Analysis of Gene Systems across Brain Disorders

Snow Bach, BA (Mod.)

Submitted for the award of
MSc

School of Mathematical Sciences,
Dublin City University

Under the supervision of

Professor Paolo Guasoni
School of Mathematical Sciences, Dublin City University

Professor Daniela Tropea
School of Medicine, Trinity College Dublin

September 2021

Declaration of Authorship

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Master of Science is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:

ID No. : 19212013

Date:

Acknowledgements

I express my gratitude to Professor Paolo Guasoni and Professor Daniela Tropea for the unique opportunity to work with both their research groups. Their support and guidance have been invaluable for the entirety of this research project. They are a great source of inspiration.

I would also like to thank the Neuropsychiatric Genomics Group within the Trinity Translational Medicine Institute for all their help and insightful discussions. Thank you to Dr Niamh Ryan for all the encouragement and enthusiastic lessons in coding.

Gracias por el apoyo incondicional. El corazón rebosa de amor.

Cảm ơn bạn vì sự hy sinh. Con không bao giờ có thể cảm ơn đủ.

Contents

List of Figures

List of Tables

Abbreviations

AD	Alzheimer's Disease
ADHD	Attention Deficit Hyperactivity Disorder
ASD	Autism Spectrum Disorder
ASD-SFARI	Autism genes from SFARI database
BIP	Bipolar Disorder
ChIP-Seq	Chromatin Immunoprecipitation Sequencing
DEG	Differentially Expressed Genes
eQTL	Expression Quantitate Trait Loci
GO	Gene Ontology
GWA	Genome Wide Association
HTT	Huntington's Disease
MDD	Major Depressive Disorder
MeCP2	Methyl-CpG Binding Protein 2
MS	Multiple Sclerosis
PD	Parkinson's Disease
PWM	Position Weight Matrix
RNA-Seq	Transcriptomic RNA-Sequencing
RTT	Rett Syndrome
SCZ	Schizophrenia
SNP	Single nucleotide polymorphisms

Abstract

Analysis of Genes Systems across Brain Disorders

Snow Bach

This thesis analyses data from genetic and epigenetic studies of brain disorders, in order to establish potential convergences of mechanisms across different conditions. Current research highlights the common symptoms across a wide range of brain disorders. We analyse the properties of the gene regulator: Methyl-CpG binding protein 2 (MeCP2), a chromatin-binding protein and a modulator of gene expression and we establish a DNA binding model: Matrix-GC, to predict MeCP2 targets. We evaluate Matrix-GC's performance using receiver operating characteristic curves while varying a determinant binding factor: guanine-cytosine nucleotide enrichment (GC content). We show by combining a DNA binding sequence with GC content, that Matrix-GC is able to capture genes bound by MeCP2 better than random chance and binding sequence alone. Matrix-GC is applied to various brain disorders associated with MeCP2, followed by downstream enrichment analysis of molecular pathways and processes. We show three main processes to be under the control of MeCP2 across several brain disorders: neuronal transmission, development, and immunoreactivity. We further validate the performance of Matrix-GC at the single gene level by comparing MeCP2-bound genes with existing high-throughput transcriptome analysis and show that our results are statistically significant. We carry out stringent control analysis by Monte Carlo permutation to strengthen the reliability of our results. We propose the Matrix-GC as an *in silico* procedure to identify putative MeCP2 target genes and shed light on mechanisms overlapping across different brain disorders. Our method of identifying target genes has broad applications and can be implemented with other proteins that influence gene regulation. Importantly, this research provides a framework for analysing genetic data with statistical rigour which can be applied to downstream gene set analysis.

Chapter 1
Introduction

1.1 General Introduction

The field of genetics combines several areas of research in order to make sense of traits, either fully or partially inherited. Classical or transmission genetics is the oldest branch studying the passing of traits from parent to offspring, while molecular genetics focuses on function and structure. On the macro scale, population genetics looks at the movement and changes of genes within populations. An emerging branch of genetics is epigenetics. Epigenetics studies how traits are affected by non-DNA factors; that is, how environmental factors influence genes rather than the traditional dogma of genes alone influencing traits [1]. Common brain disorders show a degree of heritability through genetic variation, and this can be utilised to investigate key mechanisms involved in pathology. In particular, this research focuses on methyl-CpG binding protein 2 (MeCP2), an epigenetic regulator that controls the expression level of genes. MeCP2 is implicated in normal neuron development and has been associated with several neuropsychiatric disorders [2], making this protein an ideal candidate to study different gene systems.

A wealth of genetic data is now available and made open-access thanks to high throughput methods. This provides the opportunity to conduct our own analyses and thoroughly interrogate results. While upholding the principle of reproducibility, it is equally important to scrutinise existing data with statistical rigour.

1.1.1 Neuropsychiatric Disorders and Heritability

Neuropsychiatric disorders are a group of mental disorders affecting cognition and behaviour. Examples of common neuropsychiatric disorders are autism spectrum disorders and attention deficit hyperactivity disorder, while less common are schizophrenia and bipolar disorder.

Neuropsychiatric disorders are typically complex in genetic origin and are attributed to many genes of varying heritability. As such, they have been studied mainly through common gene mutations – genetic variants that occur and confer low risk [3]. This is rooted in the common disease- common variant hypothesis, whereby high prevalence of a disorder in a population is marked by high frequency gene variants, but with lower probability of an individual expressing disease traits. Alternatively, rare mutations that result in disease are typically associated with a single gene and are considered high-penetrance e.g. Rett Syndrome or Fragile X Syndrome. Heritability quantifies the variation of traits within a population as a result of genetic differences. By calculating heritability statistics, we can estimate that on average, a certain percentage of individual differences come from genetic differences. And this gives rise to endophenotypes – behavioural traits that can be traced back to genetic roots – in attempt to simplify heritability and focus on functional consequences [4].

Given that neuropsychiatric brain disorders share symptoms, researchers question whether or not they have overlapping genetic heritability. The Brainstorm Consortium report shared genetic risks between several disorders: attention deficit hyper-activity disorder, bipolar disorder, major depressive disorder, and schizophrenia [5]. Contrastingly, neurological disorders such as Parkinson's disease or Alzheimer's disease do not show the same degree of correlation, suggesting they are aetiologically distinct.

1.1.2 Current Methods in Genetics

Data generated from genetic studies and sequencing techniques have provided a wealth of information for geneticists to investigate disease and trait heritability. High throughput methods have greatly improved the quality and quantity of genetic data [6]. We have come leaps and bounds in under 50 years with modern DNA sequencing. To put this into perspective, the first human genome sequencing endeavour lasted a year, completing in 2001 and was backed by a $3 billion plan [7]. Today, sequencing costs are in the thousands and take under a day to complete in addition to the increase in sequencing depth capabilities and efficiency [6]. With the advent of single-cell sequencing, differences in gene expression can now be precisely captured at the cell level [8,9]. With traditional or bulk sequencing, the disadvantage is the averaging of gene expression over a heterogeneous cell population. One answer to this might be, as mentioned above, single-cell sequencing (SC-Seq) which is becoming the new gold standard to measure gene expression at the cell level. This holds great potential for research into cell dynamics. Indeed, during the COVID-19 pandemic, researchers are moving en masse to SC-Seq to capture different gene profiles in different cell populations to provide a comprehensive disease map. For now, this cutting edge technology comes with its own drawbacks, be it higher noise, increased variability or a low capture rates compared to traditional sequencing [8].

Another powerful tool is the genome-wide association (GWA) study that allows us to identify genes associated with a trait through single nucleotide polymorphisms (SNPs). Our understanding of complex disorders, whose aetiology is comprised of cumulative common variations found among the population, benefits greatly from GWA studies. As of May 2021, there are over 5000 unique GWA studies across different disorders [10]. However, criticisms of GWA studies focus on the functional relevance that may or may not come with statistical association [11] especially as more common variants are implicated in diseases. For example, the “missing” heritability of complex diseases such as neuropsychiatric disorders have yet to be explained by GWA studies [3] and instead, proposes to include rare gene variants that have higher probability of disease expression.

Despite advances in the sequencing technology, there remains another dimension of genetic data that needs to be explored: epigenetics. Epigenetics is the study of how gene expression is altered when there is no change to the DNA sequence. Epigenetic regulation is particularly important in development, where spatiotemporal interactions are critical in producing the correct cells in the right place at the right time. Epigenetic regulations occur by several processes such as chemical modifications, structural remodelling and regulatory elements [1]. Indeed, aberrant epigenetic control can have functional and behavioural consequences in the nervous system [12]. This is particularly true for complex traits that are polygenic and hypothesized to be controlled by many genes. Currently the study of epigenetics has yet to illuminate underlying mechanism of neuropsychiatric disorders. However, given how development and in particular, brain development is mediated by epigenetic processes, this warrants further research.

Systems biology focuses on combining biological systems to understand activity within a contained environment [13]. -Omic studies provide large scale data in order to quantify an entire set of molecules in an organism. Transcriptomics quantifies the expression levels of gene transcripts – messenger RNA molecules that represent genetic information, while proteomics represents the set of proteins produced. Combining various -omic studies i.e. taking a multiomic approach, provides a holistic overview of complex biological systems [14,15]. Furthermore, these studies can be supported by ontological databases which produce controlled and common lexicon that informs on the role of genes and their products as with Gene Ontology [16]. Nowadays, there is an increasing number of online tools and applications available to the traditional biologist, creating a new environment for integrative research rather than biological elements in isolation from one another [17]. Online software and applications with systems biology exist in mind to browse, visualise and analyse pathway data such as Reactome, Cytoscape and WikiPathways [18–21]. These tools can be used in -omics studies to provide a broader, high-dimensional overview of biological processes especially in disease states [15].

With this unprecedented availability of open access experimental data, theoretical analysis and mathematical modelling in molecular biology has provided new abstractions in biological data [22]. Mathematical modelling is particularly important in describing biological networks with increasing accuracy and assigning meaningful relationships between its components. There are many other public databases that are available and by taking existing data in repositories or even entire models such as biological pathway databases, useful insight can be provided in building a new model in a knowledge-based approach [22,23]. Analysis of gene signalling networks are limited by the inclusion of a given set of genes in a pathway while forgoing their respective position in such pathways. Additionally, genes with multiple functions can have diverse roles in any given pathway [13]. A valuable challenge is in understanding what interactions occur between genes and how they can be exploited in pathway analyses to provide both a comprehensive and meaningful interpretation in pathological states. From a transcriptomic point of view, this problem is then reiterated in proteomics and metabolomics. For example, when considering protein-gene interactions in transcriptional regulation, the role of cofactors and coactivators is an important one which prompts further investigation in pathological states [24].

1.2 Methyl-CpG Binding Protein 2

MeCP2 is a protein that controls gene expression levels through direct and indirect mechanisms. Generally, it acts as repressor by binding methylated CpG dinucleotides to modify chromatin, the condensed DNA-protein complex that ravels and unravels in response to specific proteins. However, it may also work as an activator by interacting with specific co-factors such as CREB1. [25].

DNA sequences are transcribed into messenger RNA; this process is facilitated by transcription factor (TF) proteins which bind to DNA and regulates the rate of transcription. TFs have binding domains which allow for specific binding along DNA promoter sequences to initiate transcription. We look at MeCP2 as a transcription factor that binds to DNA and apply a motif discovery algorithm which incorporates features that promote binding of biologically relevant sites. We also assess the performance of this model and assume certain parameters for selective binding in gene sets associated with brain disorders.

In our analysis we use a position weight matrix (PWM) to represent the DNA motifs most likely to be bound by MeCP2 allowing for redundancy in binding site specificity. PWMs are generated by combining experimental data for TFs binding DNA sequences [26] and provides an *in silico* approach to predict binding sites. Prediction methods, in particular, binary classifiers are useful in predicting useful features of datasets [27]. The traditional PWM model can be placed under this category and can be evaluated by multiple parameters to assess how accurately it predicts transcription factor binding sites.

1.2.1 MeCP2 and Brain Disorders

The *MECP2* gene is associated with Rett Syndrome (RTT): a progressive X-linked neurological disorder that primarily affects females. However, certain RTT traits or phenotypes have been observed in other neurodevelopmental syndromes. Perhaps this is mediated by MeCP2, or recapitulated through common neural pathways since MeCP2 has been shown to interact with causative genes in disorders such as Angelman Syndrome [28], Fragile X syndrome [29], and Tuberous Sclerosis [30].

Moreover, there has been a strong association between RTT and Autism Spectrum Disorders (ASD) as RTT was originally classified as part of ASD. Abnormal social behaviour, stereotypies, language use are observed in RTT and ASD. suggesting common physiology. However, ASD is complex in genetic origin with multiple genes implicated in its genetic architecture; of which, mutations in *MECP2* have been reported in patients with autism [31].

MeCP2 is involved in other neuropsychiatric and neurological conditions, although this is less studied compared to neurodevelopmental syndromes [32]. Dysregulation of MeCP2 has functional consequences in neuronal development [2] and is thought to be mediated in part by Brain-derived Neurotrophic Factor (BDNF). BDNF is a neurotrophin regulated by MeCP2 and is involved in brain development, function and connectivity [33]. BDNF-related mechanisms are dysregulated as a result of *MECP2* mutations, and BDNF altered expression has been detected in several brain disorders, including neurodevelopmental disorders, depression, and anxiety [34,35].

More studies highlight shared genetic variations across different neuropsychiatric and neurological disorders [5,36,37], suggesting similar molecular mechanisms could be responsible for common presentations and pathophysiology. To try and clarify potential common mechanism influenced by MeCP2, we examine several brain disorders (Table 1.1).

Disorder	Source	Data Type
Neuropsychiatric		
Autism (ASD)	[38]	Meta-analysis of GWAS
Autism (ASD-SFARI)	SFARI Gene, [39]	Database
Schizophrenia GWAS (SCZ)	[40]	GWAS
Attention deficit hyperactivity disorder (ADHD)	[41]	GWAS and CNV review
Major depressive disorder (MDD)	[42]	Meta-analysis of GWAS
Anorexia	[43]	GWAS
Bipolar Disorder (BIP)	[37]	GWAS
Neurological		
Parkinson's disease (PD)	[44]	Meta-analysis of GWAS
Alzheimer's disease (AD)	[45]	Meta-analysis of GWAS
Huntington's disease (HTT)	[46]	GWAS
Multiple sclerosis (MS)	[47]	GWAS review
Amyotrophic lateral sclerosis (ALS)	[48]	GWAS
Epilepsy	[49]	Meta-analysis of GWAS

Table 1.1: Neuropsychiatric and neurological brain disorders analysed in this research study.

1.3 Aims

The aim of this research thesis is to re-evaluate existing genetic data and attempt to expand our understanding of systems biology through statistically rigorous methods. We do this by investigating one specific protein involved in gene regulation: MeCP2 and its impact on several brain disorders. This project has the potential to contribute specifically to the methodology of binding site modelling and biological systems analysis.

Research disseminated in this thesis contains results previously published in Scientific Reports [50].

Chapter 2
Methods

2.1 *In Silico* Methods of Assessing Gene Binding

2.1.1 Establishing MeCP2 Binding

In order to establish a simple binding model of MeCP2 based on a binding sequence motif, we retrieve a position frequency matrix (PFM) for MeCP2 from the Cistrome database (http://cistrome.org) [51]. We use the Biostrings package in RStudio version 1.1.463, to convert the MeCP2 PFM into a position weight matrix (PWM) used to identify the MeCP2 binding motif along a sequence of DNA (Figure 2.1). Klose and colleagues previously report preferred sequences for MeCP2 binding through methyl-SELEX experiments that produce oligomers which selectively bind to a specific target. [51] They validated their findings with genes known to be bound by MeCP2: *Bdnf* and *Dlx6*.

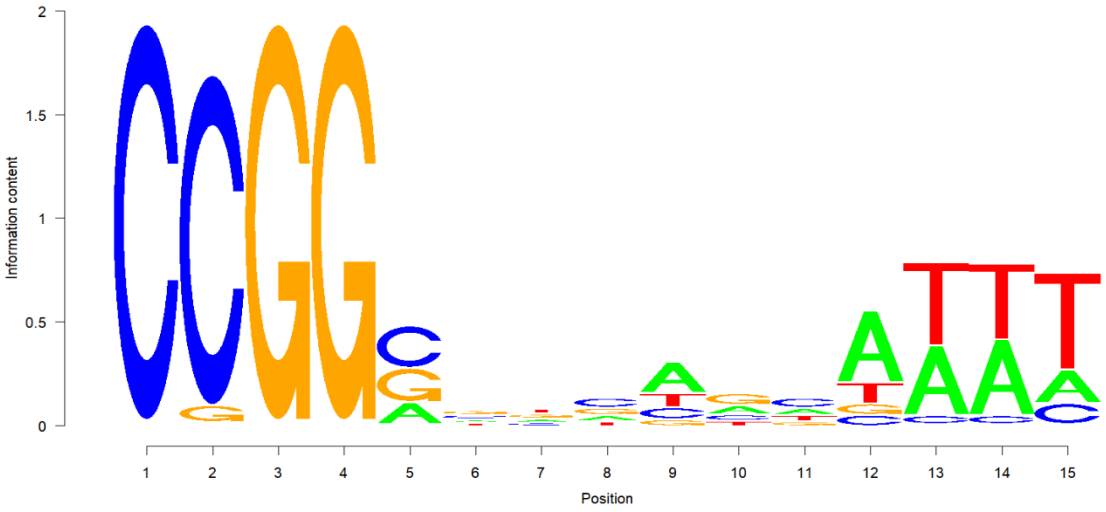

Figure 2.1: The MeCP2 sequence logo represents the conservation of sequence nucleotides for MeCP2.

The calculation of probability for a nucleotide observed along a sequence is given below: [52,53]:

$$p\ (b,i) = \frac{F_{b,i} + s(b)}{N + \sum_{i=1}^{n} s(b')}$$

$p(b,i)$ = probability of nucleotide or base b at position i
$F_{b,i}$ = counts of base b at position i
$s(b)$ = pseudocount function
N = number of sites
$b' \in \{A, C, G, T\}$

The pseudocount function is added to correct for a small sample size of binding sites and can vary with different applications [53,54].

These nucleotide probabilities are put forward into the PWM and calculated as log likelihoods:

$$PWM_{b,i} = \log_2 \frac{p(b,i)}{p(b)}$$

$PWM_{b,i}$ = PWM value of base b at position i
$p(b)$ = background probability of base b

Once there is a proposed binding site along the DNA sequence according to the PWM, a score is given by the summation of PWM values for each base at the given position.

$$S = \sum_{i=1}^{w} PWM_{l_i,i}$$

S = Score of a putative binding sequence
w = width of the PWM
l_i = base at position i of the expected motif

We retrieve MeCP2 target genes from chromatin immunoprecipitation sequencing (ChIP-Seq) data on Cistrome Data Browser (http://cistrome.org/db/). ChIP-Seq evaluates protein-DNA interactions, and by using MeCP2 data, we have a control dataset to assess our binding model. We use two sets of ChIP-Seq data from a study by Maunakea and colleagues (Cistrome ID 34392 & 34399) [55]. MeCP2 target genes on Cistrome are already scored by the BETA package indicating the regulatory potential as a putative target [56].

For positive controls, we generate sequence datasets for the top 100, 200 and 300 genes bound by MeCP2 from the IMR-90 and HCT-116 ChIP-Seq data, ranked by Cistrome BETA scoring. For negative controls, we randomly select and size-match genes from the same ChIP-Seq data with a score of 0. We define the promoter sequences as being 1000bp upstream of the transcription start site and retrieve these promoters in RStudio from the UCSC Genome Browser using the GRCh37/hg19 human reference genome. We test each sequence for the presence of the MeCP2 PWM. For every PWM match, a score is given from 0-1. This score represents how similar the motif of the PWM is on the selected sequences, compared to a random sequence.
Guanine-Cytosine nucleotide content (GC%) was previously established to be important in MeCP2 binding *in vivo* [57]. For every PWM match, we generate a sequence to include the 15bp PWM match sequence and 100bp flanking sequences, and we calculate the GC% for these 215bp sequences.

2.1.2 Receiver Operating Characteristics Curve

In order to determine the ideal PWM threshold for MeCP2 motif binding, we graph a receiver operating characteristics (ROC) curve for all datasets (Figures 3.1 and 3.2) and calculate the area under the curve (AUC). We set the minimum PWM score at 5% and stratify results based on PWM scores at increasing increments of 5%. We generate 10 random bootstrapped samples of 100, 200 and 300 negative control genes. Taking the average values, we plot the ROC curve alongside the positive controls and evaluate the results at various sequence GC%.

2.1.3 Dataset Collection

For analysis of MeCP2 interaction in different disorders, we use neuropsychiatric and neurological disorders data from multiple studies to apply our binding model (Table 2.1). For ASD-SFARI dataset [39] Gene Scoring - which assesses the strength of evidence presented for candidate ASD genes - we consider only categories S (syndromic), 1 (high confidence genes) and 2 (strong candidate genes).

For SCZ-associated genes, we use genes identified by GWAS studies. Rare variants that occur at low frequency in the population are also implicated in SCZ aetiology. However, few SCZ candidate rare variants have been confirmed with sequencing to date. Neurexin 1 is a well-known rare variant in SCZ and is also largely associated with ASD [58]. To date, SETD1A is the only genome-wide significant rare variant discovered by whole exome sequencing [59]. The identification of rare variants in SCZ is controversial, so to avoid false positive results, we do not consider SCZ rare variants in our analysis.

2.2 Methods for Functional Characterisation of Gene Sets

2.2.1 Enrichment Analyses

We employ Gene Ontology and pathway overrepresentation analysis to define functional groups of the disease gene datasets and identify any terms or pathways significantly enriched in these datasets. We carry out Gene Ontology (GO) enrichment analysis using GOrilla (http://cbl-gorilla.cs.technion.ac.il/). Discovery of over-represented GO terms is carried out in the disorder genes versus the background set which ignores ranking. The background set of genes is the entire genome minus the genes of interest. Genes are input as Ensembl identifiers with a selected p-value threshold of 0.05. To perform pathway enrichment analysis within the various gene sets, we use the ReactomePA package from Bioconductor. We input Entrez identifiers into the function call `enrichPathway`, selecting a p-value cut-off value of 0.05, controlling for false discovery rate ("fdr").

2.2.2 Network analysis

We utilise Cytoscape application version 3.7.1 [18] to carry out network analysis and visualise proteomics data from STRING (http://string-db.org) [60] using the stringAPP plugin version 1.4.2 (http://apps.cytoscape.org/apps/stringapp) [61]Ddd. Ensembl gene identifiers are input to identify any protein-protein interactions, either directly or indirectly. We apply a confidence cut-off value of 0.4 with 0 additional interactors.

2.2.3 Synaptic and Immunity Genes

To understand MeCP2 target genes involved in synaptic and immune function, we use genes as categorised by Lips and colleague [62] and genes as categorised from the ImmPort data repository (https://www.immport.org/shared/genelists), respectively, and analyse these genes using our MeCP2 binding protocol

2.2.4 Control Analysis

We generate 1000 samples of 10, 20, 50, 100, 200, 500 and 750 genes from all genes and protein-coding genes in the GRCh37/hg19 human reference genome, referred to as the genome and exome, respectively. We repeat the enrichment analysis with the 1000 samples using the ClusterProfiler [63] and ReactomePA [64] R packages. We set the p-value minimum to 1 to capture all possible results. We use the Fisher's Method [65] to combine the p-values across the 1000 trials for every pathway or GO term result:

$$X_{2n}^2 = -2\left(\sum_{i=1}^{n} \ln(P_i)\right)$$

X^2 = Chi-square test statistic
P = p-values from independent tests
$2n$ = degrees of freedom

Any overlapping significant terms and pathways from the analysis of the control datasets were excluded from the results.

For network analysis, we use randomly generated control datasets as above. We calculate the average degree of network connectivity related to the size of the datasets and to generate a range of values network connectivity associated to the size of the datasets. We run the Cytoscape analysis 20 times and we select the maximum and minimum degree of connectivity for each dataset size. This information was used to identify the brain disorders associated gene sets with a level of network connectivity different from what expected by chance. We are then able to designate protein hubs as nodes with a degree of connection, 1 degree greater than the average level of connectivity of the corresponding control gene sets size.

2.3 Validation Methods

2.3.1 SNPs in Different Brain Disorders

To identify the presence of *MECP2* single nucleotide polymorphisms (SNP) in the brain disorders, we download human *MECP2* SNP data from NCBI dbSNP (https://www.ncbi.nlm.nih.gov/snp/). We also look at Matrix-GC-derived genes and investigated if SNPs were present within the 1000bp promoter site. We compare *MECP2* SNPs to SNP from our brain disorder datasets of interest using data from NCBI ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/). We retrieve sex information of patients with these SNPs from the RettBASE: RettSyndrome.org Variation Database (mecp2.chw.edu.au).

2.3.2 Tissue Expression Analysis

We then examine the expression levels of each gene from our brain disorder gene sets using NCBI Gene (https://www.ncbi.nlm.nih.gov/gene/). We select "HPA RNA-seq for normal tissues" for analysing protein-coding genes and "RNA sequencing of total RNA from 20 human tissues" for retrieving expression data of non-coding genes. Expression data is represented as reads per kilo base per million mapped reads.

