental illness: a population-based study**. *The Lancet*, **358**(9299):2110–2112, 2001. 11

[43] U. Hiroeh, N. Kapur, R. Webb, G. Dunn, P.B. Mortensen, and L. Appleby. **Deaths from natural causes in people with mental illness: a cohort study**. *Journal of psychosomatic research*, **64**(3):275–283, 2008. 11

[44] F.C. Papadopoulos, A. Ekbom, L. Brandt, and L. Ekselius. **Excess mortality, causes of death and prognostic factors in anorexia nervosa**. *The British Journal of Psychiatry*, **194**(1):10–17, 2009. 11

[45] M. Joukamaa, M. HELIÖVAARA, P. KNEKT, A. AROMAA, R. RAITASALO, and V. LEHTINEN. **Mental disorders and cause-specific mortality**. *The British Journal of Psychiatry*, **179**(6):498–502, 2001. 11

[46] S.E. Mouridsen, H. Brønnum-Hansen, B. Rich, and T. Isager. **Mortality and causes of death in autism spectrum disorders**. *Autism*, **12**(4):403–414, 2008. 11

[47] R.M. Shavelle, D.J. Strauss, and J. Pickett. **Causes of death in autism**. *Journal of autism and developmental disorders*, **31**(6):569–576, 2001. 11

[48] U. Osby, L. Brandt, N. Correia, A. Ekbom, and P. Sparen. **Excess mortality in bipolar and unipolar disorder in Sweden**. *Archives of General Psychiatry*, **58**(9):844, 2001. 11

[49] P. Cuijpers and F. Smit. **Excess mortality in depression: a meta-analysis of community studies**. *Journal of affective disorders*, **72**(3):227–236, 2002. 11

[50] J.M. Murphy, J.D. Burke, R.R. Monson, N.J. Horton, N.M. Laird, A. Lesage, A.M. Sobol, and A.H. Leighton. **Mortality associated with depression**. *Social psychiatry and psychiatric epidemiology*, **43**(8):594–601, 2008. 11

[51] G.R. Weitoft and M. Rosén. **Is perceived nervousness and anxiety a predictor of premature mortality and severe morbidity? A longitudinal follow up of the Swedish survey of living conditions**. *Journal of epidemiology and community health*, **59**(9):794–798, 2005. 11

[52] C. Allgulander and P.W. Lavori. **Excess mortality among 3302 patients with'pure' anxiety neurosis**. *Archives of general psychiatry*, **48**(7):599, 1991. 11

[53] FW Larsen and SE Mouridsen. **The outcome in children with childhood autism and Asperger syndrome originally diagnosed as psychotic. A 30-year follow-up study of subjects hospitalized as children**. *European child & adolescent psychiatry*, **6**(4):181–190, 1997. 11

[54] M. Brinch, T. Isager, and K. Tolstrup. **Anorexia nervosa and motherhood: reproduction pattern and mothering behavior of 50 women**. *Acta Psychiatrica Scandinavica*, **77**(5):611–617, 1988. 11

[55] A.C. Svensson, P. Lichtenstein, S. Sandin, and C.M. Hultman. **Fertility of first-degree relatives of patients with schizophrenia: A three generation perspective**. *Schizophrenia research*, **91**(1-3):238–245, 2007. 11, 14

[56] M. Baron, N. Risch, and J. Mendlewicz. **Differential fertility in bipolar affective illness**. *Journal of Affective Disorders*, **4**(2):103–112, 1982. 11

[57] K.E. Williams, W.K. Marsh, and N.L. Rasgon. **Mood disorders and fertility in women: a critical review of the literature and implications for future research**. *Human reproduction update*, **13**(6):607–616, 2007. 11

[58] R.B. King. **Subfecundity and anxiety in a nationally representative sample**. *Social science & medicine*, **56**(4):739–751, 2003. 11

[59] J. Haukka, J. Suvisaari, and J. Lönnqvist. **Fertility of patients with schizophrenia, their siblings, and the general population: a cohort study from 1950 to 1959 in Finland**. *American Journal of Psychiatry*, **160**(3):460–463, 2003. 11, 14

[60] M. Weissman, P. Wickramaratne, Y. Nomura, V. Warner, D. Pilowsky, and H. Verdeli. **Offspring of depressed parents: 20 years later**. *American Journal of Psychiatry*, **163**(6):1001–1008, 2006. 11

[61] R. Webb, K. Abel, A. Pickles, L. Appleby, S. King-Hele, and P. Mortensen. **Mortality risk among offspring of psychiatric inpatients: a population-based follow-up to early adulthood**. *American Journal of Psychiatry*, **163**(12):2170–2177, 2006. 11

[62] I.N. Bespalova and J.D. Buxbaum. **Disease susceptibility genes for autism**. *Annals of medicine*, **35**(4):274–281, 2003. 12

[63] S. Folstein and M. Rutter. **Infantile autism: a genetic study of 21 twin pairs**. *Journal of Child Psychology and Psychiatry*, **18**(4):297–321, 1977. 12

[64] T. Kieseppä, T. Partonen, J. Haukka, J. Kaprio, and J. Lönnqvist. **High concordance of bipolar I disorder in a nationwide sample of twins**. *American Journal of Psychiatry*, **161**(10):1814–1821, 2004. 12

[65] P. McGuffin, F. Rijsdijk, M. Andrew, P. Sham, R. Katz, and A. Cardno. **The heritability of bipolar affective disorder and the genetic relationship to unipolar depression**. *Archives of general psychiatry*, **60**(5):497, 2003. 12, 20

[66] P.F. Sullivan, K.S. Kendler, and M.C. Neale. **Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies**. *Archives of general psychiatry*, **60**(12):1187, 2003. 12

[67] C.M. Bulik, P.F. Sullivan, F. Tozzi, H. Furberg, P. Lichtenstein, and N.L. Pedersen. **Prevalence, heritability, and prospective risk factors for anorexia nervosa**. *Archives of general psychiatry*, **63**(3):305, 2006. 12

[68] P.F. Sullivan, M.C. Neale, and K.S. Kendler. **Genetic epidemiology of major depression: review and meta-analysis**. *American Journal of Psychiatry*, **157**(10):1552–1562, 2000. 12

[69] J.M. Hettema, M.C. Neale, and K.S. Kendler. **A review and meta-analysis of the genetic epidemiology of anxiety disorders**. *American Journal of Psychiatry*, **158**(10):1568–1578, 2001. 12

[70] J.F. Crow et al. **The origins, patterns and implications of human spontaneous mutation**. *Nature Reviews Genetics*, **1**(1):40–47, 2000. 12

[71] J.F. Crow. **Development: there's something curious about paternal-age effects**. *Science's STKE*, **301**(5633):606, 2003. 12

[72] A. Reichenberg, R. Gross, M. Weiser, M. Bresnahan, J. Silverman, S. Harlap, J. Rabinowitz, C. Shulman, D. Malaspina, G. Lubin, et al. **Advancing paternal age and autism**. *Archives of general psychiatry*, **63**(9):1026, 2006. 13, 17

[73] M.S. Durkin, M.J. Maenner, C.J. Newschaffer, L.C. Lee, C.M. Cunniff, J.L. Daniels, R.S. Kirby, L. Leavitt, L. Miller, W. Zahorodny, et al. **Advanced parental age and the risk of autism spectrum disorder**. *American Journal of Epidemiology*, **168**(11):1268–1276, 2008. 13

[74] CM Hultman, S. Sandin, SZ Levine, P. Lichtenstein, and A. Reichenberg. **Advancing paternal age and risk of autism: new evidence from a population-based study and a meta-analysis of epidemiological studies**. *Molecular psychiatry*, **16**(12):1203–1212, 2010. 13

[75] T.M. Laursen, T. Munk-Olsen, M. Nordentoft, M.P. Bo, et al. **A comparison of selected risk factors for unipolar depressive disorder, bipolar affective disorder, schizoaffective disorder, and schizophrenia from a danish population-based cohort.** *The Journal of clinical psychiatry*, **68**(11):1673, 2007. 13

[76] D. Malaspina, S. Harlap, S. Fennig, D. Heiman, D. Nahon, D. Feldman, and E.S. Susser. **Advancing paternal age and the risk of schizophrenia**. *Archives of general psychiatry*, **58**(4):361, 2001. 13, 17

[77] E.M. Frans, S. Sandin, A. Reichenberg, P. Lichtenstein, N. Langstrom, and C.M. Hultman. **Advancing paternal age and bipolar disorder**. *Archives of general psychiatry*, **65**(9):1034, 2008. 13

[78] B. Devlin and S.W. Scherer. **Genetic architecture in autism spectrum disorder**. *Current opinion in genetics & development*, 2012. 13

[79] G.V. Kryukov, L.A. Pennacchio, and S.R. Sunyaev. **Most rare missense alleles are deleterious in humans: implications for complex disease and association studies**. *The American Journal of Human Genetics*, **80**(4):727–739, 2007. 13, 17

[80] R. Sandberg, R. Yasuda, D.G. Pankratz, T.A. Carter, J.A. Del Rio, L. Wodicka, M. Mayford, D.J. Lockhart, and C. Barlow. **Regional and strain-specific gene expression mapping in the adult mouse brain**. *Proceedings of the National Academy of Sciences*, **97**(20):11038, 2000. 13

[81] M.C. Keller, G. Miller, et al. **Resolving the paradox of common, harmful, heritable mental disorders: Which evolutionary genetic models work best?** *Behavioral and Brain Sciences*, **29**(4):385–404, 2006. 13, 14, 16, 17

[82] M.D. Prokosch, R.A. Yeo, and G.F. Miller. **Intelligence tests with higher-loadings show higher correlations with body symmetry: Evidence for a general fitness factor mediated by developmental stability**. *Intelligence*, **33**(2):203–213, 2005. 13

[83] TJ Crow. **Aetiology of schizophrenia; an evolutionary theory**. *International Clinical Psychopharmacology*, **10**:49, 1995. 14

[84] R.M. Nesse et al. **Natural selection and the elusiveness of happiness**. *PHILOSOPHICAL TRANSACTIONS-ROYAL SOCIETY OF LONDON SERIES B BIOLOGICAL SCIENCES*, pages 1333–1348, 2004. 14

[85] JA Sherman. **Evolutionary origin of bipolar disorder (EOBD). Psycoloquy 12**, 2001. 14

[86] S. Guisinger. **Adapted to flee famine: Adding an evolutionary perspective on anorexia nervosa.** *Psychological Review*, **110**(4):745, 2003. 14

[87] Bio116. **Sickle Cell Anemia**, September 2012. 15

[88] O. Mayo et al. **The rise and fall of the common disease-common variant (CDCV) hypothesis: how the sickle cell disease paradigm led us all astray (or did it?).** *Twin research and human genetics: the official journal of the International Society for Twin Studies*, **10**(6):793, 2007. 16

[89] RM Corbo and R. Scacchi. **Apolipoprotein E (APOE) allele distribution in the world. Is APOE* 4 a 'thrifty' allele?** *Annals of human genetics*, **63**(4):301–310, 1999. 16

[90] T. Shibre, A. Negash, G. Kullgren, D. Kebede, A. Alem, A. Fekadu, D. Fekadu, G. Medhin, and L. Jacobsson. **Perception of stigma among family members of individuals with schizophrenia and major affective disorders in rural Ethiopia**. *Social Psychiatry and Psychiatric Epidemiology*, **36**(6):299–303, 2001. 16

[91] N.J. Dingemanse, C. Both, P.J. Drent, J.M. Tinbergen, et al. **Fitness consequences of avian personalities in a fluctuating environment**. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **271**(1541):847–852, 2004. 16

[92] S.R. Jaffee and TS Price. **Gene–environment correlations: a review of the evidence and implications for prevention of mental illness**. *Molecular psychiatry*, **12**(5):432–442, 2007. 17

[93] R. Uher and P. McGuffin. **The moderation by the serotonin transporter gene of environmental adversity in the aetiology of mental illness: review and methodological analysis**. *Molecular psychiatry*, **13**(2):131–146, 2007. 17

[94] K.S. Kendler, R.C. Kessler, E.E. Walters, C. MacLean, M.C. Neale, A.C. Heath, and L.J. Eaves. **Stressful Life Events, Genetic Liability, and Onset of an Episode of Major Depression in Women**. *Focus*, **8**(3):459, 2010. 17

[95] DJ MacIntyre, DHR Blackwood, DJ Porteous, BS Pickard, and WJ Muir. **Chromosomal abnormalities and mental illness**. *Molecular psychiatry*, **8**(3):275–287, 2003. 17

[96] S.E. Folstein, B. Rosen-Sheidley, et al. **Genetics of autism: complex aetiology for a heterogeneous disorder**. *Nature Reviews Genetics*, **2**(12):943–955, 2001. 17

[97] Jonathan Sebat, B Lakshmi, Dheeraj Malhotra, Jennifer Troge, Christa Lese-Martin, Tom Walsh, Boris Yamrom, Seungtai Yoon, Alex Krasnitz, Jude Kendall, and et al. **Strong association of de novo copy number mutations with autism**. *Science*, **316**(5823):445–449, 2007. 17, 38

[98] S.M. Monroe and M.W. Reid. **Gene-environment interactions in depression research**. *Psychological Science*, **19**(10):947, 2008. 20

[99] Y.H. Chen, H.W. Lin, and H. Liu. **Two-stage analysis for gene-environment interaction utilizing both case-only and family-based analysis**. *Genetic epidemiology*, **33**(2):95–104, 2009. 20

[100] C.E. Murcray, J.P. Lewinger, and W.J. Gauderman. **Gene-environment interaction in genome-wide association studies**. *American journal of epidemiology*, **169**(2):219–226, 2009. 20

[101] R.F. Krueger. **The structure of common mental disorders**. *Archives of General Psychiatry*, **56**(10):921, 1999. 20

[102] A.G. Cardno, F.V. Rijsdijk, P.C. Sham, R.M. Murray, and P. McGuffin. **A twin study of genetic relationships between psychotic symptoms**. *American Journal of Psychiatry*, **159**(4):539–545, 2002. 20