We collate several tissues to represent specific systems for immune, digestive, urinary and reproductive systems. Immune tissues include data from: lymph node, bone marrow, spleen, adrenal, thyroid and appendix data. Digestive tissues includes data from: colon, duodenum, oesophagus, gall bladder, pancreas, liver, small intestine, salivary gland and stomach data. Reproductive tissues include data from: ovary, testis, endometrium, prostate and placenta. Urinary System includes tissues from urinary bladder and kidney data.

For each tissue, we consider a gene to be expressed if its expression level is greater or equal to 1. This convention allows us to obtain, for each disorder-tissue combination, a 2x2 contingency table for gene numbers:

	MeCP2-bound and expressed	MeCP2-bound and not expressed
not MeCP2-bound and expressed	*A*	*B*
not MeCP2-bound and not expressed	*C*	*D*

From this contingency table, the Fisher's exact test [66] calculates the exact statistical significance of the hypothesis that the proportion of expressed genes in the MeCP2-bound group is the same as the proportion of expressed genes in the not MeCP2-bound group, from the hypergeometric distribution.

2.3.3 Transcriptomic Studies

We investigate single tissue expression quantitative trait loci (eQTL) from the Genotype-Tissue Expression Portal (https://gtexportal.org/home/) for cerebellum, cortex, hippocampus tissues and whole blood, and compare them to brain disorder genes that result from our MeCP2 PWM and GC% filter. To validate our results, we use transcriptomic analyses in *Mecp2*-null mice and RTT induced pluripotent stem cells (iPSC). We use data from *Mecp2*-null mice compared to matched wildtype controls [67] considering expression analysis in blood and cerebellum tissues, and data from iPSCs from a patient with Rett Syndrome [68]. Data comes from the Gene Expression Omnibus under entries GSE129387 and GSE123753, and we carry out differential analysis with the EdgeR package (v3.14.0). Genes are not considered in the analysis where all samples show no counts. For the gene set identified by the Matrix-GC procedure, we calculate the percentage of 2-fold changed genes (fold change of ≤0.5 or ≥2). We also calculate the percentage of statistically significant differentially expressed genes (DEGs with p-value ≤ 0.05).

We evaluate the statistical significance of these sets through Monte Carlo permutation, by comparing their statistics to both the percentage of DEGs (p-value ≤0.05) and 2-fold changed genes (fold change of ≤0.5 or ≥2) in 1000 randomly selected sets of genes with equal size. These Monte Carlo samples were

selected from the set of *Mus musculus* orthologue genes for the animal studies, and from brain tissue genes in the human studies.

Chapter 3
Results

3.1 MeCP2 Binds to Brain Disorder Genes

3.1.1 Establishing a High Affinity MeCP2 Binding Procedure

To identify candidate target genes for MeCP2 binding and regulation, we establish a model that considers MeCP2 binding motif, and preferential DNA sequence features. Our procedure implements a PWM used to identify likely binding sequences [51] (Figure 2.1), combined with GC% of the gene promoter.

We validate our PWM + GC% filter by identifying genes bound by MeCP2 and we apply the model to the top scored genes in the MeCP2 ChIP-Seq IMR-90 and HCT-116 data [55]. We generate ROC curves from datasets of 100, 200 and 300 genes to determine if there was a different binding threshold at different ChIP-Seq scoring levels (Figure 3.1).

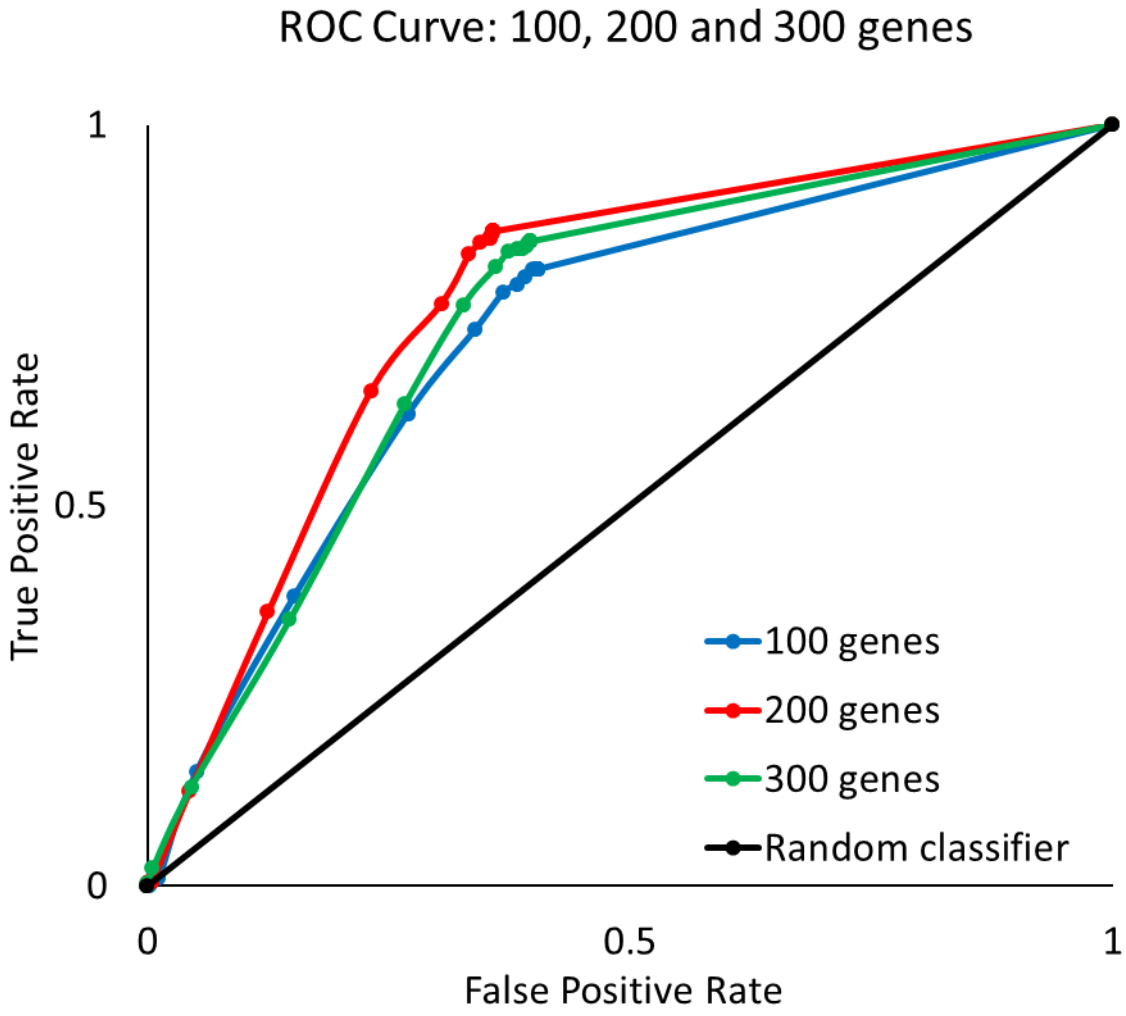

Figure 3.1: ROC curves for 100, 200 and 300 IMR-90 genes establishing a preferential PWM score threshold. The AUC is 0.725, 0.7685, 0.7419, for 100, 200 and 300 genes respectively. The random classifier has an AUC of 0.50.

We identify the ideal threshold score to be 65% for MeCP2 binding across control sets. Since MeCP2 has a higher binding potential for regions containing GC dinucleotide occurrence of ≥ 60% [57], we investigate whether the threshold score of 0.65 changes with different percentages of GC content (Figure 3.2). We combine the PWM filter with an additional GC% filter, varying the GC% threshold from 60%, to 50% and without filtering for GC content (PWM only). We determine that 60% GC content offers a reduction in false positive rate by nearly half (50% GC content: 0.657 vs. 60% GC content: 0.362 false positive rate). We report similar results when using HCT-116 ChIP-Seq data confirm that a GC content of 60% is appropriate and in line with Rube and colleagues' report [57]. In further application to brain disorder gene sets, we use a PWM score of 65% and GC content of 60% (Matrix-GC, Figure 3.3).

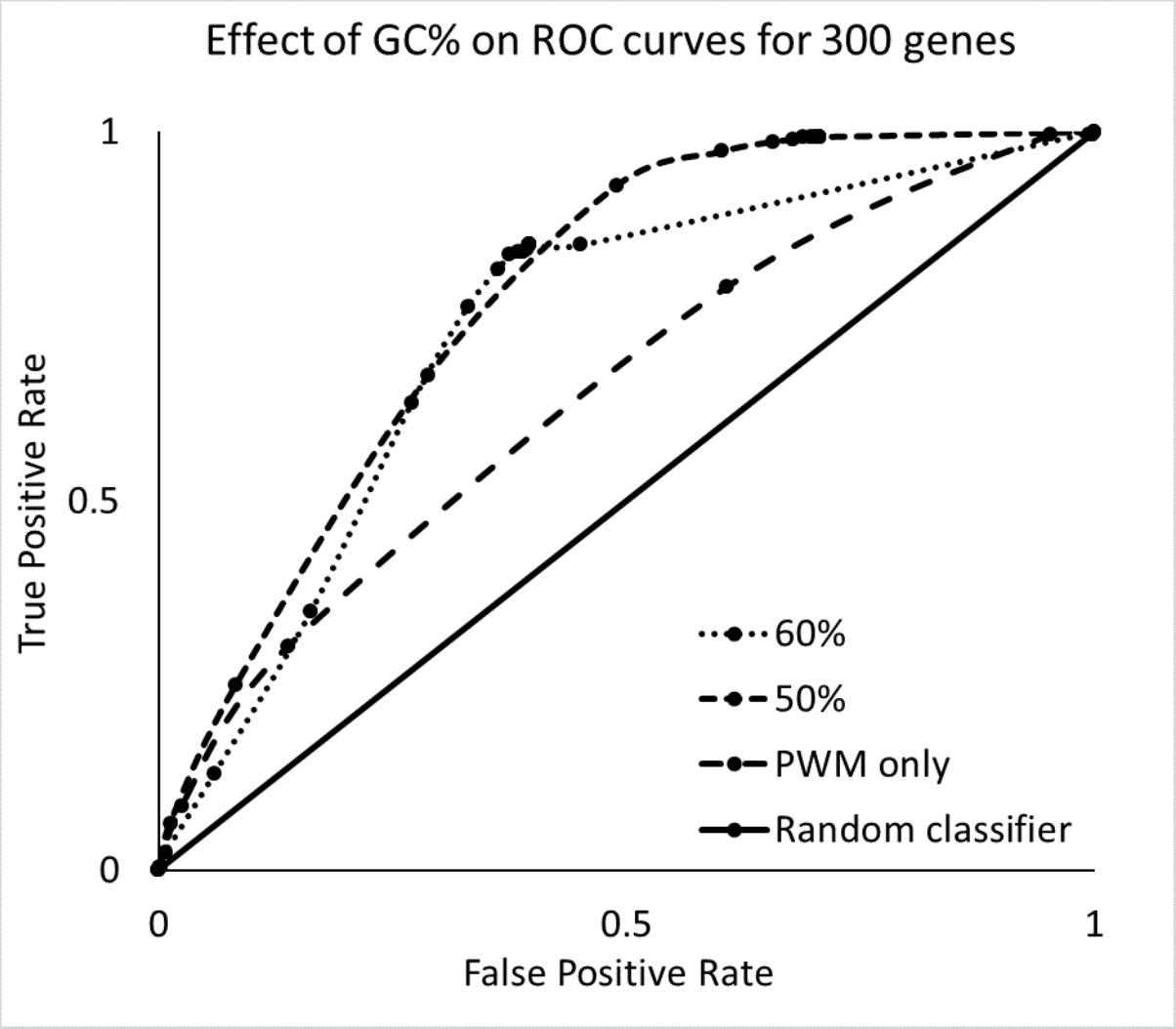

Figure 3.2: ROC curves for 300 IMR-90 genes evaluating the effects of DNA sequence GC content percentage. The AUC values are 0.7301, 0.7692 and 0.6351 for GC content percentages of 50%, 60%, and PWM only, respectively. The random classifier has an AUC of 0.50.

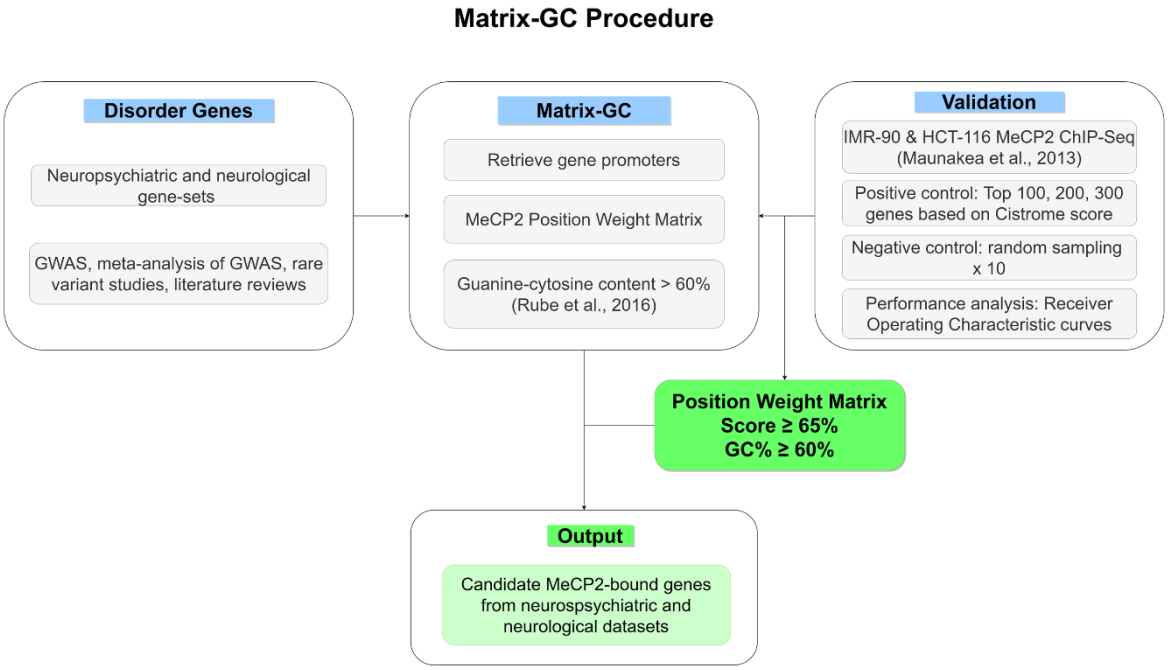

Figure 3.3: Overview of Matrix-GC procedure to detect MeCP2 transcription factor binding sites *in silico*. The Matrix-GC procedure is a combination of an MeCP2 position weight matrix and DNA sequence GC%, and was validated through positive and negative random sampling control using data from [55]. The performance of Matrix-GC was evaluated through ROC curves and is applied to the promoters of candidate genes across neurological and neuropsychiatric disorders.

We also consider MeCP2 binding to *CDKL4* gene as a negative control [69] and *S100A9 as a* positive control [67,70]. Our Matrix-GC procedure captures the same findings in mouse *Cdkl4* and human *CDKL4* orthologue and *S100A9*.

3.1.2 MeCP2 Binds to Brain Disorder Genes Preferentially

We then examine MeCP2's binding potential on gene sets associated with neuropsychiatric and neurological disorders (Table 1.1). All datasets have at least 50% of genes bound by Matrix-GC. Neurological datasets show an overall lower average percentage of MeCP2-bound genes (55.95 %) compared to neuropsychiatric disorders (67.58%). We also consider the entire genome and apply Matrix-GC to all genes in the GRCh37/hg19 human reference genome without discrimination for gene grouping, and we report an average of 39.56% genes bound by MeCP2 *in silico* across the genome (Figure 3.4).

By looking at the genome, we show that MeCP2 binds the promoters of genes associated with brain disorders more often than expected by chance.

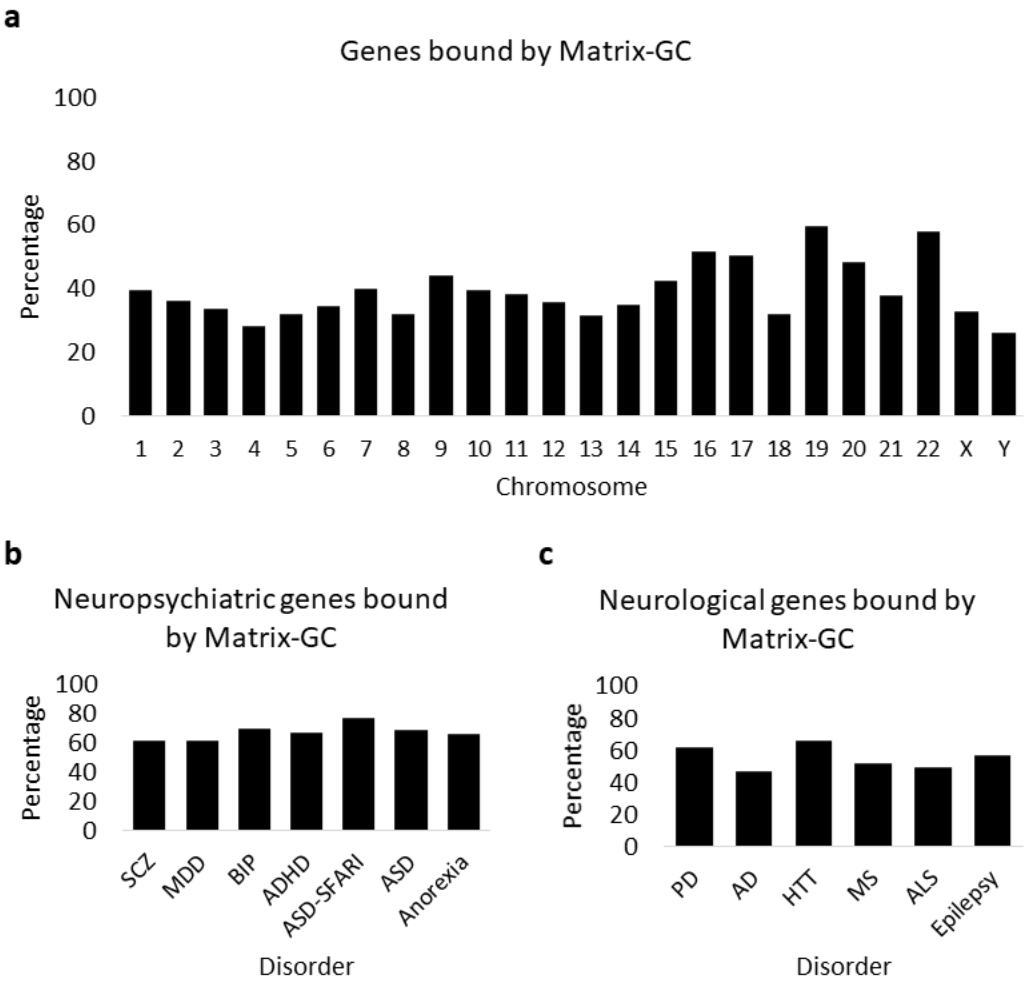

Figure 3.4: Genes bound by MeCP2 *in silico* in the genome and brain disorder datasets. **a** Percentage of genes bound by the Matrix-GC procedure in the genome. **b** Percentage of genes bound by the Matrix-GC procedure in the neuropsychiatric datasets: schizophrenia (SCZ), major depressive disorder (MDD), bipolar disorder (BIP), attention deficit hyperactivity disorder (ADHD), Autism genes from the SFARI database (ASD-SFARI), autism (ASD), and anorexia. **c** Percentage of genes bound by our Matrix-GC procedure in the neurological datasets: Parkinson's disease (PD), Alzheimer's disease (AD), Huntington's disease (HTT), multiple sclerosis (MS), amyotrophic lateral sclerosis (ALS), and epilepsy.

3.2. Functional Characterisation of Brain Disorders

3.2.1 Protein-Protein Interaction Network Analysis

We perform network analysis through Cytoscape using the StringAPP on MeCP2-bound genes filtered with the Matrix-GC procedure to map and identify central proteins or nodes that are highly connected in each disorder. First, we generate control gene sets to identify the average degree of network connectivity and use this information to identify the Matrix-GC gene sets with a significant degree of connectivity (Figure 3.5). AD, ADHD, MS, and ASD-SFARI datasets show statistically significant connected networks before and after Matrix-GC and we identify hub proteins from these datasets. (Table 3.1). After Matrix-GC, ADHD hub proteins are associated with neurotransmission processes and different neurotransmitter systems such as DRD1, DRD4, DRD5 dopamine receptors, and GRM5, GRIN2B glutamate receptors. MS hub proteins are involved in inflammatory responses such as TYK2, STAT3, CD40. Hub proteins in AD are generally associated with cell communication while in ASD-SFARI, the most connected proteins are involved in DNA processes, namely transcription. Notably, EP300 is a hub protein with the highest degree overall. EP300 is a histone acetylase protein regulated indirectly by MeCP2, likely via MEF2C [71]

Here, we report hub proteins belonging to three main functions: inflammatory responses, transcription regulation and neurotransmission across 4 brain disorders.

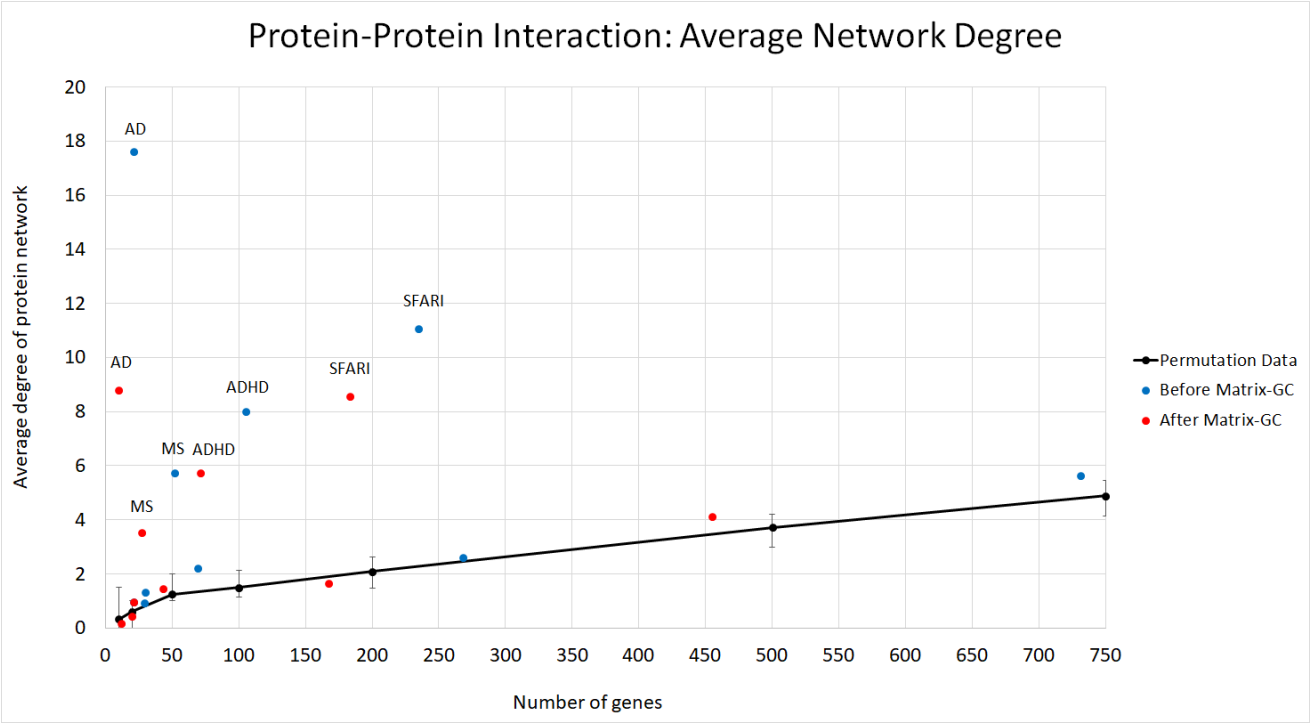

Figure 3.5: Analysis of protein network complexity in brain disorders datasets before and after Matrix-GC. We identify the average degree of connectivity. The plot line shows the mean of the average network connectivity for each control group (black). Lower and upper error bars represent the respective minimum and maximum of the average network degrees from each gene set permutation trials. Significant protein-protein networks for each gene set size are those excluded by the maximum and minimum interval for each gene set size: Autism genes from the SFARI database (SFARI), multiple sclerosis (MS), Alzheimer's disease (AD), and attention deficit hyperactivity disorder (ADHD). Significant networks are labelled above their respective plot point.

Node	Degree	Name	Disorder
EP300	45	Protein propionyltransferase p300	ASD-SFARI
CHD8	44	Chromodomain helicase DNA binding protein 8	ASD-SFARI
MECP2	31	Methyl CpG binding protein 2	ASD-SFARI
PTEN	30	Mutated in multiple advanced cancers 1	ASD-SFARI
KMT2C	28	Myeloid/lymphoid or mixed-lineage leukemia protein 3	ASD-SFARI
SIN3A	27	SIN3 transcription regulator family member A	ASD-SFARI
KDM6A	26	Ubiquitously-transcribed X chromosome tetratricopeptide repeat protein	ASD-SFARI
GRIN2B	26	Glutamate receptor, ionotropic, N-methyl D-aspartate 2B	ASD-SFARI
UBE3A	23	Human papillomavirus E6-associated protein	ASD-SFARI
SETD1B	23	Histone-lysine N-methyltransferase SETD1B	ASD-SFARI
SLC6A3	22	Solute carrier family 6 (neurotransmitter transporter), member 3	ADHD
GRM5	21	Glutamate receptor, metabotropic 5	ADHD
GRIN2B	19	Glutamate receptor, ionotropic, N-methyl D-aspartate 2B	ADHD
NPY	17	Pro-neuropeptide Y	ADHD
NTRK2	16	Neurotrophic tyrosine kinase, receptor, type 2	ADHD
DRD4	16	Dopamine receptor D4	ADHD
SNAP25	13	Synaptosomal-associated 25 kDa protein	ADHD
HTR1B	12	5-hydroxytryptamine (serotonin) receptor 1B, G protein-coupled	ADHD
DRD5	12	D1beta dopamine receptor	ADHD
CNTNAP2	12	Contactin associated protein-like 2	ADHD
CD2AP	9	Cas ligand with multiple SH3 domains	AD
ZCWPW1	9	Zinc finger CW-type and PWWP domain containing 1	AD
INPP5D	9	Phosphatidylinositol 3,4,5-trisphosphate 5-phosphatase 1	AD

Node	Degree	Name	Disorder
CELF1	9	Embryo deadenylation element-binding protein homolog	AD
BIN1	9	Box-dependent myc-interacting protein 1	AD
RIN3	9	Ras interaction/interference protein 3	AD
ABCA7	9	ATP-binding cassette, sub-family A (ABC1), member 7	AD
SORL1	9	Low-density lipoprotein receptor relative with 11 ligand-binding repeats	AD
TYK2	8	Non-receptor tyrosine-protein kinase TYK2	MS
STAT3	8	Signal transducer and activator of transcription 3 (acute-phase response factor)	MS
MYC	8	V-myc avian myelocytomatosis viral oncogene homolog	MS
CLEC16A	8	C-type lectin domain family 16, member A	MS
CD40	8	Tumor necrosis factor receptor superfamily member 5	MS
IRF8	7	Interferon consensus sequence-binding protein	MS
IL7R	7	Interleukin-7 receptor subunit alpha	MS
TNFAIP3	6	Tumor necrosis factor, alpha-induced protein 3	MS
MAPK1	4	Extracellular signal-regulated kinase 2	MS
CD6	4	T-cell differentiation antigen CD6	MS

Table 3.1: Top 10 hub proteins in significant brain disorder datasets. Top 10 hub nodes, from protein-protein interaction analysis taken from significant brain disorder gene sets, ranked by node degree. Network degrees obtained from Cytoscape and StringApp Plugin after Matrix-GC: autism genes from SFARI (ASD-SFARI), attention deficit hyperactivity disorder (ADHD), Alzheimer's disease (AD) and multiple sclerosis (MS). A node is designated as hub if it is 1 greater than the average degree of the corresponding control dataset.

3.2.2 Enrichment Analysis

We carry out GO and pathway enrichment analysis of brain disorder gene sets following the application of Matrix-GC (Figures 3.6 and 3.7).

We find significant GO Biological Process terms in ADHD, AD and ASD-SFARI gene sets before and after the binding procedure. SCZ GO terms are significant after Matrix-GC only. Epilepsy, MDD, MS, and PD-related genes show significant terms prior to Matrix-GC only. Here we report terms relating to neuronal growth, differentiation and nervous system development are significantly enriched in both ADHD and SCZ datasets. Under the regulation of MeCP2, events in the developing brain can be disturbed, contributing to SCZ phenotypes. Additionally, ADHD-related genes show significant enrichment in behaviour and learning, cell-cell communication, and catecholamine neurotransmission and metabolism, suggesting that MeCP2 may influence ADHD pathophysiology through pathways resulting in behavioural and learning difficulties in ADHD. AD-related genes detected by the Matrix-GC procedure show significantly enriched terms related to amyloid protein regulation, metabolism, protein filaments and endocytosis. The ASD-SFARI dataset has the highest number of enriched terms before and after the Matrix-GC procedure, and the most significant terms relate to nucleic acid processes. However, enrichment and network analysis based on common variants does not identify terms or pathways in the ASD gene set. It is possible that *MECP2* plays a role in ASD by coordinating functional connectivity and controlling neurotransmitter balance and cell growth as seen in other neuropsychiatric disorders [2].

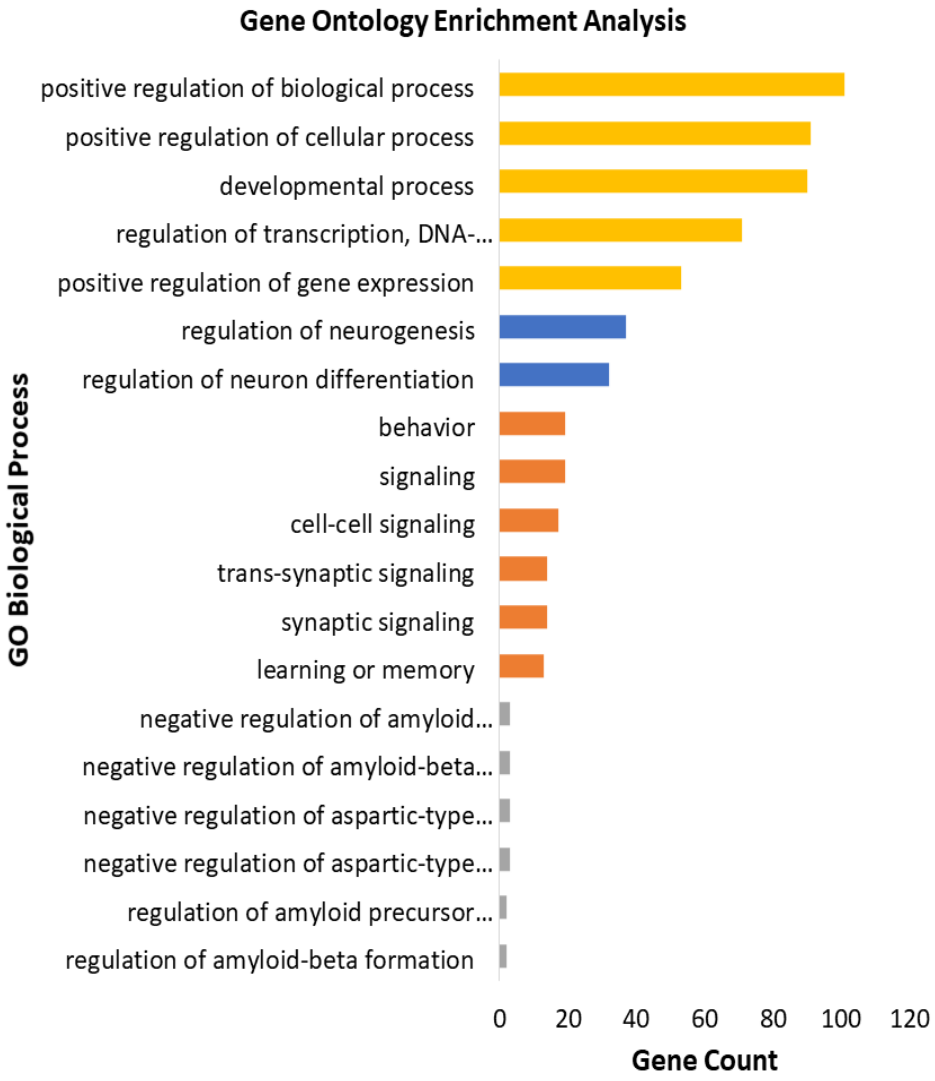

Figure 3.6: Gene Ontology enrichment analysis of neuropsychiatric and neurological disorders. Bar plot of top 5 significant results from Gene Ontology enrichment analysis after applying Matrix-GC. The following disorders are represented: autism genes from the SFARI database (ASD-SFARI, yellow), schizophrenia (SCZ, blue), attention deficit hyperactivity disorder (ADHD, orange), and Alzheimer's disease (AD, grey). Overlapping significant pathways reported in the control analysis are not represented in this plot.

For pathway analysis, we use ReactomePA R package . Only MDD, anorexia and epilepsy gene sets display enriched pathways solely before our Matrix-GC procedure. SCZ and ASD-GWAS gene sets have no enriched pathways either before or after the Matrix-GC procedure. Conversely, PD pathways are significant after Matrix-GC only, and ASD-SFARI, AD, ADHD, ALS, HTT and MS have enriched pathways both before and after the procedure (Figure 3.7).

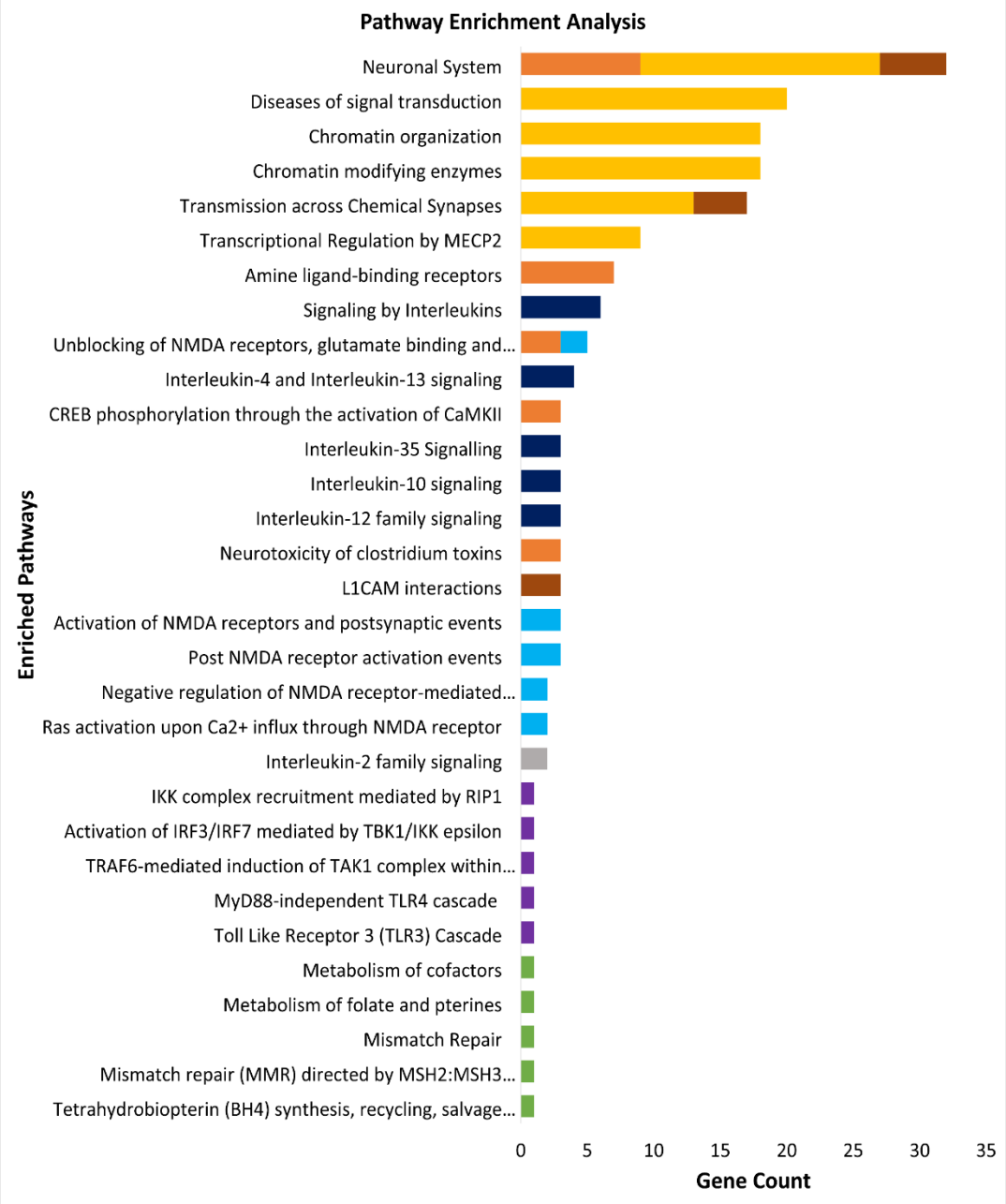

Figure 3.7: Pathway enrichment analysis of neuropsychiatric and neurological disorders. Bar plot of the top 5 significant results from pathway enrichment analysis after Matrix-GC. The following disorders are represented: attention deficit hyperactivity disorder (ADHD, orange), autism database (ASD-SFARI, yellow), bipolar disorder (BIP, brown), multiple sclerosis (MS, dark blue), Alzheimer's disease (AD, grey), Parkinson's disease (PD, light blue), and amyotrophic lateral sclerosis (ALS, purple), Huntington's disease (HD, green). Overlapping significant pathways reported in the control analysis are not represented in this plot.

“Amine ligand-binding receptors” is the most statistically significant pathway across all disorders (ADHD, adjusted p-value = 8.12×10^{-8}). Various glutamate and CREB pathways are also enriched in ADHD genes, while the ASD-SFARI dataset has the highest number of overrepresented pathways associated with chromatin organisation, growth, and neurotransmitter processes. AD and MS datasets are significantly enriched for interleukin signalling pathways before and after Matrix-GC procedure. ALS genes are enriched for Toll-like receptor processes before and after Matrix-GC, while PD-related genes are significantly associated to NMDA-related pathways.

Overall, we assert ADHD, AD and ASD-SFARI gene sets to be highly enriched for GO terms and Reactome pathways, suggesting MeCP2 influence is relevant in these disorders.

3.2.3 The Influence of MeCP2 on Synaptic and Immune Genes

Considering this involvement of MeCP2 in immune responses across different brain pathologies, we also investigate binding to immune genes using Matrix-GC. Overall, MeCP2 binds 41.8% of all immune genes (n=2466) and binds to at least 50% of genes in the cytokine, cytokine receptors, natural killer cell cytotoxicity, Transforming growth factor-beta (TGFb) family member, TGFb family member receptors and tumour necrosis factor (TNF) family members receptors (Figure 3.8). All of interferon receptor genes were captured with Matrix-GC, yet no interferon genes were bound.

We report that MeCP2 binds to 73.51 % of synaptic genes (n=1026). 50% of genes from each category are captured by Matrix-GC, while at least 80% of intracellular trafficking, RNA and protein synthesis, G-protein relay, and tyrosine kinase signalling genes are bound by MeCP2.

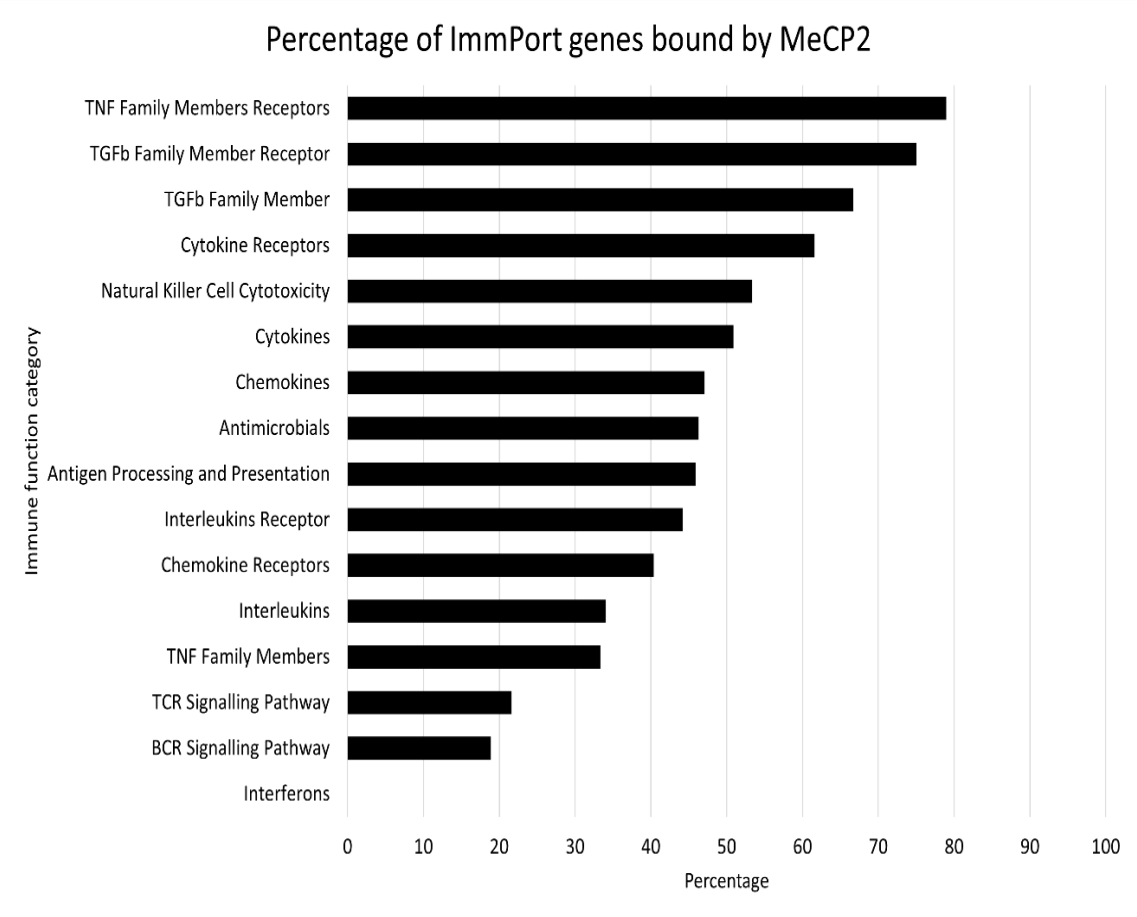

Figure 3.8: Bar plot of the percentage of categorised immune genes bound by Matrix-GC. Abbreviations: TNF = tumour necrosis factor; TGFb = tumour growth factor beta; TCR = T-cell receptor; BCR= B-cell receptor.

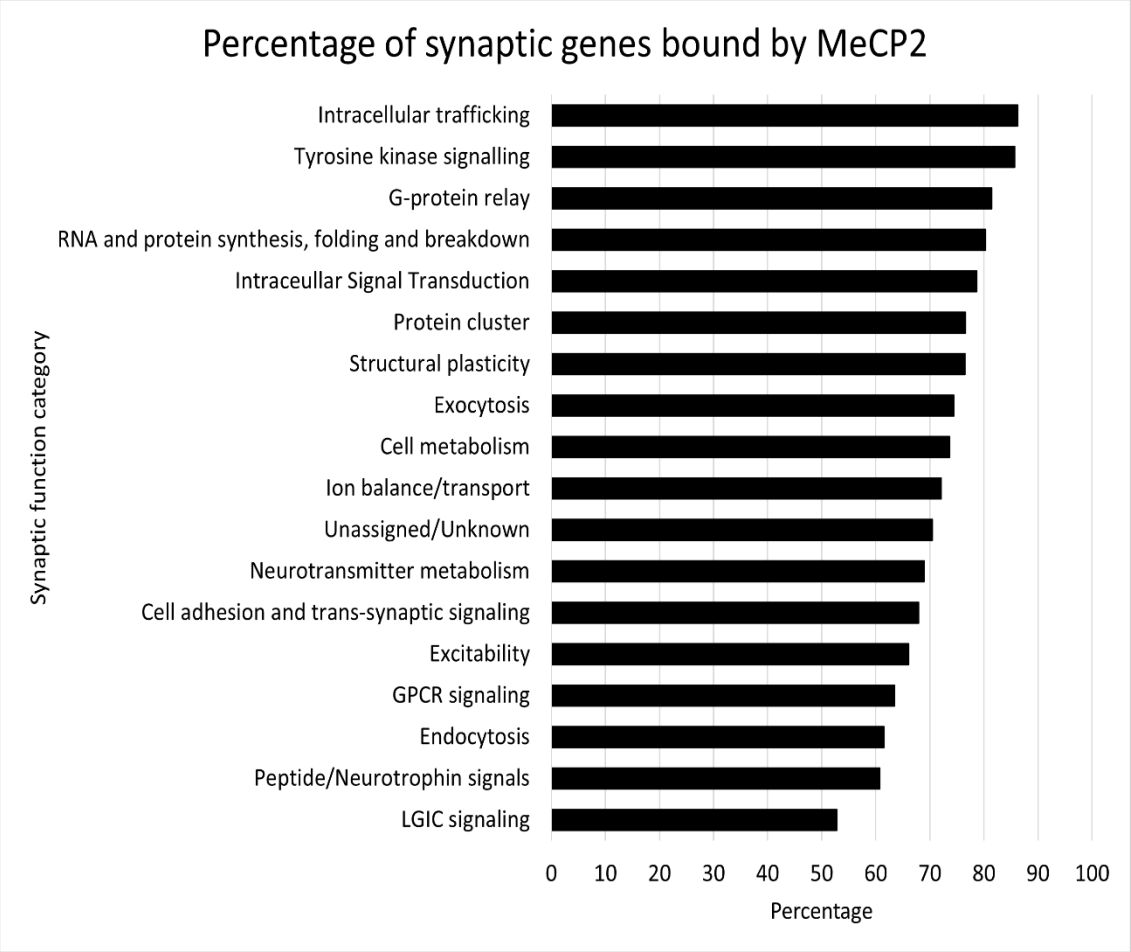

Figure 3.9: Bar plot of the percentage of categorised synaptic genes bound by Matrix-GC. Abbreviations: GPCR= G-protein coupled receptor; LCIC= Ligand-gated ion channels.

3.3. Biological Validation

3.3.1 *MECP2* Mutations are Found in Several Brain Disorders

We find over 13,000 SNPs in the *MECP2* gene from NCBI dbSNP. Several of *MECP2* SNPs are present in four brain disorders: ADHD (2 SNPs), ASD (87 SNPs), epilepsy (25 SNPs) and SCZ (12 SNPs). 16 mutations are shared between at least 2 disorders (Table 3.3)

Nucleotide/ amino acid change	dbSNP ID	Sex	Disorders
NM_004992.3(MECP2):c.872C>G (p.Ser291Cys)	rs1326849768	Unknown	ASD, Epilepsy
NM_001110792.2(MECP2):c.411C >A (p.Ile137=)	rs146107517	Female, Unknown	ASD, SCZ
NM_004992.3(MECP2):c.1178C> T (p.Pro393Leu)	rs267608402	Unknown	ASD, Epilepsy
NM_004992.3(MECP2):c.784C>G (p.Gln262Glu)	rs267608525	Unknown	ASD, Epilepsy
NM_004992.3(MECP2):c.397C>T (p.Arg133Cys)	rs28934904	Female, Unknown	ASD, Epilepsy
NM_004992.3(MECP2):c.473C>T (p.Thr158Met)	rs28934906	Female, Male	ASD, Epilepsy
NM_004992.3(MECP2):c.1189G> A (p.Glu397Lys)	rs56268439	Female, Male	ASD, SCZ
NM_004992.3(MECP2):c.553G>A (p.Gly185Ser)	rs587783135	Unknown	ASD, Epilepsy, SCZ
NM_004992.3(MECP2):c.568C>T (p.Arg190Cys)	rs587783137	Unknown	ASD, Epilepsy, SCZ
NM_001110792.2(MECP2):c.844C >T (p.Arg282Ter)	rs61750240	Female, Male	ASD, Epilepsy
NM_001110792.2(MECP2):c.916C >T (p.Arg306Ter)	rs61751362	Female, Male	ASD, Epilepsy
NM_004992.3(MECP2):c.925C>T (p.Arg309Trp)	rs61751444	Female, Male	ASD, Epilepsy
NM_001110792.2(MECP2):c.1200 _1243del (p.Pro400_Pro401insTer)	rs61752992	Female, Male	ASD, Epilepsy
NM_004992.3(MECP2):c.1214C> T (p.Pro405Leu)	rs61753016	Female, Male	ASD, Epilepsy
NM_004992.3(MECP2):c.1315G> A (p.Ala439Thr)	rs61753973	Female	ADHD, ASD
NM_001110792.2(MECP2):c.1200 _1220delinsCTGAGCCCCAGGACT TGAGCA (p.Pro401_Ser407delinsTer)	rs786205019	Female	ADHD, ASD

Table 3.2: Common *MECP2* SNPs across brain disorders. Single nucleotide polymorphisms on the *MECP2* gene that are associated with several brain disorders: attention deficit hyperactivity disorder (ADHD), autism spectrum disorders (ASD), epilepsy and schizophrenia (SCZ).

3.3.2 Tissue Expression Before and After Matrix-GC

We investigate the tissue expression for each gene and show a statistically significant difference in expression of genes in skin (epilepsy, p-value = 0.009), reproductive (BIP, p-value = 0.02), the brain (MDD, p-value = 0.0006), and immune (MDD, p-value = 0.01; ASD-SFARI, p-value = 0.02) tissues. For SCZ, we show a significant difference in genes bound by Matrix-GC and genes not bound by Matrix-GC in all tissue.

Overall, ADHD genes have the largest increases in percentage after Matrix-GC in brain (15%), fat (20%), and urinary tissues (10%) although not significant. Similarly, expression in immune tissue for epilepsy genes increases by 10% after applying Matrix-GC. We also note differences after Matrix-GC, in lung (12%), heart (17%), and skin (26%), tissues for epilepsy-related genes.