[103] K.S. Kendler, M.C. Neale, R.C. Kessler, A.C. Heath, and L.J. Eaves. **Major depression and generalized anxiety disorder: same genes,(partly) different environments?** *Archives of General Psychiatry*, **49**(9):716, 1992. 20

[104] Roger C Bland. **Psychiatric Disorders in America**. *Journal of Psychiatry and Neuroscience*, **17**(1):34–36, 1992. 22, 197

[105] M M Weissman, R C Bland, G J Canino, S Greenwald, H G Hwu, C K Lee, S C Newman, M A Oakley-Browne, M Rubio-Stipec, and P J Wickramaratne. **The cross national epidemiology of obsessive compulsive disorder. The Cross National Collaborative Group.** *The Journal of clinical psychiatry*, **55 Suppl**(0160-6689 LA - eng PT - Journal Article SB - IM):5–10, 1994. 22, 197

[106] M.M. Antony, F. Downie, and R.P. Swinson. **Diagnostic issues and epidemiology in obsessive-compulsive disorder**. *Obsessive-compulsive disorder: Theory, research and treatment*, pages 3–32, 1998. 22

[107] H. Matsunaga, K. Maebayashi, K. Hayashida, K. Okino, T. Matsui, T. Iketani, N. Kiriike, and D. Stein. **Symptom structure in Japanese patients with obsessive-compulsive disorder**. *American Journal of Psychiatry*, **165**(2):251–253, 2008. 22

[108] G Nestadt, J F Samuels, A J Romanoski, M F Folstein, and P R McHugh. **Obsessions and compulsions in the community.** *Acta Psychiatrica Scandinavica*, **89**(4):219–224, 1994. 22, 197

[109] G Nestadt, O J Bienvenu, G Cai, J Samuels, and W W Eaton. **Incidence of obsessive-compulsive disorder in adults.** *The Journal of nervous and mental disease*, **186**(7):401–406, 1998. 22, 197

[110] W. CORYELL. **Obsessive-compulsive disorder and primary unipolar depression: Comparisons of background, family history, course, and mortality**. *The Journal of Nervous and Mental Disease*, **169**(4):220, 1981. 22

[111] EH Hare, JS Price, and ETO Slater. **Fertility in obsessional neurosis**. *The British Journal of Psychiatry*, **121**(561):197–205, 1972. 22

[112] E Inouye. **Similar and Dissimilar Manifestations of Obsessive-Compulsive Neurosis in Monozygotic Twins.** *The American Journal of Psychiatry*, **121**:1171–1175, 1965. 22, 200

[113] Y. Wu, X. Liu, H. Luo, W. Deng, G. Zhao, Q. Wang, L. Zhang, X. Ma, X. Liu, R.A. Murray, et al. **Advanced paternal age increases the risk of schizophrenia and obsessive–compulsive disorder in a Chinese Han population**. *Psychiatry Research*, 2012. 22

[114] Dan Levy, Michael Ronemus, Boris Yamrom, Yoon-ha Lee, Anthony Leotta, Jude Kendall, Steven Marks, B Lakshmi, Deepa Pai, Kenny Ye, Andreas Buja, Abba Krieger, Seungtai Yoon, Jennifer Troge, Linda Rodgers, Ivan Iossifov, and Michael Wigler. **Rare de novo and transmitted copy-number variation in autistic spectrum disorders.** *Neuron*, **70**(5):886–897, June 2011. 26, 42, 43

[115] Stephan J Sanders, A Gulhan Ercan-Sencicek, Vanessa Hus, Rui Luo, Michael T Murtha, Daniel Moreno-De-Luca, Su H Chu, Michael P Moreau, Abha R Gupta, Susanne A Thomson, Christopher E Mason, Kaya Bilguvar, Patricia B S Celestino-Soper, Murim Choi, Emily L Crawford, Lea Davis, Nicole R Davis Wright, Rahul M Dhodapkar, Michael DiCola, Nicholas M DiLullo, Thomas V Fernandez, Vikram Fielding-Singh, Daniel O Fishman, Stephanie Frahm, Rouben Garagaloyan, Gerald S Goh, Sindhuja Kammela, Lambertus Klei, Jennifer K Lowe, Sabata C Lund, Anna D McGrew, Kyle A Meyer, William J Moffat, John D Murdoch, Brian J O'Roak, Gordon T Ober, Rebecca S Pottenger, Melanie J Raubeson, Youeun Song, Qi Wang, Brian L Yaspan, Timothy W Yu, Ilana R Yurkiewicz, Arthur L Beaudet, Rita M Cantor, Martin Curland, Dorothy E Grice, Murat Günel, Richard P Lifton, Shrikant M Mane, Donna M Martin, Chad A Shaw, Michael Sheldon, Jay A Tischfield, Christopher A Walsh, Eric M Morrow, David H Ledbetter, Eric Fombonne, Catherine Lord, Christa Lese Martin, Andrew I Brooks, James S Sutcliffe, Edwin H Cook, Daniel Geschwind, Kathryn Roeder, Bernie Devlin, and Matthew W State. **Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism.** *Neuron*, **70**(5):863–885, June 2011. 26, 42, 43

[116] C Lord, M Rutter, and A Le Couteur. **Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders.** *Journal of Autism and Developmental Disorders*, **24**(5):659–685, 1994. 26, 42

[117] C Lord, S Risi, L Lambrecht, E H Cook, B L Leventhal, P C DiLavore, A Pickles, and M Rutter. **The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism.** *Journal of Autism and Developmental Disorders*, **30**(3):205–223, 2000. 26, 42

[118] K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S.F.A. Grant, H. Hakonarson, and M. Bucan. **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, 2007. 26

[119] A.J. Iafrate, L. Feuk, M.N. Rivera, M.L. Listewnik, P.K. Donahoe, Y. Qi, S.W. Scherer, and C. Lee. **Detection of large-scale variation in the human genome**. *Nature genetics*, **36**(9):949–951, 2004. 27

[120] J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Månér, H. Massa, M. Walker, M. Chi, et al. **Large-scale copy number polymorphism in the human genome**. *Science*, **305**(5683):525–528, 2004. 27

[121] C.M. Smith, J.H. Finger, T.F. Hayamizu, I.J. McCright, J.T. Eppig, J.A. Kadin, J.E. Richardson, and M. Ringwald. **The mouse gene expression database (GXD): 2007 update**. *Nucleic acids research*, **35**(suppl 1):D618–D623, 2007. 28

[122] D. Näf, D.M. Krupke, J.P. Sundberg, J.T. Eppig, and C.J. Bult. **The Mouse Tumor Biology Database**. *Cancer research*, **62**(5):1235, 2002. 28

[123] A.V. Evsikov, M.E. Dolan, M.P. Genrich, E. Patek, and C.J. Bult. **MouseCyc: a curated biochemical pathways database for the laboratory mouse**. *Genome biology*, **10**(8):R84, 2009. 28

[124] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, and et al. **Gene Ontology : tool for the unification of biology. The Gene Ontology Consortium.** *Gene Expression*, **25**(1):25–29, 2000. 28, 102

[125] Kasper Lage, Niclas Tue Hansen, E Olof Karlberg, Aron C Eklund, Francisco S Roque, Patricia K Donahoe, Zoltan Szallasi, Thomas Skt Jensen, and Sren Brunak. **A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes.** *Proceedings of the National Academy of Sciences of the United States of America*, **105**(52):20870–5, 2008. 28, 102

[126] Kasper Lage, E Olof Karlberg, Zenia M Strling, Pll I Olason, Anders G Pedersen, Olga Rigina, Anders M Hinsby, Zeynep Tmer, Flemming Pociot, Niels Tommerup, and et al. **A human phenome-interactome network of protein complexes implicated in genetic disorders**. *Nature Biotechnology*, **25**(3):309–316, 2007. 28, 102

[127] The UCSC genome browser. **Schema for Conservation - Vertebrate Multiz Alignment & Conservation (46 Species)**, August 2012. 29, 312

[128] The ENCODE project. **Encyclopedia of DNA Elements**, August 2012. 29, 310

[129] V Matys, E Fricke, R Geffers, E Gssling, M Haubrock, R Hehl, K Hornischer, D Karas, A E Kel, O V Kel-Margoulis, and et al. **TRANSFAC®: transcriptional regulation, from patterns to profiles**. *Nucleic Acids Research*, **31**(1):374–378, 2003. 29, 231, 362, 405

[130] B. Soric. **Statistical "discoveries" and effect-size estimation**. *Journal of the American Statistical Association*, pages 608–610, 1989. 31

[131] Y Benjamini and Y Hochberg. **Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing Yoav Benjamini ; Yosef Hochberg Controlling the False Discovery Rate : a Practical and Powerful Approach to Multiple Testing**. *Society*, **57**(1):289–300, 1995. 31, 44

[132] W.S. Noble. **How does multiple testing correction work?** *Nature biotechnology*, **27**(12):1135–1137, 2009. 31

[133] K. Strimmer. **A unified approach to false discovery rate estimation**. *Bmc Bioinformatics*, **9**(1):303, 2008. 31

[134] A.R. Quinlan and I.M. Hall. **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics*, **26**(6):841–842, 2010. 32

[135] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W De Bakker, Mark J Daly, and et al. **PLINK: a toolset for whole genome association and population-based linkage analyses**. *Genome Research*, **81**(3):559–75, 2007. 32, 243

[136] Thayne L Sweeten, David J Posey, and Christopher J McDougle. **Brief report: autistic disorder in three children with cytomegalovirus infection.** *Journal of Autism and Developmental Disorders*, **19**(3):583–586, 1989. 35

[137] E Fombonne, C Du Mazaubrun, C Cans, and H Grandjean. **Autism and associated medical disorders in a French epidemiological survey.** *Journal of the American Academy of Child & Adolescent Psychiatry*, **36**(11):1561–1569, 1997. 35

[138] C M Hultman, P Sparen, and S Cnattingius. **Perinatal risk factors for infantile autism**. *Epidemiology*, **13**(4):417–423, 2002. 35

[139] Jeremy Veenstra-Vanderweele, Susan L Christian, and Edwin H Cook. **Autism as a paradigmatic complex genetic disorder.** *Annual Review of Genomics and Human Genetics*, **5**(2):379–405, 2004. 35

[140] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-IV-TR.* New York American Psychiatric Press Inc, 2000. 35, 54, 197, 202, 323

[141] Daniel H Geschwind. **Advances in Autism**. *Annual review of medicine*, **60**(1):367–380, February 2009. 35, 37

[142] Understanding Autism, Spectrum Disorders, and Disabilities Monitoring Network. **Autism and Developmental Disabilities Monitoring Network**. *Prevention*, **58**(Cdc):1991–1991, 2008. 35

[143] E Fombonne, P Bolton, J Prior, H Jordan, and M Rutter. **A family study of autism: cognitive patterns and levels in parents and siblings.** *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, **38**(6):667–683, 1997. 36

[144] J Piven, P Palmer, D Jacobi, D Childress, and S Arndt. **Broader autism phenotype: evidence from a family history study of multiple-incidence autism families.** *The American Journal of Psychiatry*, **154**(2):185–90, 1997. 36

[145] John N Constantino, Clara Lajonchere, Marin Lutz, Teddi Gray, Anna Abbacchi, Kathleen McKenna, Deepti Singh, and Richard D Todd. **Autistic social impairment in the siblings of children with pervasive developmental disorders.** *The American Journal of Psychiatry*, **163**(2):294–296, 2006. 36

[146] P Bolton, H Macdonald, A Pickles, P Rios, S Goode, M Crowson, A Bailey, and M Rutter. **A case-control family history study of autism.** *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, **35**(5):877–900, 1994. 36

[147] A Pickles, E Starr, S Kazak, P Bolton, K Papanikolaou, A Bailey, R Goodman, and M Rutter. **Variable expression of the autism broader phenotype: findings from extended pedigrees.** *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, **41**(4):491–502, 2000. 36

[148] A Le Couteur, A Bailey, S Goode, A Pickles, S Robertson, I Gottesman, and M Rutter. **A broader phenotype of autism: the clinical spectrum in twins.** *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, **37**(7):785–801, 1996. 36

[149] C Gillberg and E Billstedt. **Autism and Asperger syndrome: coexistence with other clinical disorders.** *Acta Psychiatrica Scandinavica*, **102**(5):321–30, 2000. 37

[150] Robin L Hansen, Sally Ozonoff, Paula Krakowiak, Kathleen Angkustsiri, Carrie Jones, Lesley J Deprey, Dung-Nghi Le, Lisa A Croen, and Irva Hertz-Picciotto. **Regression in autism: prevalence and associated factors in the CHARGE Study.** *Ambulatory pediatrics the official journal of the Ambulatory Pediatric Association*, **8**(1):25–31, 2008. 37

[151] E Simonoff, A Pickles, T Charman, S Chandler, T Loucas, and G Baird. **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 & Adolescent Psychiatry*, **47**(8):921–929, 2008. 37

[152] Michael Brimacombe, Janti Chaaban, Barbie Zimmerman-Bier, and with Xue Ming Wagner, George C. **Autism spectrum disorders: concurrent clinical disorders.** *Journal of Child Neurology*, **23**(1):6–13, 2008. 37

[153] Susan R Leekam, Carmen Nieto, Sarah J Libby, Lorna Wing, and Judith Gould. **Describing the sensory abnormalities of children and adults with autism.** *Journal of Autism and Developmental Disorders*, **37**(5):894–910, 2007. 37

[154] Eva M Jansiewicz, Melissa C Goldberg, Craig J Newschaffer, Martha B Denckla, Rebecca Landa, and Stewart H Mostofsky. **Motor signs distinguish children with high functioning autism and Asperger's syndrome from controls.** *Journal of Autism and Developmental Disorders*, **36**(5):613–621, 2006. 37

[155] Xue Ming, Michael Brimacombe, and George C Wagner. **Prevalence of motor impairment in autism spectrum disorders.** *Brain & Development*, **29**(9):565–570, 2007. 37