It is also important to consider non-coding genes (ncRNA) in our datasets because although they do not code for a protein, ncRNA may have epigenetic regulatory functions. *PVT1* is the only ncRNA bound by MeCP2 *in silico* in MS. 5 ncRNAs are found in the ADHD dataset: T*MEM161B-AS1, LINC01572, LINC00461, KDM4A-AS1, LINC02060*, of which only the first 3 are bound by Matrix-GC. 28 out of 64 ncRNAs in the SCZ dataset are identified by Matrix-GC. There is no statistically significant difference between the number of genes expressed before or after Matrix-GC. However, we do report that there is an increase in percentage of expressed genes after Matrix-GC in the cerebellum (24.11%) and whole brain (22.32%). The specific role of these ncRNAs is unknown apart from association to disorders in GWAS studies.

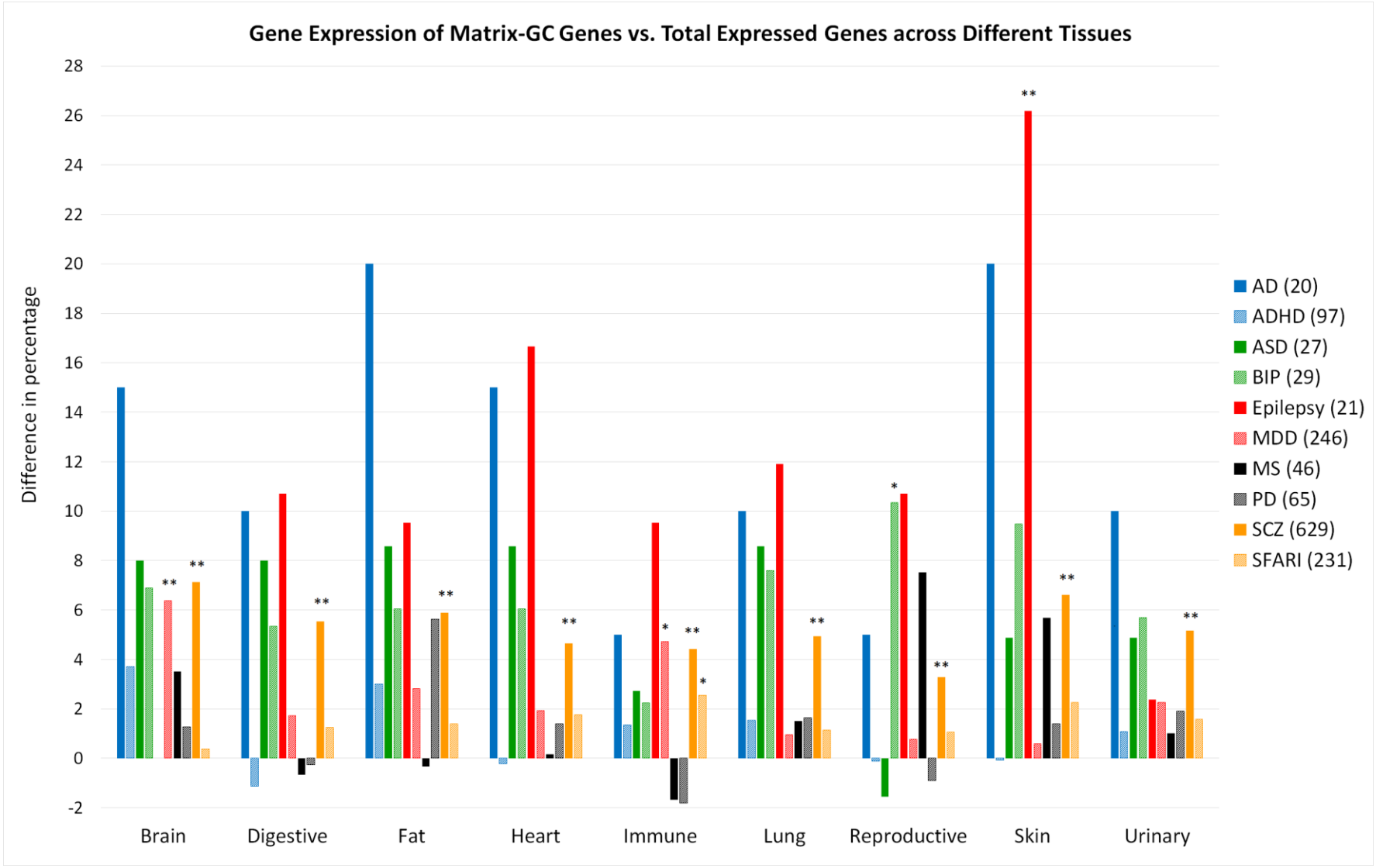
Gene Expression of Matrix-GC Genes vs. Total Expressed Genes across Different Tissues
Difference in percentage
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
-2
Brain
Digestive
Fat
Heart
Immune
Lung
Reproductive
Skin
Urinary
AD (20)
ADHD (97)
ASD (27)
BIP (29)
Epilepsy (21)
MDD (246)
MS (46)
PD (65)
SCZ (629)
SFARI (231)

Figure 3.10: Tissue expression distribution of brain disorder gene sets. RNA expression distribution, by tissue, for genes associated to brain disorders before (total) and after Matrix-GC (Matrix-GC genes). Disorders included are Autism database (ASD-SFARI, yellow), schizophrenia (SCZ, orange), Parkinson's disease (PD, grey), multiple sclerosis (MS, black), major depressive disorder (MDD, pink), epilepsy (red), autism (ASD, dark Green), Alzheimer's disease (AD, blue), attention deficit hyperactivity disorder (ADHD, light blue) and bipolar disorder (BIP, light green). Numbers in parentheses represent the total number of genes for which expression data was retrieved. The Fisher's exact test was used for statistical calculations.* represents p-value ≤ 0.05 and ** represents p-value ≤ 0.001.

3.3.3 RNAseq Data Validation

To validate candidate MeCP2 target genes identified by Matrix-GC, we hypothesise that expression of genes directly affected by MeCP2 should be altered in the mutant mice or RTT cells. We compared Matrix-GC genes to matched randomised control datasets and evaluate the expression of the candidate genes in blood and brain.

Initially, we consider the expression levels in *Mecp2*-mutant mice and matched controls using *Mecp2*$^{tm1.1Bird}$ data from cerebellum and blood (GSE129387). In particular, we look at the percentage of candidate genes whose expression was altered by at least half or double (fold change of ≤0.5 or ≥2). We investigate whether SNPs are present in the 1000bp promoter of all Matrix-GC genes. Of the total 1018 genes bound by our procedure, 380 genes are from cerebellum and 446 genes are from blood when cross-referenced with GTEx portal single-tissue eQTL data. Furthermore, out of 129 genes that are both recognised by Matrix-GC and have SNPS present in the promoter, 46 genes are confirmed by significant eQTL data to be found in blood and cerebellum tissue.

For genes expressed in the brain, the percentage of those with a fold change in the range considered is 78% and for genes with a SNP in the promoter it is 88%. The percentage of differentially expressed genes is 68% for Matrix-GC genes and increases to 83% for genes with SNPs in their promoters (Figure 3.11). Using permutation analysis we generate 1000 random control datasets of the same size of the total candidate list, and we calculate the distribution of the percentage of genes whose expression was altered by at least 2-fold. By looking at this distribution, we confirm that the expression change of our MeCP2-candidate target genes was in the 1% of the distribution and therefore p-value <0.001.

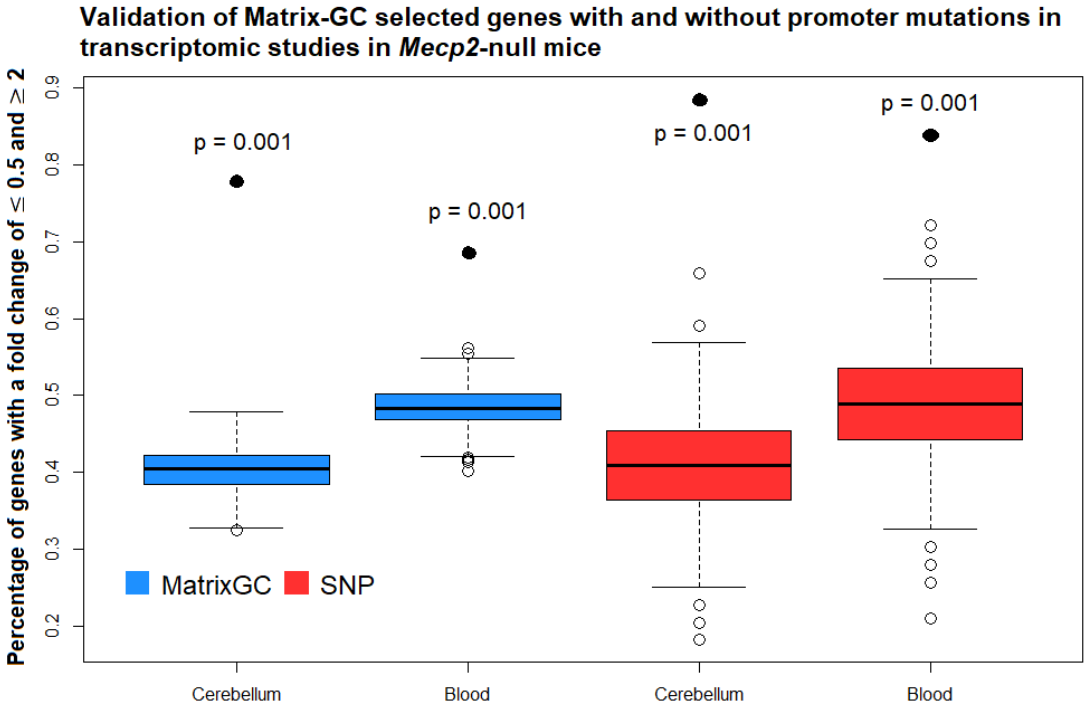

Figure 3.11: Validation of Matrix-GC genes in transcriptomic studies in *Mecp2*-null mice looking at fold change percentages. Boxplots of the percentage of genes with a 2-fold change comparing eQTL cross-referenced Matrix-GC genes (blue) and Matrix-GC genes with SNP mutations within the promoter sequences of 1000bp (red). Experimental data is from blood and cerebellum tissues in *Mecp2*$^{tm1.1Bird}$ mice. Monte Carlo permutation analysis provides the percentage of genes with a 2-fold change of (fold change of ≤0.5 or ≥2) on 1000 randomly generated and sized-matched control gene sets. The line in the middle of the boxplot represents the median of the distribution. The top and bottom borders of the boxplot represent the 3rd quartile (Q3) and 1st quartile (Q1) respectively. The upper and bottom whiskers represent Q3 + 1.5 times interquartile range, and Q1 – 1.5 times the interquartile range, respectively. Unfilled circles represent the outlier observations in each group, while filled circles represent the percentage of Matrix-GC filtered genes in each group. P-values are reported above groups where Matrix-GC filtered genes are significant compared to control data.

Following this, we then decide to evaluate and compare gene expression levels between two biological models: a mouse mutant for the for *Mecp2* gene (*Mecp2*$^{tm1.1Bird}$ data from Cerebellum and Blood (GSE129387), and cells derived from a patient with RTT (GSE123753). Both models carry similar mutations along the same region of the human *MECP2* and mouse *Mecp2* genes. In particular, we use the percentage of significant DEGs (p-value ≤ 0.05) as opposed to the fold change. DEGs are genes that are significantly different in expression between compared groups: genes in *MECP2/Mecp2* mutant cells vs. genes in wildtype or control cells. As before, we consider all the Matrix-GC genes across all datasets; of the total 1018 genes captured by Matrix-GC, 380 genes are from the cerebellum, 446 genes are from blood and 301 from the cortex when cross-referenced with GTEx portal single-tissue eQTL data.

The percentage of significant genes expressed in the mouse brain and blood was 0.049% and 0.017% respectively. Monte Carlo analysis did not show any significant results in the blood, but rather in the cerebellum (p-value <0.001) confirming the role of MeCP2 is more influential in the brain than in the blood. The percentage of genes expressed with a p-value ≤0.05 in human-derived Rett neural progenitor cells and neurons is 14.69% and 13.54% respectively. The statistical analysis reveals significant results both in neural progenitor cells and in differentiated neurons (p-value <0.005; Figure 3.12).

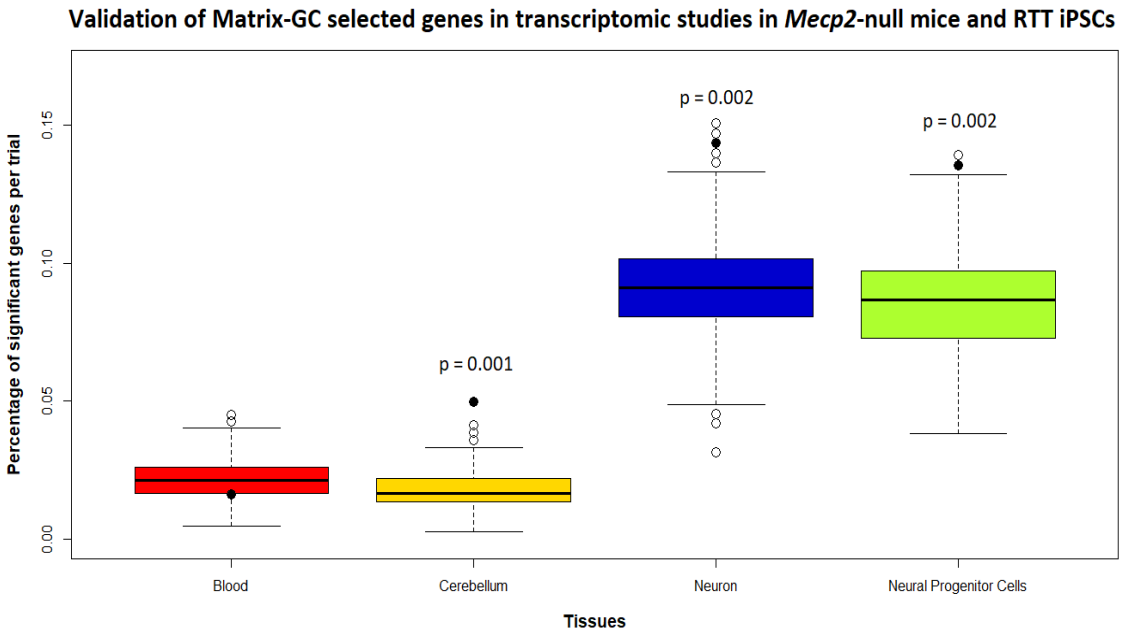

Figure 3.12: Validation of Matrix-GC selected genes in transcriptomic studies in *Mecp2*-null mice and RTT iPSCs. Boxplots of the percentage of significant genes in different experimental groups: blood in *Mecp2*$^{tm1.1Bird}$ mice (red), cerebellum in *Mecp2*$^{tm1.1Bird}$ mice (yellow), iPSCs-derived neurons from RTT patient (blue, *MECP2* deletion in exon region 3-4), iPSCs-derived neural progenitor cells from RTT patient (green, *MECP2* deletion in exon region 3-4). Monte Carlo permutation analysis provides the percentage of significant genes (p-value ≤ 0.05) on 1000 randomly generated and sized-matched control gene sets. The line in the middle of the boxplot represents the median of the distribution. The top and bottom borders of the boxplot represent the 3rd quartile (Q3) and 1st quartile (Q1) respectively. The upper and bottom whiskers represent Q3 + 1.5 times interquartile range, and Q1 – 1.5 times the interquartile range, respectively. Unfilled circles represent the outlier observations in each group, while filled circles represent the percentage of Matrix-GC filtered genes in each group. P-values are reported above groups where Matrix-GC filtered genes are significant compared to control data.

Chapter 4
Discussion

4.1 Discussion

In this study, we propose Matrix-GC, a DNA binding model that captures potential MeCP2 binding sites with moderate sensitivity. By applying this model to brain disorder datasets and utilising existing genetic data with stringent statistical analysis, we produce new results that is relevant in better understanding disease physiology.

We develop this model to identify MeCP2 binding sites using a PWM which is composed of sequence motifs that can be used to predict binding sites in gene promoter sequences. PWMs are limited in terms of discriminating true binding sites from non-binding sites as each nucleotide probability at its nucleotide position is calculated independently from its neighbour. This poses a challenge in transcription binding site analysis, and it has been proposed to expand the traditional PWM-only model with additional information to improve discrimination (Yang and Ramsey, 2015). We approach the MeCP2 PWM model with this in mind and consider the various ways to maintain, if not increase the sensitivity rate while lowering the false positive rate. Rube and colleagues show that while MeCP2 binds preferentially to methylated regions, it can bind to non-methylated DNA [57]. They report that GC% in an important sequence predictor of MeCP2 binding where there are no CpG dinucleotide repeats. By incorporating DNA sequence features that promote MeCP2 binding, Matrix-GC outperforms the PWM-only model. The additional filter of GC content reduces the amount of false positive motifs that might be bound by MeCP2 due to random chance.

To optimise our binding procedure, we consider whether the affinity of MeCP2 to bind to certain genes in definitive MeCP2 ChIP-Seq experiments has an effect on the performance. We report no discernible different between AUCs for 100, 200 and 300 genes which have different BETA scores. What did make a difference was varying sequence GC content. As expected, the PWM alone performed better than the random classifier but was outperformed when 50% and 60% GC content

was also considered. 60% GC content offers a reduction in false positive rate by nearly half compared to 50% GC content. This is in line with what Rube et al reported in their study, where MeCP2 is enriched at DNA regions with GC% of ≥60%.

We then apply Matrix-GC to gene sets derived from several genetic studies in brain disorders and show that MeCP2 binds to these gene sets at a higher percentage than it would to the genome on average. Results suggest higher involvement of MeCP2 in neuropsychiatric pathologies, although this can be attributed to the lower number of genes present in neurological datasets. However, when we investigate synaptic genes, we report a higher percentage of genes bound *in silico* by MeCP2. This strengthens the validity of our model and confirm the bias of MeCP2 towards brain-related genes, since MeCP2 regulates neuron function and plasticity [72–74].

The next question is whether genes captured by our binding protocol are biologically relevant and to this aim, we look at downstream pathways; are the genes bound by MeCP2 involved in pathological mechanisms across different brain disorders? In order to identify such mechanisms, we use the candidate MeCP2 target genes to carry out enrichment and network analysis.

Considering its role in gene regulation, and its expression in the brain and whole organism [25], it is expected that MeCP2 plays a role in several mechanisms linked to brain disorders. Furthermore, several correlations between *MECP2* expression levels and brain disorders features have been proposed. This might occur directly through *MECP2* mutations [75–77] or indirectly via MeCP2 regulating co-factors such as BDNF [34,35], and non-coding RNA action [78–80].

Our analysis reveals three main mechanisms potentially mediated by MeCP2 across brain disorders: neuronal transmission, immune-related pathways and processes for growth and development. Enrichment of dopaminergic and glutamatergic related terms and pathways in ADHD, ASD-SFARI, and PD gene sets is reinforced by dopamine and glutamate receptors hub proteins in the

protein-protein interaction network analysis. Dopaminergic dysregulation in RTT patients has been observed through reductions in dopamine itself or its metabolite, homovanillic acid [81,82]. This dysregulation leads to dyskinesia, hand stereotypies and rigidity: symptoms found also in RTT. Alterations in dopamine transmission is a feature of several neurological disorders, notably PD, but is also present in AD, ADHD, SCZ, MS and HTT [83–86].

Furthermore, increased glutamate levels [87] are observed in patients with RTT and animal models [87,88] and, glutamatergic synapses are regulated by MeCP2 [89]. NMDA glutamate receptor-related Reactome pathways are enriched in ADHD and PD sets. These results suggest MeCP2's influence in dopaminergic and glutamatergic systems has functional and behavioural consequences in several brain disorders.

Altered immunity has been reported in neuropsychiatric disorders such as SCZ, depression and ASD [90,91]. We observe one immune GO term (GO:0002292, T-cell differentiation involved in immune response) in the ASD-SFARI dataset and this stands to reason since MeCP2 is reported to alter T-lymphocyte gene expression profile [92]. AD, ALS and MS gene sets are enriched for interleukin and Toll-like receptor signalling pathways. Inflammation is already implicated in AD and its progression [93] and is also critical in MS pathology. Similar to MS, RTT displays features that are hallmarks of autoimmune disorders suggesting potential common therapeutics [94–96]. We show MeCP2 binds to immune-related genes, and in particular to S100A9, a gene already identified in transcriptomic studies on blood and brain of *Mecp2*-null mice [70]. The levels of S100A8 and S100A9 proteins are related to inflammation, and are elevated in MS.

We also determine that on average, MecP2 binds to 41.8% of immune function genes through Matrix-GC. This is a marginal increase from the genomic average of 39.56%. It is possible that MeCP2 influences immune pathways indirectly through co-factors such as CAMP Responsive Element Binding Protein 1 (CREB1) and NCoR/SMRT co-repressor complex. CREB1 influences immune pathways and favours an anti-inflammatory profile [97]. When MeCP2 is

mutated, it is unable to recruit CREB1 for necessary gene activation and can lead to an inflammatory response. NCoR/SMRT dampens down the immune response by repressing the action of liver X receptors [98]. Mutations within the interaction domain on the *MECP2* gene can prevent co-factor association and alter immune function. These two MeCP2 co-factors are examples of how MeCP2 might play a bigger role at the crossroads of immune function.

Growth and developmental processes are enriched and relate to cell cycle and proliferation. This is not surprising, given MeCP2's reputation as an epigenetic modifier in cancer [99]. Mutations in *MEF2C* affect MeCP2 function and this has been observed in epilepsy and ADHD studies [100]. Taken together, we posit that MeCP2 exerts influence in early development in SCZ and ASD. Similarly, the ADHD dataset is enriched for pathways related to neurotrophic factor signalling which mediates neuronal proliferation and maturation [101]. EP300: a transcriptional regulator is influenced indirectly by MeCP2 likely through MEF2C [71]. EP300 is a hub protein and interestingly, two antisense RNA genes *EP300-AS1* and *MEF2C-AS1* in SCZ are bound by MeCP2, while *MEF2C-AS1* has the highest expression in the brain cerebellum.

Overall, our enrichment results reveal connections between MeCP2 and different brain pathologies. MeCP2 contributes to several brain disorders directly by binding to genes that are disrupted in neuropathology. We also propose that MeCP2 can indirectly influence brain disorders by binding genes mediating other functions, including inflammation. Could it be that MeCP2, and its network of co-factors exerts controls over common homeostatic processes across several brain disorders? If so, this suggests common therapeutic targets across brain disorders.

Our next aim was to functionally validate our results at the gene level. We look at SNP mutations in MeCP2 in tandem with enrichment analysis to understand which brain disorder MeCP2 function may be altered. The presence of *MECP2* SNPs in some patients with epilepsy, is not surprising, given the presentation of epilepsy in 75% of the cases of RTT. It needs to be kept in mind however, that the effect of *MECP2* mutations in epilepsy are still not well understood [102]. The

correlation between *MECP2* mutations and epileptic phenotype in RTT has proved a challenge to describe, due to the complex nature and presentation of the disorder [103]. MeCP2 SNPs present in ASD and SCZ indicate an epigenetic link between the two disorders that is newly emerging [104].

The possible involvement of MeCP2 in ADHD has not been properly established, despite the known relationship between ADHD and ASD (and by extension, forms a tentative link to MeCP2). Here we find *MECP2* SNPs in ADHD patients, which suggests MeCP2 may be involved in pathology. This hypothesis is reinforced by immunofluorescence studies where *MECP2* is reduced in ADHD cerebral cortices [105]. A more recent study investigating predictive epigenetic markers in ADHD diagnoses reports a correlation between predictability and decreased *MECP2* mRNA levels [106].

Next, tissue expression analysis shows genes modulated by MeCP2 are also expressed outside the nervous system. For example, in the immune system, AD, ALS and MS gene sets are enriched for interleukin and Toll-like receptor signalling pathways. Subsequent gene expression analysis of AD genes shows an increase in immune tissues after Matrix-GC. While the overall tissue distribution remains steady in disorders before and after Matrix-GC, there is an increased expression of brain disorders-associated genes in brain, immune, and digestive tissue for genes associated to ALS, AD, and ADHD gene sets respectively (Figure 3.10).

Transcriptomic studies are a useful tool to validate our findings from Matrix-GC and enhance biological interpretations of the results. By looking at the complete set of gene transcripts generated within a tissue, we can take a snapshot of what genes are being utilised.

We initially evaluate a 2-fold change in genes that were bound by Matrix-GC, and propose an additional subgroup of Matrix-GC genes that also had SNP mutations within their 1000bp promoters. In both cerebellum and blood tissues, the effects were specific for the candidate genes selected though our procedure. However, we are limited in this analysis because we use a mouse transcriptomic study.

MeCP2 is regulator of neuron development and while the comparison between cerebellum and blood is intriguing, this may not be captured in human pathophysiology. Furthermore, investigating genes with SNPs in the promotor region introduces confounding factors since SNPs alone are not necessarily indicative of disease.

Therefore, by looking at statistically significant expression in genetic studies using both *Mecp2* mutant mice and iPSCs derived from a RTT patient, we provide a simple but clearer analysis of Matrix-GC filtered genes and whether they are biologically relevant. Indeed, in this secondary analysis we report that genes bound by Matrix-GC correlate with genes that are altered by MeCP2 knockout for mice cerebellum, and human neural progenitor cells and differentiated neuron. The discriminatory ability of Matrix-GC to not only bind to brain-associated genes but also to capture genes that are functionally altered under statistical rigour, proves it to be a useful MeCP2 binding model. We put forward Matrix-GC as a novel procedure to identify target genes for future studies, which can be further refined with additional predictive features.

We are able to generate a binding model that is specific to MeCP2 in this study, however it is important to mention that these methods can be applied to different binding proteins. There are databases for transcription factors that bind to DNA such as JASPAR or TRANSFAC [107,108] which are useful in this regard. It may be worthwhile to study other factors that are involved in brain growth and function e.g. CREB and FOS to complement this MeCP2 study. One of the limitations of using a PWM alone is that it can capture multiple predicted binding sites occurring by chance and does not represent a true biological site. Here, we have tried to mitigate this by including GC content into our PWM model according to data from Rube and colleagues [57]. Future studies using a similar model should explore other epigenetic features that promote true DNA binding for example DNA hypersensitivity and methylated regions. DNAse hypersensitivity sites are regions of DNA prone to being cleaved by DNAse I nuclease enzyme. As a result, the DNA structure becomes more accessible and gives way to regions of active genes. DNAse hypersensitivity can be assayed and mapped onto the genome [109] and therefore help us understand which genes a

binding factor is likely to have access to. It would also be interesting to see DNA methylation data being incorporated into binding algorithms. DNA can be tagged with methyl function groups that promote gene repression. The methylome can be mapped through bisulfate-sequencing to determine which methylated sequences MeCP2 and other proteins preferentially binds to. In particular, methylation data would be a lucrative way to assess true binding because MeCP2 recruits histone deacetylase at these sites to repress gene transcription [110]

Moreover, this does not apply solely to transcription factors and gene sequences. It would be possible to model this with protein-protein interactions whereby either a protein complex is formed or co-factor association. Protein binding motifs are similar to DNA motifs except they contain amino acids instead of nucleotides [111,112]. They can have redundancy along the protein sequence similar to MeCP2 (Figure 2.1). A protein binding model also necessitates sequence features to accurate predict true binding sites. However, instead of epigenetic factors, one may look instead at conformation or folded protein shapes that are energetically favourable.

Our methods can also be applied to general bioinformatic analyses. We provide the framework to carry out downstream gene set analysis with statistical rigour. This is particularly important to counteract the annotation bias present in these databases [113,114]. Annotation bias occurs due to over-representation of certain genes or gene products in the literature which is then implemented in the curation of bioinformatic resources. In downstream analysis, this leads to false positive results that limits interpretation. In fact, it was shown recently that the inequality of annotations in various databases is increasing over time [114]. We are limited by the information that is already available in literature and one way to correct potential noise is to have a stringent control analysis. Here, we observe that few statistically significant terms or pathways across the 1000 trials which do not affect our main findings and we are confident in our enrichment results coming from Gene Ontology and Reactome PA.

Moreover, redundancy is still a large issue in gene set analysis, from which neither the Gene Ontology nor Reactome database is spared [115–117]. This

occurs through multiple terms or pathways containing the same genes or similar sets of genes which then confounds results. Redundancy is also present through the hierarchal organisation of databases with entries being nested into parent terms and so on, which leads to the question: at which level should we be considering these results as valid without conflation and overlaps? While we do not directly address this problem in this study, we propose permutation analysis as we have carried out here, to identify where annotation biases can occur and to reduce false positives while using these tools. Recent studies have reported clustering and set cover methodologies to counteract this redundancy issue [117,118], which would be a worthwhile pursuit in our future research.

4.2 Limitations

We use ChIP-Seq data from cell lines to establish the validity of Matrix-GC. By definition, cell lines are not normal and do not have the same molecular profile as non-immortalised cells. Ideally, we would use ChIP-Seq data in primary cells from different cell populations or that are patient-derived. This is similar to the limitation relating to data from tissue expression analysis which is the poor availability of data and the unknown developmental stage at the time of tissue collection [119]. However, we do report the same results when using data from two independent cell lines, IMR-90 foetal lung fibroblasts and HCT-116 colon carcinoma cells [55].

We also need to consider the different number of genes present in the datasets, and the limited number of genes in several disorders (anorexia, epilepsy, HTT) as these datasets have lower statistical power when it comes to enrichment and network analysis. Hence the network and enrichment component of the study has a natural bias toward the disorders with more genes; this is observed with the absence of mechanisms in ASD but an abundance of mechanisms in the ASD-SFARI set, where many terms and pathways are significant.

It is difficult to carry out large scale control analysis within the Cytoscape interface including network and clustering analysis, compared to Gene Ontology and Reactome, where R packages are available. Other limitations arise from control methods implemented in this study. There is discrepancy between R packages and online interfaces. For example, the ClusterProfiler package might only be able to be updated less frequently than Gene Ontology and Reactome databases. This can lead to some genes and proteins not being represented or annotated and offsets comparative analysis. There is a trade-off between feasibility in analysing multiple large datasets using programming language, and time-consuming “manual” single set analysis into online interfaces. However, this limitation does not outweigh the benefits of repeated permutation analysis which is a valid method of controlling for false positive results and highlighting which interfaces are reliable such as Gene Ontology and Reactome.

The disadvantages of using transcriptomics studies solely in verifying outcomes of our study are tied in with the limitations of transcriptomics studies themselves. In measuring the abundance of mRNA through sequencing, this uses mRNA as a proxy for biological activity [120]. Furthermore, bulk transcriptomic studies are carried out on heterogenous tissue populations. Newer technologies allow for single-cell sequencing to be carried out however this methodology is costly and yet to be widely adopted as the standard.

Conclusion

Results from this thesis contribute to the understanding of MeCP2 and its influence in brain disorders through bioinformatic methods. By combining different genetic datasets, we are able to produce a relevant DNA binding model that feeds into downstream enrichment analysis. This strategy is not limited to MeCP2 but indeed, any protein that may influence transcription in brain disorders. Additional functional validation and rigorous statistical controls reinforces the validity of our results. Putative target genes and molecular pathways are implicated in the pathophysiology of several brain disorders which should be further investigated, namely neuroinflammation. Lastly, we provide statistical solutions to limit false positives that can be applied to gene set analysis. Future research based on these results should aim at refining model sensitivity and take advantage of the increased output of genetic studies.

Appendix

Appendix A: R Script for Matrix-GC Analysis

```
library(Biostrings)
library (mygene)
library (biomaRt)
library(regioneR)
library(BSgenome.Hsapiens.UCSC.hg19)
library(GenomicFeatures)
library(zoo)

## Create human TxDB annotation object
TxDB.HSapiens.UCSC.hg19<-makeTxDbFromUCSC( genome="hg19",
      tablename="ensGene",
      transcript_ids=NULL,circ_seqs=DEFAULT_CIRC_SEQS,
      url="http://genome.ucsc.edu/cgi-bin/",
      goldenPath_url="http://hgdownload.cse.ucsc.edu/goldenPath"
      ,taxonomyId=NA,
      miRBaseBuild=NA)

## Subset of chromosomes: exclude mitochondrial chromosomes
seqlevels(TxDB.HSapiens.UCSC.hg19)<-
      c(seqlevels(TxDB.HSapiens.UCSC.hg19)[1:24])

## Retrieve gene IDs across different systm
genes_id<- queryMany(disorder_genelist_1,
                     scopes=c("symbol"),
                     fields=("ensembl.gene"),
                     species="human")
genes_id$ensembl

which(genes_id$notfound==T)

gene_id_list<-unlist(genes_id$ensembl.gene)
```

```
gene.transcript<-subset(genes(TxDB.HSapiens.UCSC.hg19,
      columns= c("gene_id","tx_name")),
      gene_id %in% gene_id_list)

genes.tss$symbol<-queryMany(gene.transcript$gene_id,
                            scopes=c("ensemblgene"),
                            fields=c("symbol"),
                            species="human")[,4]

## Retrieve transcription start site
genes.tss = resize(gene.transcript,
                   width=1, fix='start')

## Retrieve promotor sequence: 1000 base pairs upstream of
## transcription start site
sequence<-getPromoterSeq(genes.tss, BSgenome.Hsapiens.UCSC.hg19,
                         upstream=1000,
                         downstream=0)

## Create DNA string from sequences
gene_seq_string<-lapply(sequence,DNAString)

x= 1:length(gene_seq_string)

## Convert position frequency matrix to position weight matrix
MECP2_PWM <- PWM(MECP2_PFM)

## Apply position weight matrix to sequence DNA strings
binding_function<-function(x, PWM, string) {
      matchPWM(PWM, string[[x]],
      with.score = TRUE,
      min.score= "65%")
}

gene_hits<- lapply(x, binding_function,
      MECP2_PWM,gene_seq_string)
```

```
## Extract scores from gene_hits object
score_function<- function(y, PWM, string) {
     out_mcols<- mcols(bindscore_fun(y, PWM, string))$score


return(out_mcols)
}

gene_scores<- lapply(x, score_function, MECP2_PWM,
     gene_seq_string)

## Create an empty matrix incl. chromosome, +/-strand, gene
## start, gene end coordinates GRange object
gene_list<- as.data.frame(matrix(data=0, nrow =
     length(gene.transcript)))

gene_list$V1= as.data.frame(gene.transcript)$chromosome

gene_list$V2= as.data.frame(gene.transcript)$strand

gene_list$V3= as.data.frame(gene.transcript)$start

gene_list$V4= as.data.frame(gene.transcript)$end

gene_list$V5= as.data.frame(gene.transcript)$Ensembl_IDs

gene_list$V6= queryMany(gene_list$V5,
                        scopes=c("ensemblgene"),
                        fields=c("symbol"),
                        species="human")[,4]



forward_genes <- which(gene_list[,2] == "+")

reverse_genes <- which(gene_list[,2] == "-")
```

```
## Functions to retrieve start and end coordinates across
## different binding sites

match_start_function<- function(a, matches) {
  start<- Biostrings::start(matches[[a]])
  return(start)
}
start_sites<- lapply(1:length(gene_hits), match_start_function,
      gene_hits)

match_end_function <- function(a, matches) {
  end<- Biostrings::end(matches[[a]])
  return(end)
}

end_sites <-lapply(1:length(gene_hits), match_end_function,
      gene_hits)

## Function for forward coordinates and downstream analysis
forward_coord_function<- function(gene_num,match, gene_list) {
  match_chr_F <- match[[gene_num]] -1000 + gene_list[gene_num,3]
  return(match_chr_F)
}

forward_start<- lapply(forward_genes , forward_coord_function,
      start_sites, gene_list)

forward_end <- lapply(forward_genes , forward_coord_function,
      end_sites, gene_list)

## For loop initial to create matrix: Forward strand genes
p_count = 0
p_count
for(i in 1:length(forward_start)){
  chrom2 <- gene_list[forward_genes [i],1]
  starts <- forward_start [[i]]
  ends <- forward_end [[i]]
```

```
  score<- gene_scores[forward_genes][[i]]
  ens<- gene_list[forward_genes [i],5]
  genesym<-gene_list[forward_genes [i],6]
  if (length(starts>0)){
    for(a in 1:length(starts)){
      p_count = p_count +1
    }
  }else {
    p_count = p_count +1
  }
}
p_count

full_forward_data<- as.data.frame(matrix(data=0, nrow =p_count))

full_forward_data $start = 0
full_forward_data $end = 0
full_forward_data $count = 0
full_forward_data $score=0
full_forward_data $ens=0
full_forward_data $gene=0

p_count = 0
p_count
```

```
## For loop: Forward
for(i in 1:length(forward_start)){
  print(i)
  print(p_count)

  chrom2 <- gene_list[forward_genes [i],1]
  starts <- forward_start [[i]]
  ends <- forward_end [[i]]
  score<- gene_scores[forward_genes][[i]]
  ens<- gene_list[forward_genes [i],5]
  genesym<-gene_list[forward_genes [i],6]

  if (length(starts>0)){


    for(a in 1:length(starts)){
      p_count = p_count +1
      full_forward_data $count[p_count] = i
      full_forward_data l$chr[p_count] = chrom2
      full_forward_data $start[p_count] = starts[a]
      full_forward_data $end[p_count] = ends[a]
      full_forward_data $score[p_count] = score[a]
      full_forward_data $ens[p_count]= ens[a]
      full_forward_data $gene[p_count]= genesym[a]
    }
  }else {
    #    if(length(starts)){
    print("empty!!")

    p_count = p_count +1
    full_forward_data $count[p_count] = i
    full_forward_data $chr[p_count] = 0
    full_forward_data $start[p_count] = 0
    full_forward_data $end[p_count] = 0
    full_forward_data $score[p_count]=0
    full_forward_data $ens[p_count]=0
    full_forward_data $gene[p_count]=0
```

```
  }
}

## Rename sex chromosomes
full_forward_data$chr[f_coord$chr == 23]<-"X"

full_forward_data$chr[f_coord$chr == 24]<-"Y"

## Create GRange object of sequences to calculate guanine-
## cytosine content % incl. 100bp flanks
full_forward_data <-
      full_forward_data[!(full_forward_data$chr=="0"),])

full_forward_data $chr<-paste("chr", full_forward_data$chr,
      sep = "",
      collapse = NULL)

GRange_forward_matches <- GRanges(
      seqnames = full_forward_data $chr,
      strand = "+",
      ranges = IRanges(start =  full_forward_data$start,
      width = 215)) # gives start and end

## Calculate GC%
GC_function <- function (coord_num, coordinates, genome) {
  seq<- getSeq(genome, coordinates[coord_num])
  GC_F_freq <- sum(letterFrequency(seq[1],
                                   c("G","C"),
                                   as.prob=T))
  return(GC_F_freq)
}
GC_forward<-lapply(1:length(GRange_forward_matches),
      GC_function,
      GRange_forward_matches,
      BSgenome.Hsapiens.UCSC.hg19)
```

```
full_forward_data[,2]<- full_forward_data[,2]+100

full_forward_data$GC<-unlist(GC_forward)

full_forward_data$ens<-na.locf(full_forward_data$ens)

full_forward_data$gene<-na.locf(full_forward_data$gene)

##  Repeat reverse stranded genes
reverse_coord_function <- function(gene_num,match, gene_list) {
  match_chr_R <- match[[gene_num]] + gene_list[gene_num,4]
  return(match_chr_R)
}

reverse_start <-lapply(reverse_genes, reverse_coord_function,
      match.start, gene_list)

reverse_end <-lapply(reverse_genes, reverse_coord_function,
      match.end, gene_list)
```

```
## For loop

q_count = 0

q_count

for(i in 1:length(reverse_start)){
      chrs <- gene_list[reverse_genes[i],1]
      starts_R <- reverse_start [[i]]
      ends_R <- reverse_end[[i]]
      score_R<- gene_scores[reverse_genes][[i]]
      ens_R<- gene_list[reverse_genes[i],5]
      genesym_R<-gene_list[reverse_genes[i],6]

  if (length(starts_R>0)){
    for(a in 1:length(starts_R)){
      q_count = q_count +1
    }
  }else {
    q_count = q_count +1
  }
}
q_count

full_reverse_data = as.data.frame(matrix(data=0,
      nrow = q_count))

full_reverse_data $start = 0
full_reverse_data $end = 0
full_reverse_data $count = 0
full_reverse_data $score=0
full_reverse_data $ens=0
full_reverse_data $gene=0
```

```
q_count = 0
q_count

## Reverse stranded genes loop
for(i in 1:length(reverse_start)){
  print(i)
  print(q_count)

  chrs <- gene_list[reverse_genes[i],1]
  starts_R <- reverse_start[[i]]
  ends_R <- reverse_end[[i]]
  score_R<- gene_scores[reverse_genes][[i]]
  ens_R<- gene_list[reverse_genes[i],5]
  genesym_R<-gene_list[reverse_genes[i],6]

  if (length(starts_R>0)){

    for(a in 1:length(starts_R)){
      q_count = q_count +1
      full_reverse_data$count[q_count] = i
      full_reverse_data$chr[q_count] = chrs
      full_reverse_data$start[q_count] = starts_R[a]
      full_reverse_data$end[q_count] = ends_R[a]
      full_reverse_data$score[q_count] = score_R[a]
      full_reverse_data$ens[q_count]= ens_R[a]
      full_reverse_data$gene[q_count]= genesym_R[a]
    }
  }else {
    #    if(length(starts)){
    print("empty!!")

    q_count = q_count +1
    full_reverse_data$count[q_count] = i
    full_reverse_data$chr[q_count] = 0
    full_reverse_data$start[q_count] = 0
    full_reverse_data$end[q_count] = 0
    full_reverse_data$score[q_count]=0
```

```
    full_reverse_data$ens[q_count]=0
    full_reverse_data$gene[q_count]=0
  }
}

full_reverse_data$chr[full_reverse_data$chr==23]<-"X"

full_reverse_data$chr[full_reverse_data$chr==24]<-"Y"

full_reverse_data<-
     (full_reverse_data[!(full_reverse_data$chr=="0"),])

full_reverse_data$chr<-paste("chr", full_reverse_data$chr,
     sep = "", collapse = NULL)

## Create GRange Object
GRange_reverse_matches <-
     GRanges(seqnames = full_reverse_data$chr,
     strand = "-",
     ranges = IRanges(start =  full_reverse_data$start,
     width = 215))

## Calculate GC%
GC_reverse<-lapply(1:length(GRange_reverse_matches),
     GC_function,
     GRange_reverse_matches,
     BSgenome.Hsapiens.UCSC.hg19)

full_reverse_data [,2]<- full_reverse_data [,2]+100

full_reverse_data$GC<-unlist(GC_reverse)

full_reverse_data$ens<-na.locf(full_reverse_data$ens)

full_reverse_data$gene<-na.locf(full_reverse_data$gene)
```

```
### Downstream analysis
## positive control or “normal” results
result<- as.data.frame(matrix(data=0, nrow = 19, ncol = 5))
result$V1<-seq(from=5,
               to=95,
               by=5)

colnames(result)<- c("PWM", "GC", "Genes", "FNR", "TPR")

total_data<-rbind(full_forward_data, full_reverse_data)

## Restrict to score =0.65
total_65<- total_data[(which(total_data $score >= 0.65)),]

## Matrix-GC: Score greater than or equal to 0.65
## GC% greater than or equal to 0.60
MatGC<-total_65[(which(total_65$GC >= 0.60)),]

unique(MatGC$gene)

##   Vary GC Gontent
total_GC60<-total[(which( total_data$GC >= 0.6)),]

total_GC50<-total[(which( total_data$GC >= 0.5)),]

sep<-seq(from=0.05, to=0.95,by=0.05)
```

```
## Count function for GC matches by score stratification
GC_count_function<- function (sep, total_list){
      gc_length<- length(which(total_list$score>=sep))
      return(gc_length)
}

# Count function for unique genes by score stratification
unique_genes_fun<- function (sep, total_list){
      unique_gene<-
      length(unique(total_list$gene[total_list$score>= sep]))
      return(unique_gene)
}

## 60% GC content filter
total_GC <- unlist(lapply(sep, GC_count_function,
      total_list=total_GC60))

total_gene<-unlist(lapply(sep, unique_genes_fun,
      total_list=total_GC60))

result$GC<- total_GC

result$Genes<- total_gene

## Calculate false negative rate and true positive rate
## for 100, 200, and 300 genes
result$FNR<- ((300-result$Genes)/300)

result$TPR<- (result$Genes)/300
## Repeat above  for 50% GC content and no GC content filter

## Repeat for negative control datasets
negative_con_result<- as.data.frame(matrix(data=0, nrow = 19,
      ncol = 5))
```

```
neg_result$V1<-seq(from=5,
                   to=95,
                   by=5)

colnames(neg_result)<- c("PWM", "GC", "Genes","FPR", "TNR" )

negative_total_GC60<-total[(which(negative_control$GC >= 0.6)),]

negative_total_GC50<-total[(which(negative_control$GC >= 0.5)),]

# 60% fiter
negative_con_GC <- unlist(lapply(sep, GC_count_function,
     total_list= negative_total_GC60))

negative_con_genes<-unlist(lapply(sep, unique_genes_fun,
     total_list= negative_total_GC60))

neg_result$GC<- negative_con_GC

neg_result$Genes<- negative_con_genes

## Calculate false positive rate and true negative rate
## for 100, 200, and 300 genes
neg_result$FPR<- (neg_result$Genes)/300

neg_result$TNR<- ((300-neg_result$Genes)/300)
## Repeat above  for 50% GC content and no GC content filter
```

Appendix B: R Script for Enrichment and Control Analysis

```
library(mygene)
library(clusterProfiler)
library(GenomicFeatures)
library(tidyverse)
library(ReactomePA)
library(reactome.db)
library(GO.db)
library(org.Hs.eg.db)

## Create human TxDB annotation object
TxDB.HSapiens.UCSC.hg19<-makeTxDbFromUCSC( genome="hg19",
      tablename="ensGene",
      transcript_ids=NULL,circ_seqs=DEFAULT_CIRC_SEQS,
      url="http://genome.ucsc.edu/cgi-bin/",
      goldenPath_url="http://hgdownload.cse.ucsc.edu/goldenPath"
      ,taxonomyId=NA,
      miRBaseBuild=NA)

## Subset of chromosomes: exclude mitochondrial chromosomes
seqlevels(TxDB.HSapiens.UCSC.hg19)<-
      c(seqlevels(TxDB.HSapiens.UCSC.hg19)[1:24])

## Retrieve gene IDs across different systm
genes_id<- queryMany(disorder_genelist_1,
                     scopes=c("symbol"),
                     fields=("ensembl.gene"),
                     species="human")

## Remove rows with an NA
genomeid_NA<-
      genes_id[,1:4][complete.cases(genes_id[,1:4]),]
```

```
## Random sample rows of dataframe for genomic control
## e.g. 1000 samples of 100 genes
genomecon_sample<-
      replicate(1000,       genomeid_NA[sample(nrow(genomeid_NA),
      100,replace = F),])

## Random sampling for exome genes by identifying
## protein-coding genes
protein_id<- queryMany(genome_id$gene_id,
                       scopes=c("ensembl.gene"),
                       fields=(c("ensembl.protein",
                         "symbol")),species="human")

exome_id<-protein_id[-which(protein_id$notfound==T),]

exome_id<-exome_id[-which(exome_id$ensembl=="NULL"),]

exomecon_sample<-
      replicate(1000,exome_id[sample(nrow(exome_id),
      100, replace = F),])

## Pathway Enrichment
pathway_function<- function (x) {
  enrichPathway(x$`_id`,
                organism="human",
                pvalueCutoff=1,
                pAdjustMethod = "fdr")
}

PW.df<-lapply(genomecon_sample, pathway_function)
```

```
# GO ENRICHMENT
GO_function <-function(y) {
  enrichGO(y$"_id",
           org.Hs.eg.db,
           ont = "BP",
           pvalueCutoff = 1,
           pAdjustMethod = "fdr" )

}

GO.df<-lapply(genomecon_sample, GO_function)
## Statistical Analysis with Fisher's Method
fishers_method_function<- function(x) {
     pchisq(-2 * sum(log(x)),df=2000,lower=F)
}

## Retrieving the IDs and pvalues from results .df
retrieve_function = function(n) {
  tab = data.frame(ID = genomecon_sample[[n]]$ID,
                   pvalue = genomecon_sample[[n]]$pvalue,
                   trial = n)
  tab
}
Locate_zero<- which(sapply(genomecon_sample,function (x)
     length(x$ID))==0)

genomecon_sample[loc.zero]<-NULL

results.df <-
     do.call(rbind,lapply(c(1:length(genomecon_sample)),
     retrieve_function))

pval.tbl <- pivot_wider(results.df, names_from = ID,
     values_from = pvalue)

pval.mat<-as.data.frame(pval.tbl)
```

```
## NA p-value matrix to 1000 trials/ make matrix
## rectangular
pval.mat[nrow(genomecon_sample):1000,]<-NA

set.seed(8)

pval.mat[is.na(pval.mat)]<-
      runif(length(pval.mat[is.na(pval.mat)]),
                           min=0.05, max=1)

fisher.pval.mat<-sapply(pval.mat,fishers_method_function)
```

Appendix C: R Script for Transcriptomic Data Validation Analysis

```
library(edgeR)
library(mygene)
library(dplyr)
library(GenomicFeatures)

## Re-analysis of RNA-seq data
sample_info <-factor(c(1,1,2,2))
DGE_list<-DGEList(gene_counts, group= sample_info)

## Differential expression analysis
DGE_list<-
      DGEList(DGE_list$counts[apply(DGE_list$counts,
      1,sum)!=0,], group = DGE_list$samples$group)

normCounts <- cpm(DGE_list)

pseudoNormCounts <- log2(normCounts + 1)

DGE_list <- estimateCommonDisp(DGE_list)
DGE_list <- estimateTagwiseDisp(DGE_list)

## Exact Test with No Filter
DGE_Test <- exactTest(DGE_list)

DGE_res<-topTags(DGE_Test, n=nrow(DGE_Test $table))

sum_FDR<-sum(DGE_res$table$FDR < 0.05)

## Add fold change information
DGE_res$table$FC<-2^(DGE_res$table$logFC)

sum_FC<-sum(DGE_res $table$FC <= 0.05)+
         sum(DGE_res $table$FC >= 2)
```

```
## Permutation Analysis
## Function to match eQTL data with RNAseq results
eQTL_function<-function(x,permut.data,rnaseq){
  indx<-match(permut.data[,x] , rownames(rnaseq))
  matched<-rnaseq[indx,]
  matched<-matched[-which(is.na(matched$PValue)),]
  pval.match<-matched[which( matched$PValue <= 0.05),]

}

eQTL_in_single_tissue_data

control_genes<-replicate(1000, sample(eQTL_in_single_tissue_data

     ,100, replace = F))

x<-1:1000
eQTL_res<-lapply(x, eQTL_function, control_genes, DGE_res)

y <-size of gene set

eQTL_percent<-sapply(eQTL_res, nrow)/y

mean(eQTL_percent)

percent_rank(eQTL_percent [order(eQTL_percent)])

z <- percentage of pvalue genes <= 0.05 from MatrixGC

percent_rank(c(eQTL_percent, z))
```

Appendix D: Reactome Pathway Enrichment Results

Table D.1: Reactome pathway enrichment results after Matrix-GC. For ReactomePA pathway enrichment, GeneRatio is the ratio of the number of genes from the dataset of interest related to a specific pathway to the total number of genes in the dataset of interest. BgRatio is the ratio of the total number of genes related to a specific pathway to the total number of genes [64]. Italicised grey terms represent results that are also significant after control permutation analysis.