[156] Paula Krakowiak, Beth Goodlin-Jones, Irva Hertz-Picciotto, Lisa A Croen, and Robin L Hansen. **Sleep problems in children with autism spectrum disorders, developmental delays, and typical development: a population-based study.** *Journal of Sleep Research*, **17**(2):197–206, 2008. 37

[157] Beth A Malow, Mary L Marzec, Susan G McGrew, Lily Wang, Lynnette M Henderson, and Wendy L Stone. **Characterizing sleep in children with autism spectrum disorders: a multidimensional approach.** *Sleep (Rochester)*, **29**(12):1563–1571, 2006. 37

[158] R Tuchman and I Rapin. **Epilepsy in autism**. *Lancet Neurolog*, **1**(6):352–358, 2002. 37

[159] F R Volkmar and D S Nelson. **Seizure disorders in autism.** *Journal of the American Academy of Child & Adolescent Psychiatry*, **29**(1):127–129, 1990. 37

[160] A Bailey, A Le Couteur, I Gottesman, P Bolton, E Simonoff, E Yuzda, and M Rutter. **Autism as a strongly genetic disorder: evidence from a British twin study.** *Psychological Medicine*, **25**(1):63–77, 1995. 38

[161] Brett S Abrahams and Daniel H Geschwind. **Advances in autism genetics: on the threshold of a new neurobiology.** *Nature Reviews Genetics*, **9**(5):341–355, 2008. 38

[162] M-L Jacquemont, D Sanlaville, R Redon, O Raoul, V Cormier-Daire, S Lyonnet, J Amiel, M Le Merrer, D Heron, M-C De Blois, and et al. **Array based comparative genomic hybridisation identifies high frequency of cryptic chromosomal rearrangements in patients with syndromic autism spectrum disorders**. *Journal of Medical Genetics*, **43**(11):843–849, 2006. 38

[163] L A Weiss, Y Shen, J M Korn, D E Arking, D T Miller, R Fossdal, E Saemundsen, H Stefansson, M A R Ferreira, and T Green. **Association between microdeletion and microduplication at 16p11.2 and autism**. *New England Journal of Medicine*, **358**(7):667, 2008. 38

[164] Peter Szatmari, Andrew D Paterson, Lonnie Zwaigenbaum, Wendy Roberts, Jessica Brian, Xiao-Qing Liu, John B Vincent, Jennifer L Skaug, Ann P Thompson, Lili Senman, and et al. **Mapping autism risk loci using genetic linkage and chromosomal rearrangements.** *Nature Genetics*, **39**(3):319–328, 2007. 38

[165] Jill L Silverman, Mu Yang, Catherine Lord, and Jacqueline N Crawley. **Behavioural phenotyping assays for mouse models of autism.** *Nature Reviews Neuroscience*, **11**(7):490–502, 2010. 38

[166] Ml Terranova and Giovanni Laviola. **Scoring of Social Interactions and Play in Mice During Adolescence**. *Current Protocols in Toxicology*, **13.10**:1–11, 2005. 39

[167] C S Carter, J R Williams, D M Witt, and T R Insel. **Oxytocin and social bonding.** *Annals Of The New York Academy Of Sciences*, **652**:204–211, 1992. 39, 40

[168] Hiroyuki Arakawa, Keiko Arakawa, D Caroline Blanchard, and Robert J Blanchard. **A new test paradigm for social recognition evidenced by urinary scent marking behavior in C57BL/6J mice.** *Behavioural Brain Research*, **190**(1):97–104, 2008. 39

[169] Hiroyuki Arakawa, D Caroline Blanchard, Keiko Arakawa, Christopher Dunlap, and Robert J Blanchard. **Scent marking behavior as an odorant communication in mice.** *Neuroscience & Biobehavioral Reviews*, **32**(7):1236–1248, 2008. 39

[170] K A Miczek, S C Maxson, E W Fish, and S Faccidomo. **Aggressive behavioral phenotypes in mice**. *Behavioural Brain Research*, **125**(1-2):167–181, 2001. 39

[171] Valerie J Bolivar, Samantha R Walters, and Jennifer L Phoenix. **Assessing autism-like behavior in mice: variations in social interactions among inbred strains.** *Behavioural Brain Research*, **176**(1):21–26, 2007. 39

[172] H G McFarlane, G K Kusek, M Yang, J L Phoenix, V J Bolivar, and J N Crawley. **Autism-like behavioral phenotypes in BTBR T+tf/J mice.** *Genes brain and behavior*, **7**(2):152–163, 2008. 39, 41

[173] Mu Yang, Andrew M Clarke, and Jacqueline N Crawley. **Postnatal lesion evidence against a primary role for the corpus callosum in mouse sociability.** *European Journal of Neuroscience*, **29**(8):1663–1677, 2009. 39

[174] D C Blanchard and R J Blanchard. **Ethoexperimental approaches to the biology of emotion.** *Annual Review of Psychology*, **39**:43–68, 1988. 39

[175] Kathryn K Chadman, Shiaoching Gong, Maria L Scattoni, Sarah E Boltuck, Shruti U Gandhy, Nathaniel Heintz, and Jacqueline N Crawley. **Minimal aberrant behavioral phenotypes of neuroligin-3 R451C knockin mice.** *Autism research official journal of the International Society for Autism Research*, **1**(3):147–158, 2008. 39

[176] S S Moy, J J Nadler, A Perez, R P Barbaro, J M Johns, T R Magnuson, J Piven, and J N Crawley. **Sociability and preference for social novelty in five inbred strains: an approach to assess autistic-like behavior in mice.** *Genes brain and behavior*, **3**(5):287–302, 2004. 39

[177] J J Nadler, S S Moy, G Dold, D Trang, N Simmons, A Perez, N B Young, R P Barbaro, J Piven, T R Magnuson, and et al. **Automated apparatus for quantitation of social approach behaviors in mice**. *Genes Brain and Behavior*, **3**(5):303–314, 2004. 39

[178] Mu Yang, Michael D Weber, and Jacqueline N Crawley. **Light Phase Testing of Social Behaviors: Not a Problem**. *Frontiers in neuroscience*, **2**(2):186–191, 2008. 39

[179] M L Scattoni, H G McFarlane, V Zhodzishsky, H K Caldwell, W S Young, L Ricceri, and J N Crawley. **Reduced ultrasonic vocalizations in vasopressin 1b knockout mice.** *Behavioural Brain Research*, **187**(2):371–378, 2008. 40

[180] Jules B Panksepp, Kimberly A Jochman, Joseph U Kim, Jamie J Koy, Ellie D Wilson, Qiliang Chen, Clarinda R Wilson, and Garet P Lahvis. **Affiliative Behavior, Ultrasonic Communication and Social Reward Are Influenced by Genetic Variation in Adolescent Mice**. *PLoS ONE*, **2**(4):13, 2007. 40

[181] Maria Luisa Scattoni, Shruti U Gandhy, Laura Ricceri, and Jacqueline N Crawley. **Unusual Repertoire of Vocalizations in the BTBR T+tf/J Mouse Model of Autism**. *PLoS ONE*, **3**(8):15, 2008. 40

[182] J Bakker, S Honda, N Harada, and J Balthazart. % bf Sexual partner preference requires a functional aromatase (cyp19) gene in male mice. *Hormones and Behavior*, **42**(2):158–171, 2002. 40

[183] Miranda M Lim, Zuoxin Wang, Daniel E Olazabal, Xianghui Ren, Ernest F Terwilliger, and Larry J Young. **Enhanced partner preference in a promiscuous species by manipulating the expression of a single gene**. *Nature*, **429**(6993):754–757, 2004. 40

[184] J N Ferguson, J M Aldag, T R Insel, and L J Young. **Oxytocin in the medial amygdala is essential for social recognition in the mouse.** *Journal of Neuroscience*, **21**(20):8278–8285, 2001. 40

[185] Todd H Ahern, Meera E Modi, James P Burkett, and Larry J Young. **Evaluation of two automated metrics for analyzing partner preference tests.** *Journal of Neuroscience Methods*, **182**(2):180–188, 2009. 40

[186] M A Hofer and H N Shair. **Ultrasonic vocalization, laryngeal braking, and thermogenesis in rat pups: a reappraisal.** *Behavioral neuroscience*, **107**(2):354–362, April 1993. 40

[187] Haoran Wang, Shuyin Liang, Jeffrey Burgdorf, Jurgen Wess, and John Yeomans. **Ultrasonic vocalizations induced by sex and amphetamine in M2, M4, M5 muscarinic and D2 dopamine receptor knockout mice.** *PloS one*, **3**(4):e1893, 2008. 40

[188] I Branchi, D Santucci, and E Alleva. **Ultrasonic vocalisation emitted by infant rodents: a tool for assessment of neurobehavioural development.** *Behavioural brain research*, **125**(1-2):49–56, November 2001. 40

[189] Timothy E Holy and Zhongsheng Guo. **Ultrasonic songs of male mice.** *PLoS biology*, **3**(12):e386, December 2005. 40

[190] Anna Moles, Fabrizio Costantini, Luciana Garbugino, Claudio Zanettini, and Francesca R D'Amato. **Ultrasonic vocalizations emitted during dyadic interactions in female mice: a possible index of sociability?** *Behavioural brain research*, **182**(2):223–230, September 2007. 40

[191] Maria Luisa Scattoni, Jacqueline Crawley, and Laura Ricceri. **Ultrasonic vocalizations: a tool for behavioural phenotyping of mouse models of neurodevelopmental disorders.** *Neuroscience and biobehavioral reviews*, **33**(4):508–515, April 2009. 40

[192] Hiroyuki Arakawa, Keiko Arakawa, D Caroline Blanchard, and Robert J Blanchard. **A new test paradigm for social recognition evidenced by urinary scent marking behavior in C57BL/6J mice.** *Behavioural brain research*, **190**(1):97–104, June 2008. 40

[193] Hiroyuki Arakawa, D Caroline Blanchard, Keiko Arakawa, Christopher Dunlap, and Robert J Blanchard. **Scent marking behavior as an odorant communication in mice.** *Neuroscience and biobehavioral reviews*, **32**(7):1236–1248, September 2008. 40

[194] Daniel W Wesson, Matthieu Keller, Quentin Douhard, Michael J Baum, and Julie Bakker. **Enhanced urinary odor discrimination in female aromatase knockout (ArKO) mice.** *Hormones and behavior*, **49**(5):580–586, May 2006. 40

[195] S R Wersinger, H K Caldwell, L Martinez, P Gold, S-B Hu, and W S Young. **Vasopressin 1a receptor knockout mice have a subtle olfactory deficit but normal aggression.** *Genes, brain, and behavior*, **6**(6):540–551, August 2007. 40

[196] M Wöhr, F I Roullet, and J N Crawley. **Reduced scent marking and ultrasonic vocalizations in the BTBR T+tf/J mouse model of autism.** *Genes, brain, and behavior*, **10**(1):35–43, February 2011. 40

[197] Conor M Stack, Maria A Lim, Katrina Cuasay, Madeleine M Stone, Kimberly M Seibert, Irit Spivak-Pohis, Jacqueline N Crawley, James A Waschek, and Joanna M Hill. **Deficits in social behavior and reversal learning are more prevalent in male offspring of VIP deficient female mice.** *Experimental neurology*, **211**(1):67–84, May 2008. 40

[198] Mu Yang and Jacqueline N Crawley. **Simple behavioral assessment of mouse olfaction.** *Current protocols in neuroscience / editorial board, Jacqueline N. Crawley ... [et al.]*, **Chapter 8**:Unit 8.24, July 2009. 40

[199] Maria Luisa Scattoni, Jacqueline Crawley, and Laura Ricceri. **Ultrasonic vocalizations: a tool for behavioural phenotyping of mouse models of neurodevelopmental disorders.** *Neuroscience and Biobehavioral Reviews*, **33**(4):508–15, 2009. 40, 58

[200] Bryce C Ryan, Nancy B Young, Jacqueline N Crawley, James W Bodfish, and Sheryl S Moy. **Social deficits, stereotypy and early emergence of repetitive behavior in the C58/J inbred mouse strain.** *Behavioural Brain Research*, **208**(1):178–188, 2010. 41

[201] L W Fitzgerald, A K Ratty, M Teitler, K W Gross, and S D Glick. **Specificity of behavioral and neurochemical dysfunction in the chakragati mouse: a novel genetic model of a movement disorder.** *Brain Research*, **608**(2):247–258, 1993. 41

[202] Mark H Lewis, Yoko Tanimura, Linda W Lee, and James W Bodfish. **Animal models of restricted repetitive behavior in autism.** *Behavioural Brain Research*, **176**(1):66–74, 2007. 41

[203] Jeffrey M Welch, Jing Lu, Ramona M Rodriguiz, Nicholas C Trotta, Joao Peca, Jin-Dong Ding, Catia Feliciano, Meng Chen, J Paige Adams, Jianhong Luo, and et al. **Cortico-striatal synaptic defects and OCD-like behaviours in Sapap3-mutant mice.** *Nature*, **448**(7156):894–900, 2007. 41, 200, 203

[204] Sheryl S Moy, Jessica J Nadler, Michele D Poe, Randal J Nonneman, Nancy B Young, Beverly H Koller, Jacqueline N Crawley, Gary E Duncan, and James W Bodfish. **Development of a mouse test for repetitive, restricted behaviors: relevance to autism.** *Behavioural Brain Research*, **188**(1):178–194, 2008. 41

[205] Ayling S Ballester B Beal K Bragin E et al. Hubbard TJ, Aken BL. **Ensembl 2009**. *Nucleic Acids Res*, **37**:D690–697, 2009. 43

[206] Janan T Eppig, Carol J Bult, James A Kadin, Joel E Richardson, and Judith A Blake. **The Mouse Genome Database (MGD): from genes to micea community resource for mouse biology**. *Nucleic Acids Research*, **33**(Database Issue):D471–D475, 2005. 43

[207] Cynthia L Smith, Carroll-Ann W Goldsmith, and Janan T Eppig. **The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information**. *Genome Biology*, **6**(1):R7, 2005. 43, 60

[208] Caleb Webber, Jayne Y Hehir-Kwa, Duc-Quang Nguyen, Bert B A de Vries, Joris A Veltman, and Chris P Ponting. **Forging Links between Human Mental Retardation–Associated CNVs and Mouse Gene Knockout Models**. *PLoS genetics*, 2009. 44, 46, 59

[209] G Baird, E. Simonoff, A. Pickles, Loucas T. Chandler, S., D. Meldrum, and T Charman. **Prevalence of disorders of the autism spectrum in a population cohort of children in South Thames: the Special Needs and Autism Project (SNAP).** *Lancet*, **368**:210–215, 2006. 54

[210] D. Green, T. Charman, A. Pickles, S. Chandler, T. Loucas, E. Simonoff, and G. Baird. **Impairment in movement skills of children with autistic spectrum disorders.** *Dev Med Child Neurol*, **51**:311–316, 2009. 55

[211] X. Ming, M. Brimacombe, and G.C. Wagner. **Prevalence of motor impairment in autism spectrum disorders.** *Brain Dev*, **29**:565–570, 2007. 55

[212] K.L. Staples and G. Reid. **Fundamental movement skills and autism spectrum disorders.** *J Autism Dev Disord*, **40**:209–217, 2010. 55

[213] P Giovanardi Rossi, A Posar, and A Parmeggiani. **Epilepsy in adolescents and young adults with autistic disorder.** *Brain and Development*, **22**(2):102–106, 2000. 55

[214] R F Tuchman, I Rapin, and S Shinnar. **Autistic and dysphasic children. II: Epilepsy.** *Pediatrics*, **88**(6):1219–1225, 1991. 55

[215] M L Cepeda, Y M Yang, C C Price, and A Shah. **Psychiatric disorders in children and adolescents with mental retardation and active epilepsy.** *Archives of Neurology*, **90**(3):284–287, 1996. 55

[216] Ivy M Andersen, Joanna Kaczmarska, Susan G McGrew, and Beth A Malow. **Anxiety disorders in children and adolescents with autistic spectrum disorders: a meta-analysis.** *Clinical Child and Family Psychology Review*, **23**(3):482–485, 2011. 55

[217] F W Larsen and S E Mouridsen. **The outcome in children with childhood autism and Asperger syndrome originally diagnosed as psychotic. A 30-year follow-up study of subjects hospitalized as children.** *European child adolescent psychiatry*, **6**(4):181–190, 1997. 55

[218] A Klin. **Auditory brainstem responses in autism: brainstem dysfunction or peripheral hearing loss?** *Journal of Autism and Developmental Disorders*, **23**(1):15–35, 1993. 55

[219] Magdalini Hitoglou, Athina Ververi, Alexandros Antoniadis, and Dimitrios I Zafeiriou. **Childhood autism and auditory system abnormalities.** *Pediatric Neurology*, **42**(5):309–314, 2010. 55

[220] U Rosenhall, V Nordin, M Sandstrm, G Ahlsn, and C Gillberg. **Autism and hearing loss.** *Journal of Autism and Developmental Disorders*, **29**(5):349–357, 1999. 55

[221] Erika A Bosman, Andrew C Penn, John C Ambrose, Ross Kettleborough, Derek L Stemple, and Karen P Steel. **Multiple mutations in mouse Chd7 provide models for CHARGE syndrome.** *Human Molecular Genetics*, **14**(22):3463–3476, 2005. 55

[222] Francesca Vitelli, Antonella Viola, Masae Morishima, Tiziano Pramparo, Antonio Baldini, and Elizabeth Lindsay. **TBX1 is required for inner ear morphogenesis.** *Human Molecular Genetics*, **12**(16):2041–2048, 2003. 55

[223] J Kim, E Y Kim, J S Lee, W S Lee, and H N Kim. **Temporal bone CT findings in Cornelia de Lange syndrome.** *Ajnr American Journal Of Neuroradiology*, **29**(3):569–573, 2008. 55

[224] C. Scheiber, A. Hirschfelder, S. Grbel, H. Peters, and H. Olze. **Bilateral cochlear implantation in children with Noonan syndrome.** *Int J Pediatr Otorhinolaryngol*, 2009. 55

[225] Amy Isaacs Donaldson, Krista S Heavner, and Teresa A Zwolan. **Measuring progress in children with autism spectrum disorder who have cochlear implants.** *Archives of otolaryngologyhead neck surgery*, **130**(5):666–671, 2004. 55

[226] Thomas Bourgeron. **A synaptic trek to autism.** *Current Opinion in Neurobiology*, **19**(2):231–234, 2009. 56, 114

[227] X Gai, H M Xie, J C Perin, N Takahashi, K Murphy, A S Wenocur, M Darcy, R J OHara, E Goldmuntz, D E Grice, and et al. **Rare structural variation of synapse and neurotransmission genes in autism.** *Molecular Psychiatry*, **17**:1–10, 2011. 56, 101, 114

[228] Paul H Patterson. **Modeling autistic features in animals.** *Pediatric Research*, **69**(5 Pt 2):34R–40R, 2011. 56

[229] S E P Smith, Y D Zhou, G Zhang, Z Jin, D C Stoppel, and M P Anderson. **Increased Gene Dosage of Ube3a Results in Autism Traits and Decreased Glutamate Synaptic Transmission in Mice**. *Science Translational Medicine*, **3**(103):103ra97–103ra97, 2011. 57, 58

[230] G Horev, J Ellegood, J P Lerch, Y E E Son, L Muthuswamy, H Vogel, A M Krieger, A Buja, R M Henkelman, M Wigler, and et al. **Dosage-dependent phenotypes in models of 16p11.2 lesions found in autism**. *Proceedings of the National Academy of Sciences*, pages 1–6, 2011. 57

[231] A C Reichelt, R J Rodgers, and S J Clapcote. **The role of neurexins in schizophrenia and autistic spectrum disorder.** *Neuropharmacology*, pages 1–8, 2011. 57

[232] B. Papp, C. Pál, and L.D. Hurst. **Dosage sensitivity and the evolution of gene families in yeast**. *Nature*, **424**(6945):194–197, 2003. 57

[233] N. Zheng, P. Wang, P.D. Jeffrey, and N.P. Pavletich. **Structure of a c-Cbl–UbcH7 complex: RING domain function in ubiquitin-protein ligases**. *Cell*, **102**(4):533–539, 2000. 58

[234] R.S. Kang, C.M. Daniels, S.A. Francis, S.C. Shih, W.J. Salerno, L. Hicke, and I. Radhakrishnan. **Solution structure of a CUE-ubiquitin complex reveals a conserved mode of ubiquitin binding**. *Cell*, **113**(5):621–630, 2003. 58

[235] Akira Uematsu, Takefumi Kikusui, Takashi Kihara, Toshihiro Harada, Masahiro Kato, Kiyoshi Nakano, Osamu Murakami, Nobuyoshi Koshida, Yukari Takeuchi, and Yuji Mori. **Maternal approaches to pup ultrasonic vocalizations produced by a nanocrystalline silicon thermo-acoustic emitter.** *Brain Research*, **1163**:91–99, 2007. 58

[236] Q Xu, B Modrek, and C Lee. **Genome-wide detection of tissue-specific alternative splicing in the human transcriptome.** *Nucleic Acids Research*, **30**(17):3754–3766, 2002. 59

[237] Gene Yeo, Dirk Holste, Gabriel Kreiman, and Christopher B Burge. **Variation in alternative splicing across human tissues**. *Genome Biology*, **5**(10):R74, 2004. 59

[238] Kathryn K Chadman, Mu Yang, and Jacqueline N Crawley. **Criteria for validating mouse models of psychiatric diseases.** *American journal of medical genetics Part B Neuropsychiatric genetics the official publication of the International Society of Psychiatric Genetics*, **150B**(1):1–11, 2009. 60

[239] C Webber. **Functional enrichment analysis with structural variants: pitfalls and strategies.** *Cytogenetic and Genome Research*, **135**(3-4):277–85, 2011. 61, 62

[240] Purvesh Khatri, Marina Sirota, and Atul J Butte. **Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges**. *PLoS Computational Biology*, **8**(2):e1002375, 2012. 97, 99

[241] Sheng Zhong, Lu Tian, Cheng Li, K F Storch, and W H Wong. **Comparative analysis of gene sets in the gene ontology space under the multiple hypothesis testing framework**. *Proceedings IEEE Computational Systems Bioinformatics Conference CSB IEEE Computational Systems Bioinformatics Conference*, **14**(Csb):425–435, 2004. 98

[242] Ftima Al-Shahrour, Ramn Daz-Uriarte, and Joaqun Dopazo. **Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information.** *Bioinformatics*, **21**(13):2988–2993, 2005. 99

[243] J J Goeman, S A Van De Geer, F De Kort, and H C Van Houwelingen. **A global test for groups of genes: testing association with a clinical outcome**. *Bioinformatics*, **20**(1):93–99, 2004. 99

[244] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and et al. **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(43):15545–15550, 2005. 99

[245] Lu Tian, Steven A Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S Kohane, and Peter J Park. **Discovering statistically significant pathways in expression profiling studies**. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(38):13544–9, 2005. 99

[246] Seon-Young Kim and David J Volsky. **PAGE: Parametric Analysis of Gene Set Enrichment**. *BMC Bioinformatics*, **6**(1):144, 2005. 99

[247] Marit Ackermann and Korbinian Strimmer. **A general modular framework for gene set enrichment analysis**. *BMC Bioinformatics*, **10**(1):47, 2009. 99

[248] Zhen Jiang and Robert Gentleman. **Extensions to gene set enrichment.** *Bioinformatics*, **23**(3):306–313, 2007. 99

[249] William T Barry, Andrew B Nobel, and Fred A Wright. **Significance analysis of functional categories in gene expression studies: a structured permutation approach.** *Bioinformatics*, **21**(9):1943–1949, 2005. 99

[250] Bradley Efron and Robert Tibshirani. **On testing the significance of sets of genes**. *Annals of Applied Statistics*, **1**(1):107–129, 2006. 99

[251] Jelle J Goeman and Peter Bhlmann. **Analyzing gene expression data in terms of gene sets: methodological issues.** *Bioinformatics*, **23**(8):980–987, 2007. 99

[252] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. **The KEGG resource for deciphering the genome**. *Nucleic Acids Research*, **32**(Database issue):D277–D280, 2004. 100

[253] G Joshi-Tope, M Gillespie, I Vastrik, P DEustachio, E Schmidt, B De Bono, B Jassal, G R Gopinath, G R Wu, L Matthews, and et al. **Reactome: a knowledgebase of biological pathways**. *Nucleic Acids Research*, **33**(Database Issue):D428–D432, 2005. 100

[254] P D Karp, M Riley, S M Paley, and A Pellegrini-Toole. **The MetaCyc Database**. *Nucleic Acids Research*, **30**(1):59–61, 2002. 100

[255] A M Huerta, H Salgado, D Thieffry, and J Collado-Vides. **RegulonDB: a database on transcriptional regulation in Escherichia coli.** *Nucleic Acids Research*, **26**(1):55–59, 1998. 100

[256] Joruhrer Rahnenf, Francisco S Domingues, Jochen Maydt, and Thomas Lengauer. **Calculating the Statistical Significance of Changes in Pathway Activity From Gene Expression Data**. *Statistical Applications in Genetics and Molecular Biology*, **3**(1), 2004. 100

[257] Sorin Draghici, Purvesh Khatri, Adi Laurentiu Tarca, Kashyap Amin, Arina Done, Calin Voichita, Constantin Georgescu, and Roberto Romero. **A systems biology approach for pathway level analysis.** *Genome Research*, **17**(10):1537–1545, 2007. 100

[258] Ali Shojaie and George Michailidis. **Analysis of Gene Sets Based on the Underlying Regulatory Network**. *Journal of computational biology a journal of computational molecular cell biology*, **16**(3):407–426, 2009. 100

[259] Anna Bauer-Mehren, Laura I Furlong, and Ferran Sanz. **Pathway databases and tools for their exploitation: benefits, current limitations and challenges**. *Molecular Systems Biology*, **5**(290):290, 2009. 100

[260] Saumyendra N Basu, Ravi Kollu, and Sharmila Banerjee-Basu. **AutDB: a gene reference resource for autism research**. *Nucleic Acids Research*, **37**(Database issue):D832–D836, 2009. 101, 203

[261] Hisham Al-Mubaid and Rajit K Singh. **A text-mining technique for extracting gene-disease associations from the biomedical literature.** *International Journal of Bioinformatics Research and Applications*, **6**(3):270–286, 2010. 101

[262] C Cano, T Monaghan, A Blanco, D P Wall, and L Peshkin. **Collaborative text-annotation resource for disease-centered relation extraction from biomedical text.** *Journal of Biomedical Informatics*, **42**(5):967–977, 2009. 101

[263] Ingrid Petric, Tanja Urbancic, Bojan Cestnik, and Marta Macedoni-Luksic. **Literature mining method RaJoLink for uncovering relations between biomedical concepts.** *Journal of Biomedical Informatics*, **42**(2):219–227, 2009. 101

[264] Yasunari Sakai, Chad A Shaw, Brian C Dawson, Diana V Dugas, Zaina Al-Mohtaseb, David E Hill, and Huda Y Zoghbi. **Protein interactome reveals converging molecular pathways among autism disorders.** *Science translational medicine*, **3**(86):86ra49, 2011. 101, 114

[265] Terrence F Meehan, Christopher J Carr, Jeremy J Jay, Carol J Bult, Elissa J Chesler, and Judith A Blake. **Autism candidate genes via mouse phenomics.** *Journal of Biomedical Informatics*, **44 Suppl 1**(March):1–7, 2011. 101

[266] Irina Voineagu, Xinchen Wang, Patrick Johnston, Jennifer K Lowe, Yuan Tian, Steve Horvath, Jonathan Mill, Rita M Cantor, Benjamin J Blencowe, and Daniel H Geschwind. **Transcriptomic analysis of autistic brain reveals convergent molecular pathology.** *Nature*, **474**(7351):380–384, 2011. 101