Alzheimer's Disease Pathway Enrichment					
Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue
Interleukin-2 family signaling	2/8	44/10554	0.000468	0.014512	0.006899
Attention Deficit Hyperactivity Disorder Pathway Enrichment					
Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue
Amine ligand-binding receptors	7/46	42/10554	4.46E-10	8.12E-08	6.62E-08
GPCR ligand binding	*13/46*	*457/10554*	*4.36E-08*	*3.97E-06*	*3.24E-06*
Class A/1 (Rhodopsin-like receptors)	*10/46*	*324/10554*	*9.88E-07*	*5.99E-05*	*4.89E-05*
Neurotoxicity of clostridium toxins	3/46	10/10554	9.1E-06	0.000414	0.000338
Neuronal System	9/46	368/10554	2.43E-05	0.000883	0.00072
Unblocking of NMDA receptors, glutamate binding and activation	3/46	20/10554	8.39E-05	0.002181	0.001778

CREB phosphorylation through the activation of CaMKII	3/46	20/10554	8.39E-05	0.002181	0.001778
Ras activation upon Ca2+ influx through NMDA receptor	3/46	22/10554	0.000113	0.002562	0.002089
Signaling by NTRK2 (TRKB)	3/46	25/10554	0.000167	0.00337	0.002748
Uptake and actions of bacterial toxins	3/46	31/10554	0.00032	0.00582	0.004746
CREB phosphorylation through the activation of Ras	3/46	32/10554	0.000352	0.005821	0.004747
Transmission across Chemical Synapses	6/46	227/10554	0.000423	0.006412	0.005229
Protein-protein interactions at synapses	4/46	87/10554	0.00054	0.007558	0.006163
Class C/3 (Metabotropic glutamate/pheromone receptors)	3/46	38/10554	0.000587	0.007637	0.006228
Post NMDA receptor activation events	3/46	40/10554	0.000684	0.008298	0.006767
Signaling by NTRKs	4/46	99/10554	0.000879	0.009696	0.007907
Activation of NMDA receptors and postsynaptic events	3/46	44/10554	0.000906	0.009696	0.007907
Serotonin receptors	2/46	12/10554	0.001193	0.012063	0.009837
G alpha (i) signalling events	7/46	396/10554	0.001499	0.014359	0.01171
Transcriptional Regulation by MECP2	3/46	55/10554	0.001735	0.015788	0.012875
Neurexins and neuroligins	3/46	56/10554	0.001828	0.015839	0.012916

Ion transport by P-type ATPases	3/46	57/10554	0.001923	0.01591	0.012975
Na+/Cl- dependent neurotransmitter transporters	2/46	19/10554	0.003032	0.023989	0.019563
Synaptic adhesion-like molecules	2/46	21/10554	0.003702	0.028077	0.022896
Neurotransmitter receptors and postsynaptic signal transmission	4/46	156/10554	0.004627	0.033686	0.027471
HSF1-dependent transactivation	2/46	24/10554	0.004826	0.03378	0.027548
Amyotrophic Lateral Sclerosis Pathway Enrichment					
Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue
TRAF6-mediated induction of TAK1 complex within TLR4 complex	1/1	17/10619	0.001601	0.006278	NA
Activation of IRF3/IRF7 mediated by TBK1/IKK epsilon	1/1	18/10619	0.001695	0.006278	NA
IKK complex recruitment mediated by RIP1	1/1	25/10619	0.002354	0.006278	NA
Toll Like Receptor 3 (TLR3) Cascade	1/1	94/10619	0.008852	0.012431	NA
MyD88-independent TLR4 cascade	1/1	99/10619	0.009323	0.012431	NA
TRIF(TICAM1)-mediated TLR4 signaling	1/1	99/10619	0.009323	0.012431	NA
Toll Like Receptor 4 (TLR4) Cascade	1/1	129/10619	0.012148	0.013883	NA
Bipolar Disorder Pathway Enrichment					

Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue
Neuronal System	5/15	413/10654	0.000186	0.02621	0.017023
Transmission across Chemical Synapses	4/15	272/10654	0.000454	0.026344	0.01711
L1CAM interactions	3/15	119/10654	0.000561	0.026344	0.01711
Huntington's Disease Pathway Enrichment					
Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue
Tetrahydrobiopterin (BH4) synthesis, recycling, salvage and regulation	1/2	12/10554	0.002273	0.009084	0.000797
Mismatch repair (MMR) directed by MSH2:MSH3 (MutSbeta)	1/2	14/10554	0.002651	0.009084	0.000797
Mismatch Repair	1/2	15/10554	0.002841	0.009084	0.000797
Metabolism of folate and pterines	1/2	17/10554	0.003219	0.009084	0.000797
Metabolism of cofactors	1/2	20/10554	0.003787	0.009084	0.000797
Activation of E2F1 target genes at G1/S	1/2	28/10554	0.005299	0.009084	0.000797
G1/S-Specific Transcription	1/2	28/10554	0.005299	0.009084	0.000797
Metabolism of water-soluble vitamins and cofactors	1/2	123/10554	0.023174	0.032896	0.002886

G1/S Transition	1/2	131/10554	0.024672	0.032896	0.002886
Mitotic G1-G1/S phases	1/2	149/10554	0.028038	0.033645	0.002951
Metabolism of vitamins and cofactors	1/2	189/10554	0.035497	0.038724	0.003397
Multiple Sclerosis Pathway Enrichment					
Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue
Interleukin-35 Signalling	3/18	12/10554	9.08E-07	0.000213	0.000169
Interleukin-4 and Interleukin-13 signaling	4/18	108/10554	2.84E-05	0.003339	0.002647
Interleukin-10 signaling	3/18	47/10554	6.45E-05	0.004737	0.003756
Signaling by Interleukins	6/18	462/10554	8.06E-05	0.004737	0.003756
Interleukin-12 family signaling	3/18	58/10554	0.000121	0.005027	0.003986
Interleukin-6 signaling	2/18	11/10554	0.00015	0.005027	0.003986
Interleukin-27 signaling	2/18	11/10554	0.00015	0.005027	0.003986
RAF-independent MAPK1/3 activation	2/18	23/10554	0.000681	0.017204	0.01364
Interleukin-6 family signaling	2/18	24/10554	0.000742	0.017204	0.01364
Growth hormone receptor signaling	2/18	24/10554	0.000742	0.017204	0.01364

Interleukin-20 family signaling	2/18	25/10554	0.000805	0.017204	0.01364
Interleukin-7 signaling	2/18	36/10554	0.001672	0.032752	0.025967
Parkinson's Disease Pathway Enrichment					
Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue
Post NMDA receptor activation events	3/26	81/10619	0.00098	0.042878	0.036342
Ras activation upon Ca2+ influx through NMDA receptor	2/26	22/10619	0.001292	0.042878	0.036342
Unblocking of NMDA receptors, glutamate binding and activation	2/26	23/10619	0.001413	0.042878	0.036342
Negative regulation of NMDA receptor-mediated neuronal transmission	2/26	23/10619	0.001413	0.042878	0.036342
Activation of NMDA receptors and postsynaptic events	3/26	93/10619	0.001461	0.042878	0.036342
Long-term potentiation	2/26	25/10619	0.001671	0.042878	0.036342
Autism genes from SFARI database Pathway Enrichment					
Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue
Chromatin modifying enzymes	18/126	274/10619	3.38E-09	1.07E-06	8.30E-07
Chromatin organization	18/126	274/10619	3.38E-09	1.07E-06	8.30E-07

Diseases of signal transduction	20/126	389/10619	2.80E-08	5.89E-06	4.58E-06
Transcriptional Regulation by MECP2	9/126	62/10619	4.18E-08	6.60E-06	5.13E-06
Neuronal System	18/126	413/10619	1.68E-06	0.000195	0.000151
Regulation of MECP2 expression and activity	6/126	33/10619	2.11E-06	0.000195	0.000151
GABA synthesis, release, reuptake and degradation	5/126	19/10619	2.21E-06	0.000195	0.000151
Signaling by NTRKs	9/126	99/10619	2.47E-06	0.000195	0.000151
Cohesin Loading onto Chromatin	4/126	10/10619	3.76E-06	0.000264	0.000205
MAPK family signaling cascades	14/126	294/10619	9.71E-06	0.000536	0.000417
MAPK1/MAPK3 signaling	13/126	255/10619	9.94E-06	0.000536	0.000417
SUMO E3 ligases SUMOylate target proteins	11/126	182/10619	1.02E-05	0.000536	0.000417
SUMOylation	11/126	188/10619	1.38E-05	0.000664	0.000517
Intracellular signaling by second messengers	14/126	305/10619	1.47E-05	0.000664	0.000517
Mitotic Telophase/Cytokinesis	4/126	14/10619	1.73E-05	0.000727	0.000565
Transmission across Chemical Synapses	13/126	272/10619	1.98E-05	0.000738	0.000574
PKMTs methylate histone lysines	7/126	71/10619	1.98E-05	0.000738	0.000574

Transcriptional regulation by RUNX1	12/126	239/10619	2.59E-05	0.000908	0.000706
Neurotransmitter release cycle	6/126	51/10619	2.88E-05	0.000958	0.000745
RAF activation	5/126	32/10619	3.38E-05	0.001069	0.000832
Signaling by NTRK1 (TRKA)	7/126	78/10619	3.68E-05	0.001097	0.000853
RAF/MAP kinase cascade	12/126	249/10619	3.87E-05	0.001097	0.000853
MECP2 regulates neuronal receptors and channels	4/126	17/10619	3.99E-05	0.001097	0.000853
Cardiac conduction	9/126	141/10619	4.39E-05	0.001155	0.000899
Neurexins and neuroligins	6/126	56/10619	4.95E-05	0.001251	0.000973
PIP3 activates AKT signaling	12/126	264/10619	6.83E-05	0.00166	0.001291
Signaling by RAS mutants	6/126	60/10619	7.34E-05	0.001718	0.001337
Transcriptional Regulation by TP53	14/126	365/10619	0.000105	0.002367	0.001841
RUNX1 regulates genes involved in megakaryocyte differentiation and platelet function	7/126	97/10619	0.000149	0.003238	0.002519
Glutamate Neurotransmitter Release Cycle	4/126	24/10619	0.000167	0.003406	0.002649
HSF1-dependent transactivation	4/126	24/10619	0.000167	0.003406	0.002649

Phase 0 - rapid depolarisation	5/126	46/10619	0.000202	0.003983	0.003098
Establishment of Sister Chromatid Cohesion	3/126	11/10619	0.000251	0.004808	0.00374
Energy dependent regulation of mTOR by LKB1-AMPK	4/126	29/10619	0.000357	0.00663	0.005157
CREB1 phosphorylation through NMDA receptor-mediated activation of RAS signaling	4/126	30/10619	0.000408	0.007364	0.005728
Oncogenic MAPK signaling	6/126	83/10619	0.000443	0.007782	0.006053
Ion homeostasis	5/126	56/10619	0.000511	0.008526	0.006632
TP53 Regulates Metabolic Genes	6/126	86/10619	0.000536	0.008526	0.006632
Loss of function of MECP2 in Rett syndrome	3/126	14/10619	0.00054	0.008526	0.006632
Pervasive developmental disorders	3/126	14/10619	0.00054	0.008526	0.006632
Protein-protein interactions at synapses	6/126	87/10619	0.000571	0.008797	0.006842
Spry regulation of FGF signaling	3/126	16/10619	0.000816	0.012277	0.00955
Muscle contraction	9/126	209/10619	0.000841	0.012366	0.009618
Acetylcholine Neurotransmitter Release Cycle	3/126	17/10619	0.000982	0.013794	0.010729
Platelet sensitization by LDL	3/126	17/10619	0.000982	0.013794	0.010729

RUNX1 interacts with co-factors whose precise effect on RUNX1 targets is not known	4/126	38/10619	0.001021	0.014026	0.01091
Signaling by Receptor Tyrosine Kinases	14/126	458/10619	0.001043	0.014031	0.010913
RORA activates gene expression	3/126	18/10619	0.001168	0.014769	0.011488
Serotonin Neurotransmitter Release Cycle	3/126	18/10619	0.001168	0.014769	0.011488
Norepinephrine Neurotransmitter Release Cycle	3/126	18/10619	0.001168	0.014769	0.011488
mTOR signalling	4/126	40/10619	0.001241	0.015214	0.011834
PI3K/AKT Signaling in Cancer	6/126	101/10619	0.001252	0.015214	0.011834
Costimulation by the CD28 family	5/126	70/10619	0.00142	0.016615	0.012924
Circadian Clock	5/126	70/10619	0.00142	0.016615	0.012924
Integration of energy metabolism	6/126	108/10619	0.001767	0.020307	0.015795
CTLA4 inhibitory signaling	3/126	21/10619	0.001856	0.020943	0.016289
Negative regulation of the PI3K/AKT network	6/126	110/10619	0.00194	0.021516	0.016735
Estrogen-dependent gene expression	7/126	150/10619	0.002029	0.022109	0.017197
ERK/MAPK targets	3/126	22/10619	0.00213	0.022438	0.017452

Ras activation upon Ca2+ influx through NMDA receptor	3/126	22/10619	0.00213	0.022438	0.017452
NOTCH1 Intracellular Domain Regulates Transcription	4/126	47/10619	0.002272	0.022797	0.017732
Signaling by moderate kinase activity BRAF mutants	4/126	47/10619	0.002272	0.022797	0.017732
Paradoxical activation of RAF signaling by kinase inactive BRAF	4/126	47/10619	0.002272	0.022797	0.017732
Dopamine Neurotransmitter Release Cycle	3/126	23/10619	0.002429	0.023256	0.018089
Unblocking of NMDA receptors, glutamate binding and activation	3/126	23/10619	0.002429	0.023256	0.018089
Negative regulation of NMDA receptor-mediated neuronal transmission	3/126	23/10619	0.002429	0.023256	0.018089
Regulation of lipid metabolism by Peroxisome proliferator-activated receptor alpha (PPARalpha)	6/126	119/10619	0.002884	0.027205	0.021161
Nuclear Events (kinase and transcription factor activation)	3/126	25/10619	0.0031	0.027989	0.02177
Signaling by NTRK2 (TRKB)	3/126	25/10619	0.0031	0.027989	0.02177
Long-term potentiation	3/126	25/10619	0.0031	0.027989	0.02177
Activation of anterior HOX genes in hindbrain development during early embryogenesis	6/126	122/10619	0.003264	0.028654	0.022288
Activation of HOX genes during differentiation	6/126	122/10619	0.003264	0.028654	0.022288

Notch-HLH transcription pathway	3/126	28/10619	0.004303	0.037252	0.028975
Ion transport by P-type ATPases	4/126	57/10619	0.004595	0.03864	0.030054
Negative regulation of FGFR3 signaling	3/126	29/10619	0.004758	0.03864	0.030054
Signaling by NOTCH1 PEST Domain Mutants in Cancer	4/126	58/10619	0.004891	0.03864	0.030054
Signaling by NOTCH1 in Cancer	4/126	58/10619	0.004891	0.03864	0.030054
Constitutive Signaling by NOTCH1 PEST Domain Mutants	4/126	58/10619	0.004891	0.03864	0.030054
Signaling by NOTCH1 HD+PEST Domain Mutants in Cancer	4/126	58/10619	0.004891	0.03864	0.030054
Constitutive Signaling by NOTCH1 HD+PEST Domain Mutants	4/126	58/10619	0.004891	0.03864	0.030054
ESR-mediated signaling	8/126	223/10619	0.005024	0.039202	0.030492
Regulation of TP53 Activity through Acetylation	3/126	30/10619	0.005241	0.040398	0.031422
MAPK targets/ Nuclear events mediated by MAP kinases	3/126	31/10619	0.005753	0.043287	0.033669
Negative regulation of FGFR4 signaling	3/126	31/10619	0.005753	0.043287	0.033669
Regulation of FOXO transcriptional activity by acetylation	2/126	10/10619	0.005906	0.043913	0.034156
PTEN Regulation	6/126	140/10619	0.00638	0.046882	0.036466
Negative regulation of FGFR1 signaling	3/126	33/10619	0.006864	0.049865	0.038786

Appendix E: Gene Ontology Enrichment Results

Table E.1: Gene Ontology results after Matrix-GC. For GORilla Gene Ontology, enrichment is defined as (b/n)/(B/N) with N: the total number of genes, B: the total number of genes related to a specific GO term, n: the number of genes in the dataset of interest, and b: the number of genes in the dataset of interest that related to a specific GO term [121]. Italicised grey terms represent results that are also significant after control permutation analysis.

Alzheimer's Disease Gene Ontology Enrichment							
Description	P-value	FDR q-value	Enrichment	N	B	n	b
negative regulation of amyloid precursor protein catabolic process	3.56E-08	0.000272	414.02	17941	13	10	3
negative regulation of amyloid-beta formation	2.05E-08	0.000314	489.3	17941	11	10	3
negative regulation of aspartic-type peptidase activity	2.8E-07	0.000857	1794.1	17941	2	10	2
regulation of amyloid precursor protein catabolic process	3.62E-07	0.000925	199.34	17941	27	10	3
regulation of amyloid-beta formation	1.91E-07	0.000975	244.65	17941	22	10	3
negative regulation of aspartic-type endopeptidase activity involved in amyloid precursor protein catabolic process	2.8E-07	0.00107	1794.1	17941	2	10	2
regulation of cellular amide metabolic process	*1.7E-06*	*0.00373*	*21.01*	*17941*	*427*	*10*	*5*

negative regulation of cellular amide metabolic process	2.76E-06	0.00528	36.61	17941	196	10	4
negative regulation of protein metabolic process	7.01E-06	0.0119	10.19	17941	1056	10	6
regulation of aspartic-type endopeptidase activity involved in amyloid precursor protein catabolic process	7.82E-06	0.012	448.52	17941	8	10	2
regulation of aspartic-type peptidase activity	0.00001	0.014	398.69	17941	9	10	2
regulation of actin cytoskeleton organization	*0.000018*	*0.023*	*22.78*	*17941*	*315*	*10*	*4*
regulation of actin filament-based process	*2.49E-05*	*0.0293*	*20.98*	*17941*	*342*	*10*	*4*
endocytosis	4.29E-05	0.0469	18.26	17941	393	10	4
Attention Deficit Hyperactivity Disorder Gene Ontology Enrichment							
Description	P-value	FDR q-value	Enrichment	N	B	n	b
behavior	1.49E-14	2.29E-10	10.09	17942	528	64	19
signaling	8.86E-14	6.78E-10	9.14	17942	583	64	19
learning or memory	2.16E-12	5.51E-09	15.25	17942	239	64	13
cell-cell signaling	2.85E-12	6.25E-09	9.01	17942	529	64	17
trans-synaptic signaling	2.13E-12	6.53E-09	13.13	17942	299	64	14

synaptic signaling	2.13E-12	8.16E-09	13.13	17942	299	64	14
learning	1.96E-12	9.98E-09	22.35	17942	138	64	11
cognition	1.16E-11	2.22E-08	13.35	17942	273	64	13
response to ammonium ion	1.93E-11	3.29E-08	29.34	17942	86	64	9
system process	4.29E-11	5.47E-08	4.95	17942	1303	64	23
regulation of catecholamine metabolic process	4.24E-11	5.9E-08	88.53	17942	19	64	6
regulation of dopamine metabolic process	4.24E-11	6.49E-08	88.53	17942	19	64	6
dopamine metabolic process	2.72E-10	3.21E-07	67.28	17942	25	64	6
response to cocaine	1.37E-09	1.39E-06	52.56	17942	32	64	6
cell communication	1.28E-09	1.4E-06	6.09	17942	783	64	17
regulation of trans-synaptic signaling	1.6E-09	1.44E-06	8.95	17942	407	64	13
modulation of chemical synaptic transmission	1.56E-09	1.49E-06	8.98	17942	406	64	13
response to xenobiotic stimulus	2.44E-09	2.08E-06	17.16	17942	147	64	9
chemical synaptic transmission	3.93E-09	3.01E-06	11.01	17942	280	64	11
anterograde trans-synaptic signaling	3.93E-09	3.17E-06	11.01	17942	280	64	11

catechol-containing compound metabolic process	4.82E-09	3.36E-06	43.13	17942	39	64	6
catecholamine metabolic process	4.82E-09	3.52E-06	43.13	17942	39	64	6
G protein-coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger	6.6E-09	4.4E-06	12.46	17942	225	64	10
regulation of amine transport	2.26E-08	1.44E-05	22.82	17942	86	64	7
adenylate cyclase-modulating G protein-coupled receptor signaling pathway	2.89E-08	1.77E-05	12.94	17942	195	64	9
response to toxic substance	3.29E-08	1.87E-05	8.96	17942	344	64	11
response to anesthetic	3.26E-08	1.92E-05	31.74	17942	53	64	6
nervous system process	3.79E-08	2.07E-05	5.22	17942	860	64	16
regulation of neurotransmitter levels	9.73E-08	5.14E-05	9.38	17942	299	64	10
response to drug	1.37E-07	7.02E-05	5.58	17942	703	64	14
locomotory behavior	1.7E-07	8.39E-05	13.04	17942	172	64	8
phenol-containing compound metabolic process	2.69E-07	0.000129	22.43	17942	75	64	6
regulation of cellular amine metabolic process	3.68E-07	0.000171	21.29	17942	79	64	6
response to alkaloid	5.3E-07	0.000239	20.02	17942	84	64	6

dopamine catabolic process	8.59E-07	0.000366	140.17	17942	6	64	3
regulation of response to drug	8.53E-07	0.000373	18.48	17942	91	64	6
regulation of ion transport	1.11E-06	0.000461	5.64	17942	597	64	12
catecholamine catabolic process	1.5E-06	0.000589	120.15	17942	7	64	3
catechol-containing compound catabolic process	1.5E-06	0.000605	120.15	17942	7	64	3
regulation of catecholamine secretion	1.79E-06	0.000685	24.59	17942	57	64	5
cerebral cortex development	2.13E-06	0.000794	23.76	17942	59	64	5
divalent inorganic cation homeostasis	2.47E-06	0.000901	6.58	17942	426	64	10
memory	2.89E-06	0.00103	15.02	17942	112	64	6
phenol-containing compound catabolic process	3.58E-06	0.00119	93.45	17942	9	64	3
regulation of synaptic transmission, glutamatergic	3.45E-06	0.0012	21.56	17942	65	64	5
neuron migration	3.54E-06	0.00121	14.5	17942	116	64	6
regulation of tube diameter	3.91E-06	0.00125	14.25	17942	118	64	6
regulation of blood vessel diameter	3.91E-06	0.00128	14.25	17942	118	64	6
negative regulation of secretion by cell	*4.34E-06*	*0.00133*	*10.55*	*17942*	*186*	*64*	*7*

multicellular organismal process	4.3E-06	0.00135	2.38	17942	3183	64	27
circulatory system process	4.65E-06	0.0014	10.44	17942	188	64	7
response to amphetamine	4.91E-06	0.00145	35.04	17942	32	64	4
regulation of blood vessel size	5.46E-06	0.00158	13.46	17942	125	64	6
neuron cellular homeostasis	5.57E-06	0.00158	33.98	17942	33	64	4
regulation of tube size	5.72E-06	0.00159	13.35	17942	126	64	6
response to organic cyclic compound	6.11E-06	0.00167	4.78	17942	704	64	12
associative learning	6.56E-06	0.00173	18.94	17942	74	64	5
regulation of neurotransmitter transport	6.55E-06	0.00176	13.04	17942	129	64	6
dopamine transport	0.000007	0.00182	76.46	17942	11	64	3
regulation of nervous system development	7.59E-06	0.00191	4.29	17942	849	64	13
cellular calcium ion homeostasis	7.52E-06	0.00192	6.66	17942	379	64	9
adult behavior	*7.81E-06*	*0.00193*	*12.65*	*17942*	*133*	*64*	*6*
positive regulation of nervous system development	8.35E-06	0.00203	5.73	17942	489	64	10
regulation of dopamine secretion	8.9E-06	0.00213	30.31	17942	37	64	4