[267] Santitham Prom-On, Atthawut Chanthaphan, Jonathan Hoyin Chan, and Asawin Meechai. **Enhancing biological relevance of a weighted gene co-expression network for functional module identification.** *Journal of Bioinformatics and Computational Biology*, **9**(1):111–129, 2011. 101

[268] Eyal Ben-David and Sagiv Shifman. **Networks of Neuronal Genes Affected by Common and Rare Variants in Autism Spectrum Disorders**. *PLoS Genetics*, **8**(3):e1002556, 2012. 101

[269] Y A N Kou, Catalina Betancur, Huilei Xu, Joseph D Buxbaum, and A V I M A Ayan. **Network- and Attribute-Based Classifiers can Prioritize Genes and Pathways for Autism Spectrum Disorders and Intellectual Disability**. *American Journal of Medical Genetics*, **142**(April):130–142, 2012. 101

[270] Stephan J Sanders, Michael T Murtha, Abha R Gupta, John D Murdoch, Melanie J Raubeson, A Jeremy Willsey, A Gulhan Ercan-Sencicek, Nicholas M DiLullo, Neelroop N Parikshak, Jason L Stein, and et al. **De novo mutations revealed by whole-exome sequencing are strongly associated with autism**. *Nature*, **advance on**:1–6, 2012. 101

[271] Fran Supek, Matko Bonjak, Nives kunca, and Tomislav muc. **REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms**. *PLoS ONE*, **6**(7):9, 2011. 102, 108, 238

[272] Elizabeth J Rossin, Kasper Lage, Soumya Raychaudhuri, Ramnik J Xavier, Diana Tatar, Yair Benita, Chris Cotsapas, and Mark J Daly. **Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology**. *PLoS Genetics*, **7**(1):13, 2011. 102, 103

[273] P Shannon, A Markiel, O Ozier, N S Baliga, J T Wang, D Ramage, N Amin, B Schwikowski, and T Ideker. **Cytoscape: software environment for integrated models of biomolecular interaction networks**. *Genome Research*, **13**:2498–2504, 2003. 103

[274] Brian J ORoak, Laura Vives, Santhosh Girirajan, Emre Karakoc, Niklas Krumm, Bradley P Coe, Roie Levy, Arthur Ko, Choli Lee, Joshua D Smith, and et al. **Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations**. *Nature*, **485**(7397):246–250, 2012. 115, 193

[275] Benjamin M Neale, Yan Kou, Li Liu, Avi Maayan, Kaitlin E Samocha, Aniko Sabo, Chiao-Feng Lin, Christine Stevens, Li-San Wang, Vladimir Makarov, and et al. **Patterns and rates of exonic de novo mutations in autism spectrum disorders**. *Nature*, **485**(7397):5–9, 2012. 115, 193

[276] Simon J Mitchell and Timothy A Ryan. **Munc18-dependent regulation of synaptic vesicle exocytosis by syntaxin-1A in hippocampal neurons.** *Neuropharmacology*, **48**(3):372–380, 2005. 116

[277] M A Cervinski, J D Foster, and R A Vaughan. **Syntaxin 1A regulates dopamine transporter activity, phosphorylation and surface expression.** *Neuroscience*, **170**(2):408–416, 2010. 116

[278] Bor Luen Tang, Heon Yung Gee, and Min Goo Lee. **The cystic fibrosis transmembrane conductance regulator?s expanding SNARE interactome.** *Traffic Copenhagen Denmark*, **12**(4):364–371, 2011. 116

[279] Kazuhiko Nakamura, Ayyappan Anitha, Kazuo Yamada, Masatsugu Tsujii, Yoshimi Iwayama, Eiji Hattori, Tomoko Toyota, Shiro Suda, Noriyoshi Takei, Yasuhide Iwata, and et al. **Genetic and expression analyses reveal elevated expression of syntaxin 1A (STX1A) in high functioning autism.** *The international journal of neuropsychopharmacology official scientific journal of the Collegium Internationale Neuropsychopharmacologicum CINP*, **11**(8):1073–1084, 2008. 116

[280] Christopher W Tsang, Michael Fedchyshyn, John Harrison, Hong Xie, Jing Xue, Phillip J Robinson, Lu-Yang Wang, and William S Trimble. **Superfluous role of mammalian septins 3 and 5 in neuronal development and synaptic transmission.** *Molecular and cellular biology*, **28**(23):7012–7029, December 2008. 117

[281] Crestina L Beites, Kristen A Campbell, and William S Trimble. **The septin Sept5/CDCrel-1 competes with alpha-SNAP for binding to the SNARE complex.** *The Biochemical journal*, **385**(Pt 2):347–353, 2005. 117

[282] R J Barnard, A Morgan, and R D Burgoyne. **Domains of alpha-SNAP required for the stimulation of exocytosis and for N-ethylmalemide-sensitive fusion protein (NSF) binding and activation.** *Molecular Biology of the Cell*, **7**(5):693–701, 1996. 117

[283] NCBI. **Gene ID: 9342**, August 2012. 117

[284] G. Robert Parton and Kai Simons. **The multiple faces of caveolae**. *Nature Reviews Molecular Cell Biology*, **8**(3):185–194, 2007. 117

[285] Johanna M Magga, Jason G Kay, Alice Davy, Nadine P Poulton, Stephen M Robbins, and Janice E A Braun. **ATP dependence of the SNARE/caveolin 1 interaction in the hippocampus.** *Biochemical and Biophysical Research Communications*, **291**(5):1232–1238, 2002. 118

[286] J E Braun and D V Madison. **A novel SNAP25-caveolin complex correlates with the onset of persistent synaptic potentiation.** *J Neurosci*, **20**(16):5997–6006, 2000. 118

[287] Q Su, S Mochida, J H Tian, R Mehta, and Z H Sheng. **SNAP-29: a general SNARE protein that inhibits SNARE disassembly and is implicated in synaptic transmission.** *Proc Natl Acad Sci U S A*, **98**(24):14038–43, 2001. 118

[288] L Edelmann, P I Hanson, E R Chapman, and R Jahn. **Synaptobrevin binding to synaptophysin: a potential mechanism for controlling the exocytotic fusion machine.** *the The European Molecular Biology Organization Journal*, **14**(2):224–231, 1995. 118

[289] H T McMahon, M Missler, C Li, and T C Sdhof. **Complexins: cytosolic proteins that regulate SNAP receptor function.** *Cell*, **83**(1):111–119, 1995. 118

[290] Volker Haucke, Erwin Neher, and Stephan J Sigrist. **Protein scaffolds in the coupling of synaptic exocytosis and endocytosis.** *Nature Reviews Neuroscience*, **12**(3):127–138, 2011. 118

[291] Kimberly W Raines, Guan-Liang Cao, Eun Kyoung Lee, Gerald M Rosen, and Paul Shapiro. **Neuronal nitric oxide synthase-induced S-nitrosylation of H-Ras inhibits calcium ionophore-mediated extracellular-signal-regulated kinase activity**. *The Biochemical journal*, **397**(2):329–336, 2006. 120, 124

[292] Irina Nikonenko, Bernadett Boda, Sylvain Steen, Graham Knott, Egbert Welker, and Dominique Muller. **PSD-95 promotes synaptogenesis and multiinnervated spine formation through nitric oxide signaling.** *The Journal of Cell Biology*, **183**(6):1115–27, 2008. 120

[293] Kah Fei Wan, Balwinder S Sambi, Rothwelle Tate, Catherine Waters, and Nigel J Pyne. **The inhibitory gamma subunit of the type 6 retinal cGMP phosphodiesterase functions to link c-Src and G-protein-coupled receptor kinase 2 in a signaling unit that regulates p42/p44 mitogen-activated protein kinase by epidermal growth factor.** *The Journal of Biological Chemistry*, **277**(20):271–279, 2002. 120

[294] Ravi S Muddashetty, Vijayalaxmi C Nalavadi, Christina Gross, Xiaodi Yao, Lei Xing, Oskar Laur, Stephen T Warren, and Gary J Bassell. **Reversible inhibition of PSD-95 mRNA translation by miR-125a, FMRP phosphorylation, and mmGluR signaling.** *Molecular Cell*, **42**(5):673–688, 2011. 120

[295] H J Chen, M Rojas-Soto, A Oguni, and M B Kennedy. **A synaptic Ras-GTPase activating protein (p135 SynGAP) inhibited by CaM kinase II.** *Neuron*, **20**(5):895–904, 1998. 120

[296] M J Hart, Y Maru, D Leonard, O N Witte, T Evans, and R A Cerione. **A GDP dissociation inhibitor that serves as a GTPase inhibitor for the Ras-like protein CDC42Hs.** *Science*, **258**(5083):812–815, 1992. 120

[297] Joyce Van De Leemput, Jayanth Chandran, Melanie A Knight, Lynne A Holtzclaw, Sonja Scholz, Mark R Cookson, Henry Houlden, Katrina Gwinn-Hardy, Hon-Chung Fung, Xian Lin, and et al. **Deletion at ITPR1 Underlies Ataxia in Mice and Spinocerebellar Ataxia 15 in Humans**. *PLoS Genetics*, **3**(6):7, 2007. 120

[298] David Communi, Valerie Vanweyenberg, and Christophe Erneux. **D-myo-inositol 1,4,5-trisphosphate 3-kinase A is activated by receptor activation through a calcium:calmodulin-dependent protein kinase II phosphorylation mechanism**. *The EMBO Journal*, **16**(8):1943–1952, 1997. 120

[299] P J Woodring and J C Garrison. **Expression, purification, and regulation of two isoforms of the inositol 1,4,5-trisphosphate 3-kinase.** *The Journal of Biological Chemistry*, **272**(48):30447–30454, 1997. 120

[300] D Communi, V Dewaste, and C Erneux. **Calcium-calmodulin-dependent protein kinase II and protein kinase C-mediated phosphorylation and activation of D-myo-inositol 1,4, 5-trisphosphate 3-kinase B in astrocytes**. *Journal of Biological Chemistry*, **274**(21):14734–14742, 1999. 120

[301] M J Schell, C Erneux, and R F Irvine. **Inositol 1,4,5-trisphosphate 3-kinase A associates with F-actin and dendritic spines via its N terminus.** *The Journal of Biological Chemistry*, **276**(40):37537–37546, 2001. 121

[302] N Emptage, T V Bliss, and A Fine. **Single synaptic events evoke NMDA receptor-mediated release of calcium from internal stores in hippocampal dendritic spines.** *Neuron*, **22**(1):115–124, 1999. 121

[303] R Yuste, A Majewska, and K Holthoff. **From form to function: calcium compartmentalization in dendritic spines.** *Nature Neuroscience*, **3**(7):653–659, 2000. 121

[304] P Ruth, G X Wang, I Boekhoff, B May, A Pfeifer, R Penner, M Korth, H Breer, and F Hofmann. **Transfected cGMP-dependent protein kinase suppresses calcium transients by inhibition of inositol 1,4,5-trisphosphate production.** *Proceedings of the National Academy of Sciences of the United States of America*, **90**(7):2623–2627, 1993. 121

[305] K M Cadigan and R Nusse. **Wnt signaling: a common theme in animal development**. *Genes & Development*, **11**(24):3286–3305, 1997. 122

[306] Sarah R Gilman, Ivan Iossifov, Dan Levy, Michael Ronemus, Michael Wigler, and Dennis Vitkup. **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, 2011. 122

[307] Patricia C Salinas and Yimin Zou. **Wnt signaling in neural circuit assembly.** *Annual Review of Neuroscience*, **31**(1):339–358, 2008. 122

[308] Silvana B Rosso, Daniel Sussman, Anthony Wynshaw-Boris, and Patricia C Salinas. **Wnt signaling through Dishevelled, Rac and JNK regulates dendritic development.** *Nature Neuroscience*, **8**(1):34–42, 2005. 122

[309] P C Salinas, C Fletcher, N G Copeland, N A Jenkins, and R Nusse. **Maintenance of Wnt-3 expression in Purkinje cells of the mouse cerebellum depends on interactions with granule cells.** *Development Cambridge England*, **120**(5):1277–86, 1994. 122

[310] W Zeng, K A Wharton, J A Mack, K Wang, M Gadbaw, K Suyama, P S Klein, and M P Scott. **naked cuticle encodes an inducible antagonist of Wnt signalling.** *Nature*, **403**(6771):789–795, 2000. 122

[311] L Zeng, F Fagotto, T Zhang, W Hsu, T J Vasicek, W L Perry, J J Lee, S M Tilghman, B M Gumbiner, and F Costantini. **The mouse Fused locus encodes Axin, an inhibitor of the Wnt signaling pathway that regulates embryonic axis formation.** *Cell*, **90**(1):181–192, 1997. 122

[312] Whitney M Lum, Joshua K Robertson, and Terence J Van Raay. **Dishevelled2 is a stable protein during early zebrafish development.** *Zebrafish*, **8**(2):65–71, 2011. 122

[313] Yi Zhang, Soek Ying Neo, Jiahuai Han, and Sheng-Cai Lin. **Dimerization Choices Control the Ability of Axin and Dishevelled to Activate c-Jun N-terminal Kinase/Stress-activated Protein Kinase**. *Journal of Biological Chemistry*, **11**(275(32)):25008–14, 2000. 122

[314] Igor Schneider, Patricia N Schneider, Sarah W Derry, Shengda Lin, Lacy J Barton, Trudi Westfall, and Diane C Slusarski. **Zebrafish Nkd1 promotes Dvl degradation and is required for left-right patterning.** *Developmental Biology*, **348**(1):22–33, 2010. 122

[315] Sewoon Kim and Eek-hoon Jho. **The protein stability of Axin, a negative regulator of Wnt signaling, is regulated by Smad ubiquitination regulatory factor 2 (Smurf2).** *The Journal of Biological Chemistry*, **285**(47):36420–36426, 2010. 124

[316] Baldwin C Mak, Heidi L Kenerson, Lauri D Aicher, Elizabeth A Barnes, and Raymond S Yeung. **Aberrant beta-catenin signaling in tuberous sclerosis.** *The American journal of pathology*, **167**(1):107–116, 2005. 124

[317] Yinghua Zhang, Yan Sun, Fei Wang, Zhongping Wang, Yuwen Peng, and Ruixi Li. **Downregulating the Canonical Wnt/-catenin Signaling Pathway Attenuates the Susceptibility to Autism-like Phenotypes by Decreasing Oxidative Stress.** *Neurochemical Research*, 2012. 124

[318] Nathan D Okerlund and Benjamin N R Cheyette. **Synaptic Wnt signaling-a contributor to major psychiatric disorders?** *Journal of neurodevelopmental disorders*, **3**(2):162–174, 2011. 124

[319] Marjelo A Mines, Christopher J Yuskaitis, Margaret K King, Eleonore Beurel, and Richard S Jope. **GSK3 Influences Social Preference and Anxiety-Related Behaviors during Social Interaction in a Mouse Model of Fragile X Syndrome and Autism**. *PLoS ONE*, **5**(3):12, 2010. 124

[320] Walter Alexander. **American Psychiatric Association**. *Pharmacy and Therapeutics*, **33**(6):364–367, 2008. 197

[321] Katharine A Phillips. **The obsessive-compulsive spectrums.** *The Psychiatric clinics of North America*, **25**(4):791–809, 2002. 197

[322] D J Stein. **Neurobiology of the obsessive-compulsive spectrum disorders.** *Biological Psychiatry*, **47**(4):296–304, 2000. 197

[323] B DellOsso, A C Altamura, E Mundo, D Marazziti, and E Hollander. **Diagnosis and treatment of obsessive-compulsive disorder and related disorders.** *International Journal of Clinical Practice*, **61**(1):98–104, 2000. 197

[324] Oms. **Global burden disease**. *httpwwwwhointhealthinfobodenindexhtml*, 2008. 197

[325] Alicia Kaplan and Eric Hollander. **A review of pharmacologic treatments for obsessive-compulsive disorder.** *Psychiatric services Washington DC*, **54**(8):1111–1118, 2003. 198, 200

[326] Michael A Jenike. **Clinical practice. Obsessive-compulsive disorder.** *The New England Journal of Medicine*, **350**(3):259–265, 2004. 198, 200

[327] R Hoehn-Saric and B D Greenberg. **Psychobiology of obsessive-compulsive disorder: anatomical and physiological considerations**. *International Review of Psychiatry*, **9**(1):15–30, 1997. 198

[328] L R Baxter, M E Phelps, J C Mazziotta, B H Guze, J M Schwartz, and C E Selin. **Local cerebral glucose metabolic rates in obsessive-compulsive disorder. A comparison with rates in unipolar depression and in normal controls.** *Archives of General Psychiatry*, **44**(3):211–218, 1987. 198

[329] R Baxter, M Schwartz, E Phelps, Barry H Guze, Lynn Fairbanks, and D Ph. **Cerebral glucose metabolic rates in nondepressed patients with obsessive-compulsive disorder**. *Psychiatry Interpersonal and Biological Processes*, **145**(12):1560–1563, 1988. 198

[330] S E Swedo, M B Schapiro, C L Grady, D L Cheslow, H L Leonard, A Kumar, R Friedland, S I Rapoport, and J L Rapoport. **Cerebral glucose metabolism in childhood-onset obsessive-compulsive disorder**. *Archives of General Psychiatry*, **46**(6):518–523, 1989. 198

[331] G V Sawle, N F Hymas, A J Lees, and R S Frackowiak. **OBSESSIONAL SLOWNESS - FUNCTIONAL-STUDIES WITH POSITRON EMISSION TOMOGRAPHY**. *Brain: A journal of neurology*, **114 (Pt 5)**(Pt 5):2191–2202, 1991. 198

[332] S L Rauch, M A Jenike, N M Alpert, L Baer, H C Breiter, C R Savage, and A J Fischman. **Regional cerebral blood flow measured during symptom provocation in obsessive-compulsive disorder using oxygen 15-labeled carbon dioxide and positron emission tomography.** *Archives of General Psychiatry*, **51**(1):62–70, 1994. 198

[333] P K Mcguire, C J Bench, C D Frith, I M Marks, R S Frackowiak, and R J Dolan. **Functional anatomy of obsessive-compulsive phenomena.** *The British Journal of Psychiatry*, **164**(4):459–468, 1994. 198

[334] L R Baxter, J M Schwartz, J C Mazziotta, M E Phelps, J J Pahl, B H Guze, and L Fairbanks. **Local cerebral glucose metabolic rates in obsessive-compulsive disorder. Patients treated with clomipramine.** *Archives of General Psychiatry*, **145**(9):1560–1563, 1990. 198

[335] L R Baxter, J M Schwartz, K S Bergman, M P Szuba, B H Guze, J C Mazziotta, A Alazraki, C E Selin, H K Ferng, and P Munford. **Caudate glucose metabolic rate changes with both drug and behavior therapy for obsessive-compulsive disorder.** *Archives of General Psychiatry*, **49**(9):681–689, 1992. 198

[336] D Perani, C Colombo, S Bressi, A Bonfanti, F Grassi, S Scarone, L Bellodi, E Smeraldi, and F Fazio. **[18F]FDG PET study in obsessive-compulsive disorder. A clinical/metabolic correlation study after treatment.** *The British Journal of Psychiatry*, **166**(2):244–250, 1995. 198

[337] S Saxena, A L Brody, K M Maidment, J J Dunkin, M Colgan, S Alborzian, M E Phelps, and L R Baxter. **Localized orbitofrontal and subcortical metabolic changes and predictors of response to paroxetine treatment in obsessive-compulsive disorder.** *Neuropsychopharmacology*, **21**(6):683–693, 1999. 198

[338] J M Schwartz, P W Stoessel, L R Baxter, K M Martin, and M E Phelps. **Systematic changes in cerebral glucose metabolic rate after successful behavior modification treatment of obsessive-compulsive disorder.** *Archives of General Psychiatry*, **53**(2):109–113, 1996. 198

[339] P Mindus and H Nyman. **Normalization of personality characteristics in patients with incapacitating anxiety disorders after capsulotomy.** *Acta Psychiatrica Scandinavica*, **83**(4):283–291, 1991. 198

[340] Christopher Pittenger, John H Krystal, and Vladimir Coric. **Glutamate-modulating drugs as novel pharmacotherapeutic agents in the treatment of obsessive-compulsive disorder.** *NeuroRx the journal of the American Society for Experimental NeuroTherapeutics*, **3**(1):69–81, 2006. 198

[341] D Ebert, O Speck, A Knig, M Berger, J Hennig, and F Hohagen. **1H-magnetic resonance spectroscopy in obsessive-compulsive disorder: evidence for neuronal loss in the cingulate gyrus and the right striatum.** *Psychiatry Research*, **74**(3):173–176, 1997. 198

[342] R Bartha, M B Stein, P C Williamson, D J Drost, R W Neufeld, T J Carr, G Canaran, M Densmore, G Anderson, and A R Siddiqui. **A short echo 1H spectroscopy and volumetric MRI study of the corpus striatum in patients**

with obsessive-compulsive disorder and comparison subjects. *The American Journal of Psychiatry*, **155**(11):1584–1591, 1998. 198

[343] D R Rosenberg and M S Keshavan. **A.E. Bennett Research Award. Toward a neurodevelopmental model of of obsessive–compulsive disorder.** *Biological Psychiatry*, **43**(9):623–640, 1998. 198

[344] Jonathan T Ting and Guoping Feng. **Glutamatergic Synaptic Dysfunction and Obsessive-Compulsive Disorder**. *Current chemical genomics*, **2**:62–75, 2008. 199

[345] Speculations O N The, Functional Anatomy, O F Basal, and Ganglia Disorders. **SPECULATIONS ON THE FUNCTIONAL ANATOMY OF BASAL**. *Young*, pages 73–94, 1983. 198

[346] George Paxinos. *The rat nervous system*. Elsevier Academic Press, 2004. 198

[347] A M Graybiel. **The basal ganglia.** *Current Biology*, **10**(14):R509–R511, 2000. 198

[348] Sibel Tekin and Jeffrey L Cummings. **Frontal-subcortical neuronal circuits and clinical neuropsychiatry: an update.** *Journal of Psychosomatic Research*, **53**(2):647–654, 2002. 200

[349] D Hoyeri and A Pazosi. **Serotonin Receptors in the Human Brain. I. Characterization and Autoradiographic Localization of 5-HT**. *Science*, **376**:85–96, 1986. 200

[350] Gerald Nestadt, Marco Grados, and Jack F Samuels. **Genetics of obsessive-compulsive disorder.** *The Psychiatric clinics of North America*, **33**(1):141–158, 2010. 200

[351] Danil S Van Grootheest, Danille C Cath, Aartjan T Beekman, and Dorret I Boomsma. **Twin studies on obsessive-compulsive disorder: a review.** *Twin research and human genetics the official journal of the International Society for Twin Studies*, **8**(5):450–458, 2005. 200

[352] Marco A Grados, John Walkup, and Samuel Walford. **Genetics of obsessive-compulsive disorders: new findings and challenges.** *Brain & Development*, **25 Suppl 1**:S55–S61, 2003. 200

[353] David L Pauls. **The genetics of obsessive compulsive disorder: a review of the evidence.** *American Journal Of Medical Genetics Part C Seminars In Medical Genetics*, **148C**(2):133–139, 2008. 200

[354] San M J Hemmings and Dan J Stein. **The current status of association studies in obsessive-compulsive disorder.** *The Psychiatric clinics of North America*, **29**(2):411–444, 2006. 200

[355] S Zchner, J R Wendland, A E Ashley-Koch, A L Collins, K N Tran-Viet, K Quinn, K C Timpano, M L Cuccaro, M A Pericak-Vance, D C Steffens, and et al. **Multiple rare SAPAP3 missense variants in trichotillomania and OCD.** *Molecular Psychiatry*, **14**(1):6–9, 2009. 200

[356] O J Bienvenu, Y Wang, Y Y Shugart, J M Welch, M A Grados, A J Fyer, S L Rauch, J T McCracken, S A Rasmussen, D L Murphy, and et al. **Sapap3 and pathological grooming in humans: Results from the OCD collaborative genetics study.** *American journal of medical genetics Part B Neuropsychiatric genetics the official publication of the International Society of Psychiatric Genetics*, **150B**(5):710–720, 2009. 200

[357] N H Dodman, E K Karlsson, A Moon-Fanelli, M Galdzicka, M Perloski, L Shuster, K Lindblad-Toh, and E I Ginns. **A canine chromosome 7 locus confers compulsive disorder susceptibility.** *Molecular Psychiatry*, **15**(1):8–10, 2010. 201, 203, 237, 324, 325, 331, 346, 347, 365, 367, 373, 374

[358] High-throughput Sequencing and Preparation Oligonucleotide Kit. **Multiplexed Sequencing with the Illumina Genome Analyzer System**. *Analyzer*, **2010**(6):1–4, 2008. 202

[359] Keiko Satoh, Masakazu Takeuchi, Yoshiya Oda, Maki Deguchi-Tawarada, Yoshimasa Sakamoto, Kaho Matsubara, Takeshi Nagasu, and Yoshimi Takai. **Identification of activity-regulated proteins in the postsynaptic density fraction.** *Genes to cells devoted to molecular cellular mechanisms*, **7**(2):187–197, 2002. 203

[360] Bryen A Jordan, Brian D Fernholz, Muriel Boussac, Chongfeng Xu, Gabriela Grigorean, Edward B Ziff, and Thomas A Neubert. **Identification and verification of novel rodent postsynaptic density proteins.** *Molecular cellular proteomics MCP*, **3**(9):857–871, 2004. 203

[361] Ed S Lein, Michael J Hawrylycz, Nancy Ao, Mikael Ayres, Amy Bensinger, Amy Bernard, Andrew F Boe, Mark S Boguski, Kevin S Brockway, Emi J Byrnes, and et al. **Genome-wide atlas of gene expression in the adult mouse brain.** *Nature*, **445**(7124):168–176, 2007. 203

[362] Anatol C Kreitzer and Robert C Malenka. **Striatal plasticity and basal ganglia circuit function.** *Neuron*, **60**(4):543–554, 2008. 203

[363] Jack Samuels, Yin Yao Shugart, Marco A Grados, Virginia L Willour, O Joseph Bienvenu, Benjamin D Greenberg, James A Knowles, James T McCracken, Scott L Rauch, Dennis L Murphy, and et al. **Significant linkage to compulsive hoarding on chromosome 14 in families with obsessive-compulsive disorder: results from the OCD Collaborative Genetics Study.** *The American Journal of Psychiatry*, **164**(3):493–499, 2007. 203

[364] Jennifer M Chou-Green, Todd D Holscher, Mary F Dallman, and Susan F Akana. **Compulsive behavior in the 5-HT2C receptor knockout mouse.** *Physiology & Behavior*, **78**(4-5):641–649, 2003. 203

[365] P Jia, J Sun, A Y Guo, and Z Zhao. **SZGR: a comprehensive schizophrenia gene resource.** *Molecular Psychiatry*, **15**(5):453–462, 2010. 203

[366] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and et al. **The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Research*, **20**(9):1297–1303, 2010. 203, 326, 333, 387

[367] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo Del Angel, Manuel A Rivas, Matt Hanna, and et al. **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nature Genetics*, **43**(5):491–8, 2011. 203, 242, 326, 333

[368] Manuel A Rivas, Mlissa Beaudoin, Agnes Gardet, Christine Stevens, Yashoda Sharma, Clarence K Zhang, Gabrielle Boucher, Stephan Ripke, David Ellinghaus, Noel Burtt, and et al. **Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease**. *Nature Genetics*, **43**(11):1066–1073, 2011. 203, 212

[369] Manuel A. Rivas and Mark J. Daly. **Syzygy Documentation Release 1.1.0**, 2011. 203, 204

[370] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nature Protocols*, **4**(1):44–57, 2009. 204