regulation of neuron differentiation	9.13E-06	0.00215	5.06	17942	609	64	11
trans-synaptic signaling, modulating synaptic transmission	9.31E-06	0.00216	70.09	17942	12	64	3
negative regulation of secretion	9.61E-06	0.0022	9.34	17942	210	64	7
calcium ion homeostasis	9.86E-06	0.00222	6.44	17942	392	64	9
response to organonitrogen compound	1.07E-05	0.00237	4.52	17942	744	64	12
sensitization	1.25E-05	0.00259	280.34	17942	2	64	2
feeding behavior	1.22E-05	0.0026	16.69	17942	84	64	5
innate vocalization behavior	1.25E-05	0.00263	280.34	17942	2	64	2
catecholamine transport	1.21E-05	0.00264	64.69	17942	13	64	3
regulation of neurogenesis	1.22E-05	0.00264	4.46	17942	754	64	12
cellular divalent inorganic cation homeostasis	0.000013	0.00266	6.21	17942	406	64	9
regulation of biological quality	1.36E-05	0.00275	2.24	17942	3379	64	27
vascular process in circulatory system	*1.61E-05*	*0.00321*	*11.14*	*17942*	*151*	*64*	*6*
response to amine	1.64E-05	0.00321	26.08	17942	43	64	4

metal ion homeostasis	0.00002	0.00389	5.18	17942	541	64	10
positive regulation of neurogenesis	2.06E-05	0.00394	5.87	17942	430	64	9
response to nitrogen compound	0.000024	0.00454	4.17	17942	807	64	12
positive regulation of neuron differentiation	2.47E-05	0.00462	6.67	17942	336	64	8
vocalization behavior	2.84E-05	0.00524	49.47	17942	17	64	3
regulation of synaptic plasticity	3.15E-05	0.00575	9.89	17942	170	64	6
response to bronchodilator	3.24E-05	0.00585	21.99	17942	51	64	4
regulation of neurotransmitter uptake	0.000034	0.00606	46.72	17942	18	64	3
negative regulation of serotonin secretion	3.75E-05	0.00638	186.9	17942	3	64	2
norepinephrine-epinephrine vasoconstriction involved in regulation of systemic arterial blood pressure	3.75E-05	0.00645	186.9	17942	3	64	2
trans-synaptic signaling by endocannabinoid, modulating synaptic transmission	3.75E-05	0.00653	186.9	17942	3	64	2
trans-synaptic signaling by lipid, modulating synaptic transmission	3.75E-05	0.0066	186.9	17942	3	64	2
ammonium ion metabolic process	3.95E-05	0.00665	9.5	17942	177	64	6

transmission of nerve impulse	4.03E-05	0.00671	44.26	17942	19	64	3
negative regulation of synaptic transmission	4.38E-05	0.00722	20.39	17942	55	64	4
regulation of cell development	4.73E-05	0.00755	3.89	17942	865	64	12
cellular metal ion homeostasis	4.78E-05	0.00755	5.27	17942	479	64	9
negative regulation of neurotransmitter transport	4.73E-05	0.00762	42.05	17942	20	64	3
monoamine transport	4.73E-05	0.0077	42.05	17942	20	64	3
regulation of transport	5.22E-05	0.00808	2.86	17942	1666	64	17
developmental process	5.17E-05	0.00809	1.92	17942	4530	64	31
cation homeostasis	5.55E-05	0.0085	4.6	17942	610	64	10
inorganic ion homeostasis	6.44E-05	0.00977	4.51	17942	621	64	10
positive regulation of cell development	6.55E-05	0.00984	5.06	17942	499	64	9
adenylate cyclase-activating G protein-coupled receptor signaling pathway	6.85E-05	0.0102	11.68	17942	120	64	5
multicellular organismal signaling	7.03E-05	0.0104	18.09	17942	62	64	4
regulation of membrane potential	7.22E-05	0.0105	5.74	17942	391	64	8

neurotransmitter uptake	7.29E-05	0.0105	36.57	17942	23	64	3
positive regulation of blood circulation	7.97E-05	0.0114	17.52	17942	64	64	4
neurotransmitter transport	0.00008	0.0114	11.3	17942	124	64	5
regulation of multicellular organismal process	8.14E-05	0.0114	2.25	17942	2872	64	23
cell migration	8.43E-05	0.0117	3.97	17942	776	64	11
cellular homeostasis	9.03E-05	0.0122	3.94	17942	782	64	11
phospholipase C-activating G protein-coupled receptor signaling pathway	8.99E-05	0.0123	16.99	17942	66	64	4
negative regulation of ion transport	8.96E-05	0.0124	11.04	17942	127	64	5
system development	9.41E-05	0.0126	4.31	17942	650	64	10
second-messenger-mediated signaling	9.96E-05	0.0133	6.48	17942	303	64	7
positive regulation of multicellular organismal process	0.00011	0.0145	2.82	17942	1592	64	16
locomotion	0.000115	0.015	3.54	17942	949	64	12
regulation of localization	0.000118	0.0154	2.32	17942	2536	64	21
regulation of secretion by cell	0.000122	0.0157	4.18	17942	671	64	10

cellular cation homeostasis	0.000128	0.0164	4.63	17942	545	64	9
regulation of cell communication	0.000131	0.0166	2.07	17942	3384	64	25
cAMP-mediated signaling	0.000137	0.0172	10.08	17942	139	64	5
ion homeostasis	0.000141	0.0176	4.1	17942	683	64	10
positive regulation of synaptic transmission	0.000147	0.0181	9.94	17942	141	64	5
positive regulation of vasoconstriction	0.000148	0.0181	29	17942	29	64	3
cellular ion homeostasis	0.000151	0.0184	4.53	17942	557	64	9
regulation of signaling	0.000157	0.0187	2.05	17942	3420	64	25
acid secretion	0.000156	0.0188	14.75	17942	76	64	4
vasoconstriction	0.000164	0.0193	28.03	17942	30	64	3
reproductive behavior	0.000164	0.0195	28.03	17942	30	64	3
regulation of developmental process	0.000189	0.0206	2.32	17942	2420	64	20
nonassociative learning	0.000186	0.0207	93.45	17942	6	64	2
negative regulation of transport	0.000187	0.0207	4.99	17942	449	64	8
positive regulation of neurotransmitter uptake	0.000186	0.0208	93.45	17942	6	64	2

negative regulation of oxidoreductase activity	0.000181	0.0209	27.13	17942	31	64	3
dopamine receptor signaling pathway	0.000181	0.021	27.13	17942	31	64	3
regulation of serotonin secretion	0.000186	0.021	93.45	17942	6	64	2
trans-synaptic signaling by endocannabinoid	0.000186	0.0211	93.45	17942	6	64	2
regulation of synaptic vesicle transport	0.000181	0.0212	14.19	17942	79	64	4
central nervous system neuron development	0.000199	0.0212	26.28	17942	32	64	3
trans-synaptic signaling by lipid	0.000186	0.0213	93.45	17942	6	64	2
fear response	0.000199	0.0214	26.28	17942	32	64	3
cellular response to catecholamine stimulus	0.000199	0.0215	26.28	17942	32	64	3
cellular response to monoamine stimulus	0.000199	0.0217	26.28	17942	32	64	3
regulation of neuron projection development	0.000205	0.0217	4.93	17942	455	64	8
response to chemical	0.000222	0.0225	2.29	17942	2449	64	20
cell projection organization	0.000217	0.0226	3.89	17942	720	64	10
positive regulation of cell projection organization	0.000218	0.0226	5.7	17942	344	64	7
regulation of secretion	0.000222	0.0226	3.88	17942	722	64	10

G protein-coupled receptor signaling pathway	0.000216	0.0227	3.1	17942	1174	64	13
cell motility	0.000221	0.0227	3.56	17942	866	64	11
phospholipase C-activating dopamine receptor signaling pathway	0.00026	0.0257	80.1	17942	7	64	2
positive regulation of neuron projection development	0.000263	0.0258	6.73	17942	250	64	6
cyclic-nucleotide-mediated signaling	0.000264	0.0258	8.76	17942	160	64	5
vocal learning	0.00026	0.0259	80.1	17942	7	64	2
imitative learning	0.00026	0.026	80.1	17942	7	64	2
observational learning	0.00026	0.0262	80.1	17942	7	64	2
response to catecholamine	0.000284	0.0274	23.36	17942	36	64	3
response to monoamine	0.000284	0.0276	23.36	17942	36	64	3
regulation of plasma membrane bounded cell projection organization	0.0003	0.0286	4.13	17942	611	64	9
regulation of anatomical structure size	0.000302	0.0286	5.41	17942	363	64	7
regulation of insulin secretion	*0.000305*	*0.0286*	*8.5*	*17942*	*165*	*64*	*5*
regulation of multicellular organismal development	0.0003	0.0287	2.48	17942	1921	64	17

positive regulation of protein kinase activity	0.000316	0.0295	4.62	17942	485	64	8
positive regulation of neurotransmitter transport	0.000334	0.031	22.13	17942	38	64	3
righting reflex	0.000346	0.031	70.09	17942	8	64	2
regulation of cell projection organization	0.000338	0.0312	4.06	17942	621	64	9
regulation of dopamine uptake involved in synaptic transmission	0.000346	0.0312	70.09	17942	8	64	2
catecholamine uptake	0.000346	0.0314	70.09	17942	8	64	2
dopamine uptake	0.000346	0.0315	70.09	17942	8	64	2
regulation of catecholamine uptake involved in synaptic transmission	0.000346	0.0317	70.09	17942	8	64	2
negative regulation of blood vessel diameter	0.000361	0.0322	21.56	17942	39	64	3
response to oxygen-containing compound	0.000386	0.0342	2.92	17942	1246	64	13
signal transduction	0.000398	0.0351	1.82	17942	4306	64	28
response to organic substance	0.000407	0.0356	2.42	17942	1971	64	17
cellular chemical homeostasis	0.000417	0.0363	3.95	17942	639	64	9
regulation of neurotransmitter secretion	0.00043	0.0372	11.33	17942	99	64	4

reactive oxygen species metabolic process	0.000447	0.0384	11.21	17942	100	64	4
regulation of calcium ion-dependent exocytosis	0.0005	0.0428	10.89	17942	103	64	4
positive regulation of kinase activity	0.000521	0.0443	4.29	17942	523	64	8
adenylate cyclase-activating dopamine receptor signaling pathway	0.000553	0.0456	56.07	17942	10	64	2
norepinephrine metabolic process	0.000553	0.0458	56.07	17942	10	64	2
regulation of uterine smooth muscle contraction	0.000553	0.0461	56.07	17942	10	64	2
learned vocalization behavior or vocal learning	0.000553	0.0463	56.07	17942	10	64	2
behavioral response to ethanol	0.000553	0.0466	56.07	17942	10	64	2
long-term synaptic potentiation	0.000552	0.0467	18.69	17942	45	64	3
anatomical structure development	0.000584	0.0478	2.02	17942	3053	64	22
visual learning	0.000589	0.048	18.28	17942	46	64	3
Schizophrenia Gene Ontology Enrichment							
Description	P-value	FDR q-value	Enrichment	N	B	n	b
regulation of neuron differentiation	3.04E-06	0.0234	2.45	17961	612	383	32

regulation of neurogenesis	3.04E-06	0.0467	2.27	17961	764	383	37
Autism genes from SFARI database Gene Ontology Enrichment							
Description	P-value	FDR q-value	Enrichment	N	B	n	b
positive regulation of biological process	2.38E-17	3.66E-13	2.05	17961	5521	160	101
developmental process	9.87E-17	7.59E-13	2.21	17961	4573	160	90
positive regulation of gene expression	2.3E-15	1.18E-11	3.24	17961	1834	160	53
regulation of transcription, DNA-templated	2.98E-15	1.15E-11	2.52	17961	3164	160	71
positive regulation of cellular process	3.94E-15	1.21E-11	2.07	17961	4928	160	91
chromatin organization	4.23E-15	1.08E-11	5.6	17961	621	160	31
regulation of cellular macromolecule biosynthetic process	5.49E-15	1.21E-11	2.36	17961	3612	160	76
regulation of nucleic acid-templated transcription	7.3E-15	1.4E-11	2.48	17961	3218	160	71
regulation of RNA biosynthetic process	8.06E-15	1.38E-11	2.47	17961	3224	160	71
regulation of RNA metabolic process	2.1E-14	3.23E-11	2.38	17961	3448	160	73
positive regulation of metabolic process	3.57E-14	4.99E-11	2.4	17961	3317	160	71
regulation of macromolecule biosynthetic process	3.59E-14	4.6E-11	2.28	17961	3736	160	76

positive regulation of RNA metabolic process	4.4E-14	5.2E-11	3.36	17961	1572	160	47
regulation of biosynthetic process	9.67E-14	1.06E-10	2.2	17961	3978	160	78
regulation of cellular biosynthetic process	1.17E-13	1.2E-10	2.21	17961	3904	160	77
positive regulation of nucleic acid-templated transcription	1.4E-13	1.34E-10	3.38	17961	1494	160	45
positive regulation of RNA biosynthetic process	1.43E-13	1.29E-10	3.38	17961	1495	160	45
regulation of transcription by RNA polymerase II	1.87E-13	1.6E-10	2.67	17961	2476	160	59
cellular component organization	2.5E-13	2.02E-10	2.02	17961	4775	160	86
regulation of nucleobase-containing compound metabolic process	4.14E-13	3.18E-10	2.22	17961	3734	160	74
positive regulation of transcription, DNA-templated	4.35E-13	3.18E-10	3.41	17961	1416	160	43
cellular component organization or biogenesis	4.51E-13	3.15E-10	2	17961	4822	160	86
positive regulation of nucleobase-containing compound metabolic process	5.42E-13	3.62E-10	3.08	17961	1750	160	48
anatomical structure development	5.86E-13	3.75E-10	2.4	17961	3090	160	66
positive regulation of cellular metabolic process	3.56E-12	2.19E-09	2.36	17961	3047	160	64

cognition	4.88E-12	2.88E-09	7.78	17961	274	160	19
positive regulation of macromolecule metabolic process	5.23E-12	2.98E-09	2.34	17961	3073	160	64
positive regulation of macromolecule biosynthetic process	5.8E-12	3.18E-09	2.98	17961	1730	160	46
regulation of cell differentiation	8.64E-12	4.58E-09	3	17961	1682	160	45
positive regulation of cellular biosynthetic process	9.54E-12	4.89E-09	2.89	17961	1824	160	47
positive regulation of biosynthetic process	1.63E-11	8.07E-09	2.85	17961	1852	160	47
regulation of gene expression	1.63E-11	7.83E-09	1.99	17961	4463	160	79
regulation of primary metabolic process	3.49E-11	1.62E-08	1.81	17961	5587	160	90
positive regulation of transcription by RNA polymerase II	3.87E-11	1.75E-08	3.54	17961	1110	160	35
regulation of nitrogen compound metabolic process	5.39E-11	2.37E-08	1.82	17961	5429	160	88
positive regulation of nitrogen compound metabolic process	7.42E-11	3.17E-08	2.3	17961	2931	160	60
negative regulation of cellular biosynthetic process	7.68E-11	3.19E-08	3.09	17961	1455	160	40
negative regulation of biosynthetic process	1.23E-10	4.99E-08	3.04	17961	1478	160	40
regulation of cellular metabolic process	1.38E-10	5.43E-08	1.77	17961	5718	160	90

system development	1.79E-10	6.89E-08	4.42	17961	661	160	26
negative regulation of cellular process	1.9E-10	7.11E-08	1.94	17961	4396	160	76
covalent chromatin modification	2.35E-10	8.62E-08	6.22	17961	343	160	19
negative regulation of cellular macromolecule biosynthetic process	2.37E-10	8.46E-08	3.16	17961	1313	160	37
behavior	2.72E-10	9.51E-08	4.93	17961	524	160	23
negative regulation of macromolecule biosynthetic process	3.17E-10	1.08E-07	3.06	17961	1392	160	38
learning or memory	4.61E-10	1.54E-07	7.48	17961	240	160	16
regulation of developmental process	4.68E-10	1.53E-07	2.41	17961	2422	160	52
modulation of chemical synaptic transmission	6.58E-10	2.11E-07	5.5	17961	408	160	20
regulation of trans-synaptic signaling	6.87E-10	2.15E-07	5.49	17961	409	160	20
regulation of cell development	7.82E-10	2.4E-07	3.72	17961	875	160	29
regulation of neurogenesis	8.18E-10	2.47E-07	3.97	17961	764	160	27
regulation of neuron projection development	9.34E-10	2.76E-07	5.1	17961	462	160	21
regulation of neuron differentiation	1.04E-09	3.01E-07	4.4	17961	612	160	24

negative regulation of RNA biosynthetic process	1.04E-09	2.97E-07	3.29	17961	1127	160	33
negative regulation of nucleic acid-templated transcription	1.04E-09	2.92E-07	3.29	17961	1127	160	33
histone modification	1.07E-09	2.93E-07	6.07	17961	333	160	18
regulation of multicellular organismal process	1.24E-09	3.34E-07	2.21	17961	2891	160	57
negative regulation of transcription by RNA polymerase II	1.25E-09	3.31E-07	3.89	17961	779	160	27
regulation of plasma membrane bounded cell projection organization	1.39E-09	3.61E-07	4.34	17961	621	160	24
negative regulation of RNA metabolic process	1.44E-09	3.7E-07	3.17	17961	1205	160	34
negative regulation of transcription, DNA-templated	1.47E-09	3.7E-07	3.33	17961	1080	160	32
regulation of metabolic process	1.49E-09	3.68E-07	1.65	17961	6478	160	95
regulation of cell projection organization	1.9E-09	4.64E-07	4.27	17961	631	160	24
positive regulation of neurogenesis	2.07E-09	4.98E-07	5.15	17961	436	160	20
regulation of macromolecule metabolic process	2.24E-09	5.3E-07	1.68	17961	6002	160	90
negative regulation of cellular metabolic process	2.24E-09	5.22E-07	2.37	17961	2373	160	50

regulation of cellular component organization	2.44E-09	5.59E-07	2.39	17961	2301	160	49
positive regulation of nervous system development	3.31E-09	7.49E-07	4.75	17961	496	160	21
cellular protein modification process	3.65E-09	8.14E-07	2.2	17961	2809	160	55
protein modification process	3.65E-09	8.02E-07	2.2	17961	2809	160	55
negative regulation of nucleobase-containing compound metabolic process	4.5E-09	9.74E-07	2.96	17961	1326	160	35
regulation of cell morphogenesis	4.62E-09	9.87E-07	4.91	17961	457	160	20
regulation of synaptic plasticity	5.77E-09	1.21E-06	8.34	17961	175	160	13
negative regulation of nitrogen compound metabolic process	7.33E-09	1.52E-06	2.37	17961	2222	160	47
negative regulation of biological process	1E-08	2.06E-06	1.74	17961	5158	160	80
regulation of nervous system development	1.01E-08	2.04E-06	3.53	17961	859	160	27
positive regulation of neuron differentiation	1.02E-08	2.04E-06	5.61	17961	340	160	17
social behavior	1.1E-08	2.17E-06	18.33	17961	49	160	8
intraspecies interaction between organisms	1.1E-08	2.14E-06	18.33	17961	49	160	8
macromolecule modification	1.41E-08	2.72E-06	2.1	17961	3000	160	56

cellular protein metabolic process	1.47E-08	2.79E-06	2.07	17961	3088	160	57
positive regulation of cell projection organization	1.57E-08	2.95E-06	5.45	17961	350	160	17
multicellular organismal process	2.2E-08	4.08E-06	2.03	17961	3208	160	58
positive regulation of cell development	2.55E-08	4.66E-06	4.44	17961	506	160	20
multicellular organism growth	3.04E-08	5.5E-06	12.95	17961	78	160	9
plasma membrane bounded cell projection morphogenesis	3.42E-08	6.11E-06	7.19	17961	203	160	13
cell projection organization	3.9E-08	6.89E-06	3.65	17961	738	160	24
cell projection morphogenesis	4.31E-08	7.52E-06	7.05	17961	207	160	13
positive regulation of synaptic transmission	4.32E-08	7.45E-06	8.95	17961	138	160	11
learning	5.01E-08	8.55E-06	8.82	17961	140	160	11
regulation of cellular process	5.9E-08	9.97E-06	1.37	17961	9909	160	121
positive regulation of excitatory postsynaptic potential	7E-08	1.17E-05	26.94	17961	25	160	6
positive regulation of neuron projection development	7.25E-08	0.000012	6.14	17961	256	160	14
positive regulation of cellular component organization	8E-08	1.31E-05	2.95	17961	1143	160	30

cellular macromolecule metabolic process	9.51E-08	1.54E-05	1.79	17961	4339	160	69
multi-organism behavior	9.65E-08	1.54E-05	14.03	17961	64	160	8
animal organ development	9.86E-08	1.56E-05	2.92	17961	1154	160	30
cell part morphogenesis	1.14E-07	0.000018	6.49	17961	225	160	13
negative regulation of metabolic process	1.17E-07	1.82E-05	2.01	17961	3012	160	54
regulation of dendritic spine development	1.23E-07	1.89E-05	13.61	17961	66	160	8
regulation of developmental growth	1.26E-07	1.92E-05	5.41	17961	311	160	15
chromatin remodeling	1.33E-07	2.01E-05	8.02	17961	154	160	11
regulation of biological process	1.36E-07	2.03E-05	1.32	17961	10785	160	127
histone lysine methylation	1.75E-07	2.59E-05	13.02	17961	69	160	8
neuron projection morphogenesis	2.14E-07	3.13E-05	6.8	17961	198	160	12
brain development	2.26E-07	3.27E-05	6.77	17961	199	160	12
regulation of multicellular organismal development	2.37E-07	0.000034	2.33	17961	1929	160	40
visual behavior	2.45E-07	3.48E-05	16.04	17961	49	160	7
peptidyl-lysine modification	2.49E-07	3.51E-05	5.55	17961	283	160	14

regulation of cell morphogenesis involved in differentiation	2.49E-07	3.47E-05	5.55	17961	283	160	14
developmental growth	2.49E-07	3.44E-05	5.55	17961	283	160	14
growth	3.21E-07	0.000044	5.44	17961	289	160	14
negative regulation of gene expression	3.56E-07	4.84E-05	2.29	17961	1959	160	40
negative regulation of macromolecule metabolic process	4.56E-07	6.15E-05	2.01	17961	2793	160	50
anatomical structure morphogenesis	4.72E-07	0.000063	2.65	17961	1311	160	31
regulation of dendrite development	5.05E-07	6.69E-05	8.02	17961	140	160	10
regulation of cell cycle	5.3E-07	6.96E-05	2.83	17961	1111	160	28
peptidyl-amino acid modification	5.48E-07	7.14E-05	3.27	17961	790	160	23
cellular developmental process	5.53E-07	7.14E-05	2.06	17961	2558	160	47
cellular response to stimulus	6.22E-07	7.96E-05	2.03	17961	2652	160	48
regulation of growth	6.87E-07	8.73E-05	3.62	17961	621	160	20
biological regulation	6.95E-07	8.75E-05	1.28	17961	11406	160	130
peptidyl-lysine methylation	7.45E-07	9.31E-05	10.82	17961	83	160	8

nervous system process	7.94E-07	9.84E-05	3.1	17961	870	160	24
positive regulation of neurological system process	8.01E-07	9.85E-05	13.55	17961	58	160	7
regulation of signaling	8.29E-07	0.000101	1.88	17961	3284	160	55
modulation of excitatory postsynaptic potential	8.42E-07	0.000102	18.2	17961	37	160	6
regulation of dendrite morphogenesis	9.8E-07	0.000118	10.44	17961	86	160	8
positive regulation of dendritic spine development	9.92E-07	0.000118	17.72	17961	38	160	6
histone methylation	1.07E-06	0.000127	10.32	17961	87	160	8
positive regulation of calcium ion-dependent exocytosis	1.23E-06	0.000144	25.51	17961	22	160	5
cellular component morphogenesis	1.24E-06	0.000144	4.25	17961	423	160	16
regulation of dendritic spine morphogenesis	1.58E-06	0.000183	16.43	17961	41	160	6
regulation of cell growth	1.6E-06	0.000184	4.43	17961	380	160	15
protein methylation	1.91E-06	0.000217	8.02	17961	126	160	9
protein alkylation	1.91E-06	0.000216	8.02	17961	126	160	9
positive regulation of dendrite development	2.17E-06	0.000243	11.73	17961	67	160	7
acidic amino acid transport	2.43E-06	0.00027	15.31	17961	44	160	6

cellular response to external stimulus	2.56E-06	0.000283	4.93	17961	296	160	13
central nervous system development	2.64E-06	0.00029	7.71	17961	131	160	9
rhythmic process	2.75E-06	0.0003	5.35	17961	252	160	12
regulation of cellular component biogenesis	2.89E-06	0.000313	2.96	17961	871	160	23
glutamate secretion	3.63E-06	0.00039	20.79	17961	27	160	5
regulation of system process	3.89E-06	0.000415	3.68	17961	518	160	17
positive regulation of multicellular organismal process	4.03E-06	0.000427	2.31	17961	1602	160	33
protein metabolic process	4.15E-06	0.000436	1.75	17961	3728	160	58
positive regulation of cell differentiation	4.79E-06	0.000501	2.88	17961	898	160	23
cell communication	6.18E-06	0.000642	3.02	17961	781	160	21
macromolecule methylation	6.41E-06	0.000661	5.42	17961	228	160	11
protein phosphorylation	6.94E-06	0.000711	2.9	17961	852	160	22
regulation of mitotic cell cycle	7.23E-06	0.000736	3.36	17961	602	160	18
regulation of cell communication	7.45E-06	0.000753	1.8	17961	3247	160	52
nervous system development	8.38E-06	0.000842	4.79	17961	281	160	12