[371] Bo Eskerod Madsen and Sharon R Browning. **A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic**. *PLoS Genetics*, **5**(2):11, 2009. 220

[372] Samuel Levy, Granger Sutton, Pauline C Ng, Lars Feuk, Aaron L Halpern, Brian P Walenz, Nelson Axelrod, Jiaqi Huang, Ewen F Kirkness, Gennady Denisov, and et al. **The diploid genome sequence of an individual human**. *PLoS Biology*, **5**(10):e254, 2007. 220

[373] R. Zhang and B. Su. **Small but influential: the role of microRNAs on gene regulatory network and 3 UTR evolution**. *Journal of Genetics and Genomics*, **36**(1):1–6, 2009. 230

[374] R.C. Friedman, K.K.H. Farh, C.B. Burge, and D.P. Bartel. **Most mammalian mRNAs are conserved targets of microRNAs**. *Genome research*, **19**(1):92–105, 2009. 230

[375] Iris Cheung, Hennady P Shulha, Yan Jiang, Anouch Matevossian, Jie Wang, Zhiping Weng, and Schahram Akbarian. **Developmental regulation and individual differences of neuronal H3K4me3 epigenomes in the prefrontal cortex**. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(19):8824–8829, 2010. 230, 231

[376] E Schreiber, A Tobler, U Malipiero, W Schaffner, and A Fontana. **cDNA cloning of human N-Oct3, a nervous-system specific POU domain transcription factor binding to the octamer DNA motif.** *Nucleic Acids Research*, **21**(2):253–258, 1993. 232

[377] Dashzeveg Bayarsaihan, Judit Dunai, John M Greally, Kazuhiko Kawasaki, Kenta Sumiyama, Badam Enkhmandakh, Nobuyoshi Shimizu, and Frank H Ruddle. **Isolation of the human genomic brain-2/N-Oct 3 gene (POUF3) and assignment to chromosome 6q16.** *Genomics*, **79**(2):137–143, 2002. 232

[378] A Ashworth, B P Williams, A M Buchberg, P N Goodfellow, E Solomon, J Potter, and K R Willison. **The gene for the ubiquitous octamer-binding protein Oct-1 is on human chromosome 1, region cen-q32, and near Ly-22 and Ltw-4 on mouse chromosome 1.** *Genomics*, **6**(3):666–672, 1990. 232

[379] A C Strm, M Forsberg, P Lillhager, and G Westin. **The transcription factors Sp1 and Oct-1 interact physically to regulate human U2 snRNA gene expression.** *Nucleic Acids Research*, **24**(11):1981–1986, 1996. 232

[380] Lei Zheng, Robert G Roeder, and Yan Luo. **S phase activation of the histone H2B promoter by OCA-S, a coactivator complex that contains GAPDH as a key component.** *Cell*, **114**(2):255–266, 2003. 232

[381] P Carninci, T Kasukawa, S Katayama, J Gough, M C Frith, N Maeda, R Oyama, T Ravasi, B Lenhard, C Wells, and et al. **The transcriptional landscape of the mammalian genome**. *Science*, **309**(5740):1559–1563, 2005. 234

[382] Sam Griffiths-Jones. **Annotating noncoding RNA genes.** *Annual Review of Genomics and Human Genetics*, **8**(1):279–298, 2007. 234, 235

[383] Dmitri Parkhomchuk, Tatiana Borodina, Vyacheslav Amstislavskiy, Maria Banaru, Linda Hallen, Sylvia Krobitsch, Hans Lehrach, and Alexey Soldatov. **Transcriptome analysis by strand-specific sequencing of complementary DNA**. *Nucleic Acids Research*, **37**(18):e123, 2009. 234

[384] J L Biedler, L Helson, and B A Spengler. **Morphology and growth, tumorigenicity, and cytogenetics of human neuroblastoma cells in continuous culture.** *Cancer Research*, **33**(11):2643–2652, 1973. 234

[385] Ken C Pang, Stuart Stephen, Pr G Engstrm, Khairina Tajul-Arifin, Weisan Chen, Claes Wahlestedt, Boris Lenhard, Yoshihide Hayashizaki, and John S Mattick. **RNAdba comprehensive mammalian noncoding RNA database**. *Nucleic Acids Research*, **33**(Database Issue):D125–D130, 2005. 235

[386] Changning Liu, Baoyan Bai, Geir Skogerb, Lun Cai, Wei Deng, Yong Zhang, Dongbo Bu, Yi Zhao, and Runsheng Chen. **NONCODE: an integrated knowledge database of non-coding RNAs**. *Nucleic Acids Research*, **33**(Database Issue):D112–D115, 2005. 235

[387] Jakob Skou Pedersen, Gill Bejerano, Adam Siepel, Kate Rosenbloom, Kerstin Lindblad-Toh, Eric S Lander, Jim Kent, Webb Miller, and David Haussler. **Identification and Classification of Conserved RNA Secondary Structures in the Human Genome**. *PLoS Computational Biology*, **2**(4):12, 2006. 235

[388] Stefan Washietl, Ivo L Hofacker, Melanie Lukasser, Alexander Httenhofer, and Peter F Stadler. **Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome.** *Nature Biotechnology*, **23**(11):1383–1390, 2005. 235

[389] E Rivas and S R Eddy. **Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs.** *Bioinformatics*, **16**(7):583–605, 2000. 235

[390] David H Mathews. **Revolutions in RNA Secondary Structure Prediction**. *Journal of Molecular Biology*, **359**(3):526–532, 2006. 235

[391] Vienna Format, Ct Format, Image Converter, Internet Explorer, and Adobe Svg. **RNAfold web server**, 2011. 235

[392] NCBI. **Gene ID: 6091**, August 2012. 237

[393] NCBI. **Gene ID: 1012**, August 2012. 237

[394] YY Shugart, J. Samuels, VL Willour, MA Grados, BD Greenberg, JA Knowles, JT McCracken, SL Rauch, DL Murphy, Y. Wang, et al. **Genomewide linkage scan for obsessive-compulsive disorder: evidence for susceptibility loci on chromosomes 3q, 7p, 1q, 15q, and 6q**. *Molecular Psychiatry*, **11**(8):763–770, 2006. 237

[395] Matthew N Bainbridge, Min Wang, Daniel L Burgess, Christie Kovar, Matthew J Rodesch, Mark D'Ascenzo, Jacob Kitzman, Yuan-Qing Wu, Irene Newsham, Todd A Richmond, Jeffrey A Jeddeloh, Donna Muzny, Thomas J Albert, and Richard A Gibbs. **Whole exome capture in solution with 3 Gbp of data**. *Genome Biology*, **11**(6):R62, 2010. 240

[396] P Menozzi, A Piazza, and L Cavalli-Sforza. **Synthetic maps of human gene frequencies in Europeans.** *Science*, **201**(4358):786–792, 1978. 244

[397] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. **Detection of nonneutral substitution rates on mammalian phylogenies**. *Genome Research*, **20**(1):110–121, 2010. 312, 361

[398] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, Ladeana W Hillier, Stephen Richards, and et al. **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome Research*, **15**(8):1034–50, 2005. 312, 361

[399] Nathan B Sutter and Elaine A Ostrander. **Dog star rising: the canine genetic system.** *Nature Reviews Genetics*, **5**(12):900–910, 2004. 317

[400] R K Wayne, E Geffen, D J Girman, K P Koepfli, L M Lau, and C R Marshall. **Molecular systematics of the Canidae.** *Systematic Biology*, **46**(4):622–653, 1997. 317

[401] C Vila. **Multiple and Ancient Origins of the Domestic Dog**. *Science*, **276**(5319):1687–1689, 1997. 317

[402] Carolyne Bardeleben, Rachael L Moore, and Robert K Wayne. **Isolation and molecular evolution of the selenocysteine tRNA (Cf TRSP) and RNase P RNA (Cf RPPH1) genes in the dog family, Canidae.** *Molecular Biology and Evolution*, **22**(2):347–359, 2005. 317

[403] Peter Savolainen, Ya-ping Zhang, Jing Luo, Joakim Lundeberg, and Thomas Leitner. **Genetic evidence for an East Asian origin of domestic dogs.** *Science*, **298**(5598):1610–3, 2002. 317

[404] Kerstin Lindblad-Toh, Claire M Wade, Tarjei S Mikkelsen, Elinor K Karlsson, David B Jaffe, Michael Kamal, Michele Clamp, Jean L Chang, Edward J Kulbokas, Michael C Zody, and et al. **Genome sequence, comparative analysis and haplotype structure of the domestic dog.** *Nature*, **438**(7069):803–819, 2005. 317, 319, 320, 330, 333, 346

[405] American Kennel Club. *The Complete Dog Book.* Howell Book House, New York, 19th edition, 1998. 317

[406] R K Wayne. **Limb morphology of domestic and wild canids: the influence of development on morphologic change.** *Journal of Morphology*, **187**(3):301–319, 1986. 317

[407] Elinor K Karlsson and Kerstin Lindblad-Toh. **Leader of the pack: gene mapping in dogs and other model organisms.** *Nature Reviews Genetics*, **9**(9):713–725, 2008. 317, 318, 319, 321

[408] E A Ostrander, F Galibert, and D F Patterson. **Canine genetics comes of age.** *Trends in Genetics*, **16**(3):117–124, 2000. 317

[409] David R Sargan. **IDID: inherited diseases in dogs: web-based information for canine inherited disease genetics.** *Mammalian genome official journal of the International Mammalian Genome Society*, **15**(6):503–506, 2004. 317

[410] Jennie L Rowell, Donna O McCarthy, and Carlos E Alvarez. **Dog models of naturally occurring cancer.** *Trends in Molecular Medicine*, **17**(7):380–388, 2011. 317

[411] D M Vail and E G MacEwen. **Spontaneously occurring tumors of companion animals as models for human cancer.** *Cancer Investigation*, **18**(8):781–92, 2000. 317

[412] Laurel J Gershwin. **Veterinary autoimmunity: autoimmune diseases in domestic animals.** *Annals Of The New York Academy Of Sciences*, **1109**(3):109–116, 2007. 317

[413] W Löscher, D Schwartz-Porsche, H H Frey, and D Schmidt. **Evaluation of epileptic dogs as an animal model of human epilepsy.** *ArzneimittelForschung*, **35**(1):82–87, 1985. 317

[414] K.L. Overall. **Natural animal models of human psychiatric conditions: assessment of mechanism and validity**. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, **24**(5):727–776, 2000. 317, 322, 323

[415] F Mueller, B Fuchs, and B Kaser-Hotz. **Comparative biology of human and canine osteosarcoma.** *Anticancer Research*, **27**(1A):155–164, 2007. 317

[416] The International and Hapmap Consortium. **ARTICLES A haplotype map of the human genome**. *Nature*, **437**(7063):1299–1320, 2005. 319

[417] K Hansen and C Khanna. **Spontaneous and genetically engineered animal models; use in preclinical cancer drug development.** *European Journal of Cancer*, **40**(6):858–880, 2004. 319

[418] Chand Khanna, Jennifer Prehn, David Hayden, Ryan D Cassaday, Jana Caylor, Shevin Jacob, Seuli M Bose, Sung-Hyeok Hong, Stephen M Hewitt, and Lee J Helman. **A randomized controlled trial of octreotide pamoate long-acting release and carboplatin versus carboplatin alone in dogs with naturally occurring osteosarcoma: evaluation of insulin-like growth factor suppression and chemotherapy.** *Clinical Cancer Research*, **8**(7):2406–2412, 2002. 319

[419] D L Hartl and A G Clark. *Principles of Population Genetics*, **37**. Sinauer Associates, 1997. 319

[420] L Peltonen, A Palotie, and K Lange. **Use of population isolates for mapping complex traits.** *Nature Reviews Genetics*, **1**(3):182–190, 2000. 319, 320

[421] Nathan B Sutter, Michael A Eberle, Heidi G Parker, Barbara J Pullar, Ewen F Kirkness, Leonid Kruglyak, and Elaine A Ostrander. **Extensive and breed-specific linkage disequilibrium in Canis familiaris.** *Genome Research*, **14**(12):2388–2396, 2004. 319, 320

[422] Molly Przeworski. **The signature of positive selection at randomly chosen loci.** *Genetics*, **160**(3):1179–1189, 2002. 319

[423] U A Luescher, D B McKeown, and J Halip. **Stereotypic or obsessive-compulsive disorders in dogs and cats.** *The Veterinary clinics of North America Small animal practice*, **21**(2):401–413, 1991. 322

[424] K L Overall. **Use of clomipramine to treat ritualistic stereotypic motor behavior in three dogs.** *Journal of the American Veterinary Medical Association*, **205**(12):1733–1741, 1994. 322, 323

[425] J L Rapoport, D H Ryland, and M Kriete. **Drug treatment of canine acral lick. An animal model of obsessive-compulsive disorder.** *Archives of General Psychiatry*, **49**(7):517–521, 1992. 323

[426] Karen L Overall and Arthur E Dunham. **Clinical features and outcome in dogs and cats with obsessive-compulsive disorder: 126 cases (1989-2000).** *Journal of the American Veterinary Medical Association*, **221**(10):1445–1452, 2002. 323

[427] C J Hewson, U A Luescher, J M Parent, P D Conlon, and R O Ball. **Efficacy of clomipramine in the treatment of canine compulsive disorder.** *Journal of the American Veterinary Medical Association*, **213**(12):1760–1766, 1998. 323

[428] A A Moon-Fanelli and N H Dodman. **Description and development of compulsive tail chasing in terriers and response to clomipramine treatment.** *Journal of the American Veterinary Medical Association*, **212**(8):1252–1257, 1998. 323

[429] K.L. Overall. *Clinical Behavioral Medicine for Small Animals*. Mosby, 1997. 323

[430] A M Graybiel and S L Rauch. **Toward a neurobiology of obsessive-compulsive disorder.** *Neuron*, **28**(2):343–7, 2000. 323

[431] Robert E Handsaker, Joshua M Korn, James Nemesh, and Steven A McCarroll. **Discovery and genotyping of genome structural polymorphism by sequencing on a population scale.** *Nature Genetics*, **43**(3):269–276, 2011. 326, 327, 350, 351, 377, 389, 393

[432] Kai Ye, Marcel H Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. **Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads**. *Bioinformatics*, **25**(21):2865–2871, 2009. 326

[433] Hugo Y K Lam, Xinmeng Jasmine Mu, Adrian M Sttz, Andrea Tanzer, Philip D Cayting, Michael Snyder, Philip M Kim, Jan O Korbel, and Mark B Gerstein. **Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library.** *Nature Biotechnology*, **28**(1):47–55, 2010. 326

[434] Jan O Korbel, Alexander Eckehart Urban, Jason P Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M Kim, Dean Palejev, Nicholas J Carriero, Lei Du, and et al. **Paired-end mapping reveals extensive structural variation in the human genome.** *Science*, **318**(5849):420–426, 2007. 326

[435] Jan O Korbel, Alexej Abyzov, Xinmeng Jasmine Mu, Nicholas Carriero, Philip Cayting, Zhengdong Zhang, Michael Snyder, and Mark B Gerstein. **PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data**. *Genome Biology*, **10**(2):R23, 2009. 326

[436] Ken Chen, John W Wallis, Michael D McLellan, David E Larson, Joelle M Kalicki, Craig S Pohl, Sean D McGrath, Michael C Wendl, Qunyuan Zhang, Devin P Locke, and et al. **BreakDancer: an algorithm for high-resolution mapping of genomic structural variation.** *Nature Methods*, **6**(9):677–681, 2009. 326

[437] Derek Y Chiang, Gad Getz, David B Jaffe, Michael J T OKelly, Xiaojun Zhao, Scott L Carter, Carsten Russ, Chad Nusbaum, Matthew Meyerson, and Eric S Lander. **High-resolution mapping of copy-number alterations with massively parallel sequencing.** *Nature Methods*, **6**(1):99–103, 2009. 326

[438] Seungtai Yoon, Zhenyu Xuan, Vladimir Makarov, Kenny Ye, and Jonathan Sebat. **Sensitive and accurate detection of copy number variants using read depth of coverage.** *Genome Research*, **19**(9):1586–1592, 2009. 326

[439] Can Alkan, Jeffrey M Kidd, Tomas Marques-Bonet, Gozde Aksay, Francesca Antonacci, Fereydoun Hormozdiari, Jacob O Kitzman, Carl Baker, Maika Malig, Onur Mutlu, and et al. **Personalized copy number and segmental duplication maps using next-generation sequencing.** *Nature Genetics*, **41**(10):1061–1067, 2009. 326

[440] Helga Thorvaldsdttir, James T Robinson, and Jill P Mesirov. **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.** *Briefings in Bioinformatics*, **Advance pu**:bbs017–, 2012. 328, 376

[441] Heng Li, Jue Ruan, and Richard Durbin. **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Research*, **18**(11):1851–1858, 2008. 329

[442] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics*, **25**(16):2078–2079, 2009. 329

[443] K Truv, O Eriksson, and M Norling. % bf SEQscoring: a tool to facilitate the interpretation of data generated with next generation sequencing technologies. *EMBnet*, 2011. 329, 344, 390

[444] Motoo Kimura. **A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences**. *Journal of Molecular Evolution*, **16**(1330):111–120, 1980. 336

[445] John Wakeley. **The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance**. *Trends in Ecology & Evolution*, **33**(4):158–162, 1996. 336

[446] Heidi G Parker, Lisa V Kim, Nathan B Sutter, Scott Carlson, Travis D Lorentzen, Tiffany B Malek, Gary S Johnson, Hawkins B DeFrance, Elaine A Ostrander, and Leonid Kruglyak. **Genetic structure of the purebred domestic dog.** *Science*, **304**(5674):1160–1164, 2004. 341, 342, 358

[447] I K Puja, D N Irion, A L Schaffer, and N C Pedersen. **The Kintamani dog: genetic profile of an emerging breed from Bali, Indonesia.** *The Journal of heredity*, **96**(7):854–859, 2005. 341

[448] Hidekazu Tanaka, Weisong Shan, Greg R. Phillips, Kirsten Arndt, Ozlem Bozdagi, Lawrence Shapiro, George W. Huntley, Deanna L. Benson, and David R. Colman. % bf Molecular Modification of N-Cadherin in Response to Synaptic Activity. *Neuron*, **25**(1):93 – 107, 2000. 346

[449] R Gingras, C Richard, M El-Alfy, C R Morales, M Potier, and A V Pshezhetsky. **Purification, cDNA cloning, and expression of a new human blood plasma glutamate carboxypeptidase homologous to N-acetyl-aspartyl-alpha-glutamate carboxypeptidase/prostate-specific membrane antigen.** *The Journal of Biological Chemistry*, **274**(17):11742–11750, 1999. 352, 364

[450] Jack Russell Terrier Club of America. **The JRTCA Registry**, July 2012. 357

[451] Jason Ernst and Manolis Kellis. **Discovery and characterization of chromatin states for systematic annotation of the human genome.** *Nature Biotechnology*, **28**(8):817–825, 2010. 361

[452] Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, Noam Shoresh, Lucas D Ward, Charles B Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, and et al. **Mapping and analysis of chromatin state dynamics in nine human cell types.** *Nature*, **473**(7345):43–49, 2011. 361

[453] Hela Mrabet Khiari, Afef Achouri, Nadia Ben Ali, Aroua Cherif, Hend Batti, Taieb Messaoud, and Amel Mrabet. **Obsessive-compulsive disorder: a new risk factor for Alzheimer disease?** *Neurological sciences official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology*, **32**(5):959–962, 2011. 362

[454] Yoav Gilad, Scott A Rifkin, and Jonathan K Pritchard. **Revealing the architecture of gene regulation: the promise of eQTL studies.** *Trends in Genetics*, **24**(8):408–415, 2008. 363

[455] Charalampos G Spilianakis, Maria D Lalioti, Terrence Town, Gap Ryol Lee, and Richard A Flavell. **Interchromosomal associations between alternatively expressed loci.** *Nature*, **435**(7042):637–645, 2005. 363

[456] K Kawamura, K Watanabe, T Suzuki, T Yamakawa, T Kamiyama, H Nakagawa, and S Tsurufuji. **cDNA cloning and expression of a novel human desmocollin.** *The Journal of Biological Chemistry*, **269**(42):26295–26302, 1994. 363

[457] Marc M Oshiro, Christina J Kim, Ryan J Wozniak, Damian J Junk, Jos L Muoz-Rodrguez, Jeanne A Burr, Matthew Fitzgerald, Sangita C Pawar, Anne E Cress, Frederick E Domann, and et al. **Epigenetic silencing of DSC3 is a common event in human breast cancer**. *Breast Cancer Research*, **7**(5):R669–R680, 2005. 363

[458] B S Meldrum. **Glutamate as a neurotransmitter in the brain: review of physiology and pathology.** *The Journal of nutrition*, **130**(4S Suppl):1007S–15S, 2000. 364

[459] Andrew I Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, and et al. **A gene atlas of the mouse and human protein-encoding transcriptomes**. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(16):6062–6067, 2004. 366

[460] NCBI. **Gene ID: 6310**, September 2012. 367

[461] Patricia C Salinas and Stephen R Price. **Cadherins and catenins in synapse development**. *Current Opinion in Neurobiology*, **15**(1):73–80, 2005. 368

[462] M Takeichi. **Cadherin cell adhesion receptors as a morphogenetic regulator.** *Science*, **251**(5000):1451–1455, 1991. 368

[463] M Takeichi, S Nakagawa, S Aono, T Usui, and T Uemura. **Patterning of cell assemblies regulated by adhesion receptors of the cadherin superfamily.** *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, **355**(1399):885–890, 2000. 368

[464] Y Shimoyama, G Tsujimoto, M Kitajima, and M Natori. **Identification of three human type-II classic cadherins and frequent heterophilic interactions between different subclasses of type-II classic cadherins.** *The Biochemical journal*, **349**(Pt 1):159–167, 2000. 368

[465] K Herrenknecht, M Ozawa, C Eckerskorn, F Lottspeich, M Lenter, and R Kemler. **The uvomorulin-anchorage protein alpha catenin is a vinculin homologue.** *Proceedings of the National Academy of Sciences of the United States of America*, **88**(20):9156–9160, 1991. 368

[466] D L Benson and H Tanaka. **N-cadherin redistribution during synaptogenesis in hippocampal neurons.** *Journal of Neuroscience*, **18**(17):6892–6904, 1998. 369

[467] Hideru Togashi, Kentaro Abe, Akira Mizoguchi, Kanna Takaoka, Osamu Chisaka, and Masatoshi Takeichi. **Cadherin regulates dendritic spine morphogenesis.** *Neuron*, **35**(1):77–89, 2002. 369

[468] Kentaro Abe, Osamu Chisaka, Frans Van Roy, and Masatoshi Takeichi. **Stability of dendritic spines and synaptic contacts is controlled by alpha N-catenin.** *Nature Neuroscience*, **7**(4):357–363, 2004. 369

[469] Yugo Fukazawa, Yoshito Saitoh, Fumiko Ozawa, Yasuhiko Ohta, Kensaku Mizuno, and Kaoru Inokuchi. **Hippocampal LTP is accompanied by enhanced F-actin content within the dendritic spine that is essential for late LTP maintenance in vivo.** *Neuron*, **38**(3):447–460, 2003. 369

[470] S. Lee, S. Hong, S. Kim, and S. Kang. **Ataxin-1 occupies the promoter region of E-cadherin in vivo and activates CtBP2-repressed promoter**. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, **1813**(5):713–722, 2011. 369

[471] K. Shimamura, S. Hirano, A.P. McMahon, and M. Takeichi. **Wnt-1-dependent regulation of local E-cadherin and alpha N-catenin expression in the embryonic mouse brain**. *Development*, **120**(8):2225–2234, 1994. 369

[472] A C Hall, F R Lucas, and P C Salinas. **Axonal remodeling and synaptic differentiation in the cerebellum is regulated by WNT-7a signaling.** *Cell*, **100**(5):525–535, 2000. 369

[473] Roberto Rosato, Jacqueline M Veltmaat, John Groffen, and Nora Heisterkamp. **Involvement of the tyrosine kinase fer in cell adhesion.** *Molecular and Cellular Biology*, **18**(10):5762–5770, 1998. 369

[474] Jose Piedra, Susana Miravet, Julio Castao, Hctor G Plmer, Nora Heisterkamp, Antonio Garca De Herreros, and Mireia Duach. **p120 Catenin-associated Fer and Fyn tyrosine kinases regulate beta-catenin Tyr-142 phosphorylation and beta-catenin-alpha-catenin Interaction.** *Molecular and Cellular Biology*, **23**(7):2287–2297, 2003. 369

[475] NCBI. **Gene ID: 9315**, August 2012. 370

[476] NCBI. **Gene ID: 83992**, August 2012. 370

[477] Qun Lu, Mercedes Paredes, Miguel Medina, Jianhua Zhou, Robert Cavallo, Mark Peifer, Lisa Orecchio, and Kenneth S Kosik. **delta-catenin, an adhesive junction-associated protein which promotes cell scattering.** *The Journal of Cell Biology*, **144**(3):519–532, 1999. 370

[478] Maria Cruz Martinez, Tomoyo Ochiishi, Michael Majewski, and Kenneth S Kosik. **Dual regulation of neuronal morphogenesis by a delta-catenin-cortactin complex and Rho.** *The Journal of Cell Biology*, **162**(1):99–111, 2003. 370

[479] B D Angst, C Marcozzi, and A I Magee. **The cadherin superfamily: diversity in form and function.** *Journal of Cell Science*, **114**(Pt 4):629–641, 2001. 370

[480] T Takeuchi, A Misaki, S B Liang, A Tachibana, N Hayashi, H Sonobe, and Y Ohtsuki. **Expression of T-cadherin (CDH13, H-Cadherin) in human brain and its characteristics as a negative growth regulator of epidermal growth factor in neuroblastoma cells.** *Journal of Neurochemistry*, **74**(4):1489–1497, 2000. 370

[481] Alejandro Arias-Vsquez, Marieke E Altink, Nanda N J Rommelse, Dorine I E Slaats-Willemse, Cathelijne J M Buschgens, Ellen A Fliers, Stephen V Faraone, Joseph A Sergeant, Jaap Oosterlaan, Barbara Franke, and et al. **CDH13 is associated with working memory performance in Attention Deficit/Hyperactivity Disorder.** *Genes brain and behavior*, **10**(8):844–51, 2011. 370

[482] C.L. Simpson, V.K. Hansen, P.C. Sham, A. Collins, J.F. Powell, A. Al-Chalabi, et al. **MaGIC: a program to generate targeted marker sets for genome-wide association studies**. *Biotechniques*, **37**:996–999, 2004. 373

[483] Ellen Goessling, Olga V. Kel-Margoulis, Alexander E. Kel, and Edgar Wingender. **MATCH^{TM} - a tool for searching transcription factor binding sites in DNA sequences. Application for the analysis of human chromosomes.**, October 2001. 405

[484] S. Djebali, C.A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, et al. **Landscape of transcription in human cells**. *Nature*, **489**(7414):101–108, 2012. 410

[485] R.E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M.T. Maurano, E. Haugen, N.C. Sheffield, A.B. Stergachis, H. Wang, B. Vernot, et al. **The accessible chromatin landscape of the human genome**. *Nature*, **489**(7414):75–82, 2012. 410

[486] S. Neph, J. Vierstra, A.B. Stergachis, A.P. Reynolds, E. Haugen, B. Vernot, R.E. Thurman, S. John, R. Sandstrom, A.K. Johnson, et al. **An expansive human regulatory lexicon encoded in transcription factor footprints**. *Nature*, **489**(7414):83–90, 2012. 410

[487] S. Neph, A.B. Stergachis, A. Reynolds, R. Sandstrom, E. Borenstein, and J.A. Stamatoyannopoulos. **Circuitry and dynamics of human transcription factor regulatory networks**. *Cell*, 2012. 410