positive regulation of glutamate receptor signaling pathway	1.02E-05	0.00101	28.06	17961	16	160	4
positive regulation of cell growth	1.05E-05	0.00104	6.52	17961	155	160	9
positive regulation of growth	1.17E-05	0.00115	5.08	17961	243	160	11
regulation of biological quality	0.000013	0.00128	1.75	17961	3401	160	53
regulation of synapse organization	1.36E-05	0.00133	5.56	17961	202	160	10
system process	1.42E-05	0.00137	2.38	17961	1318	160	28
positive regulation of developmental growth	1.49E-05	0.00143	6.24	17961	162	160	9
RNA biosynthetic process	1.67E-05	0.00159	3.3	17961	579	160	17
histone H3-K9 modification	1.68E-05	0.0016	24.95	17961	18	160	4
regulation of postsynapse organization	1.81E-05	0.00171	8.54	17961	92	160	7
positive regulation of dendrite morphogenesis	1.82E-05	0.00171	15.17	17961	37	160	5
positive regulation of neurotransmitter transport	1.82E-05	0.0017	15.17	17961	37	160	5
regulation of anatomical structure morphogenesis	2.09E-05	0.00193	2.56	17961	1054	160	24
synapse organization	2.09E-05	0.00192	5.98	17961	169	160	9

regulation of cell cycle process	2.25E-05	0.00206	2.97	17961	719	160	19
regulation of synapse structural plasticity	2.37E-05	0.00215	48.11	17961	7	160	3
regulation of TOR signaling	2.39E-05	0.00216	8.19	17961	96	160	7
regulation of synaptic transmission, glutamatergic	2.43E-05	0.00219	10.36	17961	65	160	6
positive regulation of neurotransmitter secretion	2.63E-05	0.00235	22.45	17961	20	160	4
animal organ morphogenesis	2.82E-05	0.00251	3.49	17961	483	160	15
organonitrogen compound metabolic process	2.93E-05	0.00259	1.56	17961	4740	160	66
regulation of behavior	3.16E-05	0.00277	9.9	17961	68	160	6
regulation of ion transport	3.16E-05	0.00276	3.13	17961	609	160	17
regulation of calcium ion-dependent exocytosis	3.55E-05	0.00308	7.7	17961	102	160	7
positive regulation of AMPA receptor activity	3.76E-05	0.00325	42.1	17961	8	160	3
long-term synaptic potentiation	3.86E-05	0.00331	13.05	17961	43	160	5
response to abiotic stimulus	3.95E-05	0.00338	2.59	17961	954	160	22
positive regulation of cell morphogenesis involved in differentiation	4.28E-05	0.00363	6.28	17961	143	160	8

visual learning	4.32E-05	0.00365	12.76	17961	44	160	5
regulation of synaptic vesicle cycle	4.54E-05	0.00382	7.41	17961	106	160	7
regulation of membrane potential	4.68E-05	0.00391	3.75	17961	389	160	13
methylation	4.77E-05	0.00396	4.36	17961	283	160	11
negative regulation of TOR signaling	4.83E-05	0.00399	12.47	17961	45	160	5
positive regulation of regulated secretory pathway	4.83E-05	0.00397	12.47	17961	45	160	5
associative learning	5.11E-05	0.00418	9.1	17961	74	160	6
response to radiation	5.33E-05	0.00434	3.7	17961	394	160	13
cellular response to extracellular stimulus	5.35E-05	0.00433	4.74	17961	237	160	10
cellular localization	5.36E-05	0.00432	1.98	17961	1981	160	35
negative regulation of intracellular signal transduction	5.37E-05	0.0043	3.3	17961	511	160	15
response to extracellular stimulus	5.56E-05	0.00442	3.47	17961	453	160	14
negative regulation of cell size	0.000056	0.00444	37.42	17961	9	160	3
regulation of dendrite extension	5.61E-05	0.00442	18.71	17961	24	160	4
cellular process	5.78E-05	0.00453	1.16	17961	13588	160	141

regulation of megakaryocyte differentiation	5.98E-05	0.00466	11.94	17961	47	160	5
regulation of myeloid cell differentiation	6.41E-05	0.00497	5.18	17961	195	160	9
memory	6.46E-05	0.00499	7.02	17961	112	160	7
positive regulation of mitotic cell cycle	6.59E-05	0.00507	5.91	17961	152	160	8
phosphorylation	6.92E-05	0.00529	2.43	17961	1063	160	23
macromolecule metabolic process	7.11E-05	0.00541	1.46	17961	5781	160	75
regulation of neuronal synaptic plasticity	7.32E-05	0.00554	11.45	17961	49	160	5
postsynaptic density organization	7.95E-05	0.00599	33.68	17961	10	160	3
regulation of neurotransmitter levels	8.29E-05	0.00621	4.1	17961	301	160	11
tube formation	9.49E-05	0.00708	6.6	17961	119	160	7
nitrogen compound metabolic process	0.000095	0.00705	1.4	17961	6562	160	82
regulation of organelle organization	0.000099	0.00732	2.32	17961	1163	160	24
anatomical structure formation involved in morphogenesis	0.000101	0.00741	2.85	17961	669	160	17
cellular response to chemical stimulus	0.000105	0.00769	2.09	17961	1555	160	29

face morphogenesis	0.000121	0.00882	15.48	17961	29	160	4
plasma membrane bounded cell projection organization	0.000129	0.00938	3.04	17961	553	160	15
regulation of neurological system process	0.000136	0.00979	6.24	17961	126	160	7
positive regulation of synaptic transmission, glutamatergic	0.000139	0.00997	14.97	17961	30	160	4
eye photoreceptor cell development	0.000139	0.00992	14.97	17961	30	160	4
negative regulation of synaptic transmission	0.00014	0.00993	10.02	17961	56	160	5
regulation of neurotransmitter transport	0.000143	0.0101	6.19	17961	127	160	7
postsynaptic specialization organization	0.000144	0.0101	28.06	17961	12	160	3
cellular response to organic substance	0.000149	0.0104	2.21	17961	1271	160	25
regulation of axonogenesis	0.000156	0.0109	5.22	17961	172	160	8
positive regulation of developmental process	0.000158	0.011	2.2	17961	1276	160	25
negative regulation of cell communication	0.000164	0.0114	2.19	17961	1279	160	25
negative regulation of signaling	0.00017	0.0117	2.19	17961	1282	160	25
action potential	0.000179	0.0123	9.51	17961	59	160	5

Wnt signaling pathway	0.000181	0.0124	3.75	17961	329	160	11
adult behavior	0.00019	0.0129	5.91	17961	133	160	7
response to chemical	0.000198	0.0134	1.78	17961	2463	160	39
response to light stimulus	0.000199	0.0134	4.04	17961	278	160	10
photoreceptor cell development	0.000203	0.0136	13.61	17961	33	160	4
axon guidance	0.00021	0.014	4.43	17961	228	160	9
cell surface receptor signaling pathway involved in cell-cell signaling	0.000212	0.0141	3.69	17961	335	160	11
postsynapse organization	0.000228	0.0151	13.21	17961	34	160	4
neuron projection organization	0.000228	0.0151	13.21	17961	34	160	4
regulation of neurotransmitter secretion	0.00023	0.0151	6.94	17961	97	160	6
neuron projection guidance	0.000231	0.0151	4.37	17961	231	160	9
positive regulation of actin nucleation	0.000235	0.0153	24.05	17961	14	160	3
histone H3-K9 methylation	0.000235	0.0152	24.05	17961	14	160	3
N-terminal peptidyl-lysine acetylation	0.000235	0.0152	74.84	17961	3	160	2

postsynaptic density assembly	0.000235	0.0151	74.84	17961	3	160	2
positive regulation of synapse structural plasticity	0.000235	0.0151	74.84	17961	3	160	2
response to glycoside	0.000235	0.015	74.84	17961	3	160	2
regulation of vesicle docking	0.000235	0.0149	74.84	17961	3	160	2
histone H3-K4 methylation	0.000256	0.0162	12.83	17961	35	160	4
regulation of response to stimulus	0.000264	0.0167	1.57	17961	3792	160	53
cellular response to endogenous stimulus	0.000272	0.017	2.73	17961	659	160	16
regulation of regulated secretory pathway	0.000284	0.0177	5.53	17961	142	160	7
N-terminal protein amino acid acetylation	0.000292	0.0182	22.45	17961	15	160	3
response to organic substance	0.000297	0.0184	1.86	17961	1988	160	33
glial cell differentiation	0.000304	0.0188	8.5	17961	66	160	5
post-embryonic development	0.000304	0.0187	8.5	17961	66	160	5
neuron differentiation	0.000316	0.0194	4.19	17961	241	160	9
regulation of neuron migration	0.000319	0.0194	12.14	17961	37	160	4
sister chromatid cohesion	0.000319	0.0194	12.14	17961	37	160	4

negative regulation of cell differentiation	0.000337	0.0204	2.67	17961	672	160	16
regulation of molecular function	0.000346	0.0209	1.7	17961	2710	160	41
dendritic spine morphogenesis	0.000357	0.0214	21.05	17961	16	160	3
response to endogenous stimulus	0.00037	0.0221	2.33	17961	964	160	20
positive regulation of cell cycle	0.000373	0.0222	3.45	17961	358	160	11
regulation of exocytosis	0.000391	0.0232	4.56	17961	197	160	8
amino acid transport	0.000412	0.0243	6.24	17961	108	160	6
cellular metabolic process	0.000426	0.0251	1.33	17961	7152	160	85
ATP-dependent chromatin remodeling	0.000427	0.0251	7.91	17961	71	160	5
positive regulation of mitotic sister chromatid segregation	0.00043	0.0252	19.81	17961	17	160	3
vocalization behavior	0.00043	0.0251	19.81	17961	17	160	3
positive regulation of lamellipodium assembly	0.00043	0.025	19.81	17961	17	160	3
dendrite morphogenesis	0.000432	0.0249	11.23	17961	40	160	4
response to nutrient levels	0.000444	0.0255	3.16	17961	426	160	12

phosphate-containing compound metabolic process	0.000455	0.0261	1.87	17961	1864	160	31
regulation of synaptic vesicle exocytosis	0.000456	0.026	7.8	17961	72	160	5
regulation of transmembrane transport	0.000463	0.0263	2.97	17961	491	160	13
regulation of cytoskeleton organization	0.000463	0.0262	2.97	17961	491	160	13
regulation of mitotic nuclear division	0.000464	0.0262	5.1	17961	154	160	7
NMDA glutamate receptor clustering	0.000468	0.0263	56.13	17961	4	160	2
striatal medium spiny neuron differentiation	0.000468	0.0262	56.13	17961	4	160	2
rhythmic synaptic transmission	0.000468	0.0261	56.13	17961	4	160	2
cell surface receptor signaling pathway	0.000483	0.0269	1.84	17961	1956	160	32
dicarboxylic acid transport	0.000486	0.027	7.69	17961	73	160	5
chromosome organization	0.000492	0.0272	3.34	17961	370	160	11
DNA conformation change	0.000501	0.0276	5.04	17961	156	160	7
regulation of mitotic spindle assembly	0.000513	0.0282	18.71	17961	18	160	3
short-chain fatty acid metabolic process	0.000513	0.0281	18.71	17961	18	160	3
response to stimulus	0.000515	0.0281	1.45	17961	4882	160	63

regulation of interferon-beta production	0.000571	0.031	10.44	17961	43	160	4
phosphorus metabolic process	0.000576	0.0312	1.84	17961	1890	160	31
neural tube closure	0.000585	0.0315	7.39	17961	76	160	5
neuron migration	0.000603	0.0324	5.81	17961	116	160	6
primary metabolic process	0.000607	0.0325	1.33	17961	7003	160	83
positive regulation of exocytosis	0.000621	0.0331	7.29	17961	77	160	5
acid secretion	0.000621	0.033	7.29	17961	77	160	5
regulation of phosphorylation	0.000622	0.0329	1.97	17961	1479	160	26
regulation of DNA binding	0.00063	0.0333	5.76	17961	117	160	6
cellular response to nutrient levels	0.000636	0.0335	4.24	17961	212	160	8
regulation of synaptic vesicle transport	0.000659	0.0346	7.2	17961	78	160	5
regulation of signal transduction	0.000686	0.0359	1.63	17961	2893	160	42
tube closure	0.000698	0.0364	7.1	17961	79	160	5
positive regulation of dendritic spine morphogenesis	0.000708	0.0367	16.84	17961	20	160	3
response to oxygen-containing compound	0.000723	0.0374	2.06	17961	1251	160	23

biological_process	0.000761	0.0392	1.07	17961	16509	160	157
localization	0.000775	0.0398	1.47	17961	4351	160	57
neuronal dense core vesicle exocytosis	0.000775	0.0397	44.9	17961	5	160	2
presynaptic dense core vesicle exocytosis	0.000775	0.0396	44.9	17961	5	160	2
postsynaptic specialization assembly	0.000775	0.0394	44.9	17961	5	160	2
macromolecule localization	0.000786	0.0399	1.86	17961	1754	160	29
regulation of DNA-templated transcription, elongation	0.000803	0.0406	9.55	17961	47	160	4
T cell differentiation involved in immune response	0.00082	0.0413	16.04	17961	21	160	3
central nervous system neuron differentiation	0.000828	0.0416	6.84	17961	82	160	5
neurotransmitter transport	0.00093	0.0466	5.35	17961	126	160	6
regulation of transcription regulatory region DNA binding	0.000941	0.0469	9.16	17961	49	160	4
positive regulation of dendrite extension	0.000943	0.0469	15.31	17961	22	160	3
positive regulation of supramolecular fiber organization	0.000956	0.0474	4.52	17961	174	160	7
positive regulation of molecular function	0.000989	0.0489	1.88	17961	1610	160	27

Appendix F: Protein-Protein Interaction Network Results

Table F.1: Hub proteins from brain disorders after Matrix-GC. Hub proteins are designated as having a degree of connection of at least 1 more than the average level of connectivity of the corresponding size-matched control datasets.

Node	Degree	Name/Description	Disorder
CD2AP	9	Cas ligand with multiple SH3 domains	AD
ZCWPW1	9	Zinc finger CW-type and PWWP domain containing 1	AD
INPP5D	9	Phosphatidylinositol 3,4,5-trisphosphate 5-phosphatase 1	AD
CELF1	9	Embryo deadenylation element-binding protein homolog	AD
BIN1	9	Box-dependent myc-interacting protein 1	AD
RIN3	9	Ras interaction/interference protein 3	AD
ABCA7	9	ATP-binding cassette, sub-family A (ABC1), member 7	AD
SORL1	9	Low-density lipoprotein receptor relative with 11 ligand-binding repeats	AD
EPHA1	8	Erythropoietin-producing hepatoma receptor	AD
PTK2B	8	Calcium-regulated non-receptor proline-rich tyrosine kinase	AD
SLC6A3	22	Solute carrier family 6 (neurotransmitter transporter), member 3	ADHD
GRM5	21	Glutamate receptor, metabotropic 5	ADHD
GRIN2B	19	Glutamate receptor, ionotropic, N-methyl D-aspartate 2B	ADHD
NPY	17	Pro-neuropeptide Y	ADHD
NTRK2	16	Neurotrophic tyrosine kinase, receptor, type 2	ADHD
DRD4	16	Dopamine receptor D4	ADHD
SNAP25	13	Synaptosomal-associated 25 kDa protein	ADHD
HTR1B	12	5-hydroxytryptamine (serotonin) receptor 1B, G protein-coupled	ADHD
DRD5	12	D1beta dopamine receptor	ADHD

CNTNAP2	12	Contactin associated protein-like 2	ADHD
GRM8	11	Glutamate receptor, metabotropic 8	ADHD
DRD1	11	D(1A) dopamine receptor	ADHD
CNR1	11	Cannabinoid receptor 1 (brain)	ADHD
SYP	10	Major synaptic vesicle protein p38	ADHD
MAOB	10	Amine oxidase [flavin-containing] B	ADHD
GRM7	10	Glutamate receptor, metabotropic 7	ADHD
DISC1	10	Disrupted in schizophrenia 1 protein	ADHD
RBFOX1	9	RNA binding protein, fox-1 homolog (C. elegans) 1	ADHD
HTR2C	9	5-hydroxytryptamine (serotonin) receptor 2C, G protein-coupled	ADHD
DBH	9	Dopamine beta-hydroxylase (dopamine beta-monooxygenase)	ADHD
ADRA2C	9	Alpha-2 adrenergic receptor subtype C4	ADHD
PTPRD	8	Receptor-type tyrosine-protein phosphatase delta	ADHD
ADRA1B	8	Alpha-1B adrenergic receptor	ADHD
CTNND2	7	Catenin (cadherin-associated protein), delta 2	ADHD
AUTS2	7	Autism susceptibility gene 2 protein	ADHD
ASTN2	6	Astrotactin 2	ADHD
TACR3	5	Neurokinin B receptor	ADHD
PARK2	5	Parkin RBR E3 ubiquitin protein ligase	ADHD
CHRNA7	5	Cholinergic receptor, nicotinic, alpha 7 (neuronal)	ADHD
SYT2	4	Synaptotagmin II	ADHD
NTF3	4	Nerve growth factor 2	ADHD
LPHN3	4	Calcium-independent alpha-latrotoxin receptor 3	ADHD
FOXP2	4	Trinucleotide repeat-containing gene 10 protein	ADHD
DPP6	4	Dipeptidyl aminopeptidase-related protein	ADHD
DPH2	4	S-adenosyl-L-methionine:L-histidine 3-amino-3-carboxypropyltransferase 2	ADHD

CCDC24	4	Coiled-coil domain containing 24	ADHD
CAMK2D	4	Calcium/calmodulin-dependent protein kinase type II subunit delta	ADHD
B4GALT2	4	UDP-galactose:beta-N-acetylglucosamine beta-1,4-galactosyltransferase 2	ADHD
TRIM32	3	RING-type E3 ubiquitin transferase TRIM32	ADHD
NKAIN2	3	Sodium/potassium-transporting ATPase subunit beta-1-interacting protein 2	ADHD
KDM4A	3	JmjC domain-containing histone demethylation protein 3A	ADHD
IMMP2L	3	IMP2 inner mitochondrial membrane peptidase-like (S. cerevisiae)	ADHD
SV2C	2	Synaptic vesicle glycoprotein 2C	ADHD
SLC6A9	2	Solute carrier family 6 (neurotransmitter transporter, glycine), member 9	ADHD
NDE1	2	Nuclear distribution protein nudE homolog 1	ADHD
GFOD1	2	Glucose-fructose oxidoreductase domain containing 1	ADHD
CAMK2G	2	Calcium/calmodulin-dependent protein kinase type II subunit gamma	ADHD
BAIAP2	2	Brain-specific angiogenesis inhibitor 1-associated protein 2	ADHD
ATP2C2	2	ATPase, Ca++ transporting, type 2C, member 2	ADHD
PA2G4	2	Proliferation-associated protein 2G4	Anorexia
C8orf74	2	Chromosome 8 open reading frame 74	ASD
TRANK1	4	Tetratricopeptide repeat and ankyrin repeat containing 1	BIP
ANK3	3	Ankyrin 3, node of Ranvier (ankyrin G)	BIP
NCAN	3	Chondroitin sulfate proteoglycan 3	BIP
LMAN2L	2	Lectin, mannose-binding 2-like	BIP
ADCY2	2	Adenylate cyclase 2 (brain)	BIP
RIMS1	2	Regulating synaptic membrane exocytosis protein 1	BIP
SHANK2	2	SH3 and multiple ankyrin repeat domains protein 2	BIP
EP300	15	Protein propionyltransferase p300	MDD

SF3B1	8	Pre-mRNA-splicing factor SF3b 155 kDa subunit	MDD
RBFOX1	7	RNA binding protein, fox-1 homolog (C. elegans) 1	MDD
HIST1H2BN	7	Histone cluster 1, H2bn	MDD
HIST1H2BL	7	Histone cluster 1, H2bl	MDD
GRM5	7	Glutamate receptor, metabotropic 5	MDD
PAX6	6	Aniridia type II protein	MDD
KMT2A	6	Myeloid/lymphoid or mixed-lineage leukemia protein 1	MDD
ZKSCAN4	5	Zinc finger protein with KRAB and SCAN domains 4	MDD
PGBD1	5	PiggyBac transposable element-derived protein 1	MDD
ESR2	5	Nuclear receptor subfamily 3 group A member 2	MDD
ERBB4	5	V-erb-b2 avian erythroblastic leukemia viral oncogene homolog 4	MDD
ABT1	5	Activator of basal transcription 1	MDD
TCF4	4	Class B basic helix-loop-helix protein 19	MDD
PAX5	4	B-cell-specific transcription factor	MDD
ITPR3	4	Inositol 1,4,5-trisphosphate receptor, type 3	MDD
HSPA1A	4	Heat shock 70 kDa protein 1A	MDD
HMGN4	4	High mobility group nucleosomal binding domain 4	MDD
FCF1	4	rRNA-processing protein FCF1 homolog	MDD
CNTNAP5	4	Contactin associated protein-like 5	MDD
TYK2	8	Non-receptor tyrosine-protein kinase TYK2	MS
STAT3	8	Signal transducer and activator of transcription 3 (acute-phase response factor)	MS
MYC	8	V-myc avian myelocytomatosis viral oncogene homolog	MS
CLEC16A	8	C-type lectin domain family 16, member A	MS
CD40	8	Tumor necrosis factor receptor superfamily member 5	MS

IRF8	7	Interferon consensus sequence-binding protein	MS
IL7R	7	Interleukin-7 receptor subunit alpha	MS
TNFAIP3	6	Tumor necrosis factor, alpha-induced protein 3	MS
MAPK1	4	Extracellular signal-regulated kinase 2	MS
CD6	4	T-cell differentiation antigen CD6	MS
CD58	4	Lymphocyte function-associated antigen 3	MS
TCF7	3	Transcription factor 7 (T-cell specific, HMG-box)	MS
MMEL1	3	Membrane metallo-endopeptidase-like 1	MS
IL12A	3	Cytotoxic lymphocyte maturation factor 35 kDa subunit	MS
ZMIZ1	2	Zinc finger MIZ domain-containing protein 1	MS
OLIG3	2	Class E basic helix-loop-helix protein 20	MS
METTL1	2	tRNA (guanine(46)-N(7))-methyltransferase	MS
BACH2	2	BTB and CNC homology 1, basic leucine zipper transcription factor 2	MS
MAPT	8	Microtubule-associated protein tau	PD
MCCC1	7	Methylcrotonoyl-CoA carboxylase subunit alpha, mitochondrial	PD
TMEM175	5	Endosomal/lysosomal potassium channel TMEM175	PD
SPPL2C	5	Signal peptide peptidase like 2C	PD
KANSL1	5	KAT8 regulatory NSL complex subunit 1	PD
CRHR1	5	Corticotropin releasing hormone receptor 1	PD
STH	4	Saitohin	PD
GBA	4	D-glucosyl-N-acylsphingosine glucohydrolase	PD
DGKQ	4	Diacylglycerol kinase, theta 110kDa	PD
TMEM163	2	Transmembrane protein 163	PD
ITIH4	2	Inter-alpha-trypsin inhibitor family heavy chain-related protein	PD
FAM47E	2	Family with sequence similarity 47 member E	PD
ARHGAP27	2	CIN85-associated multi-domain-containing Rho GTPase-activating protein 1	PD
EP300	34	Protein propionyltransferase p300	SCZ

PPP1CC	20	Serine/threonine-protein phosphatase PP1-gamma catalytic subunit	SCZ
RELA	18	Nuclear factor of kappa light polypeptide gene enhancer in B-cells 3	SCZ
CDC20	18	Cell division cycle protein 20 homolog	SCZ
PSMA4	17	Proteasome (prosome, macropain) subunit, alpha type, 4	SCZ
CACNB2	17	Voltage-dependent L-type calcium channel subunit beta-2	SCZ
NHP2L1	16	NHP2 non-histone chromosome protein 2-like 1 (S. cerevisiae)	SCZ
PPP4C	15	Serine/threonine-protein phosphatase 4 catalytic subunit	SCZ
RBX1	14	Ring-box 1, E3 ubiquitin protein ligase	SCZ
SF3B1	13	Pre-mRNA-splicing factor SF3b 155 kDa subunit	SCZ
PPP2R2A	13	Serine/threonine-protein phosphatase 2A 55 kDa regulatory subunit B alpha isoform	SCZ
KAT5	13	Histone acetyltransferase HTATIP	SCZ
GRIN2A	13	Glutamate receptor, ionotropic, N-methyl D-aspartate 2A	SCZ
CNNM2	13	Cyclin and CBS domain divalent metal cation transport mediator 2	SCZ
CKAP5	13	Colonic and hepatic tumor overexpressed gene protein	SCZ
RANGAP1	12	Ran GTPase activating protein 1	SCZ
HSPA9	12	Heat shock 70kDa protein 9 (mortalin)	SCZ
GATAD2A	12	GATA zinc finger domain-containing protein 2A	SCZ
RPS17	11	Small ribosomal subunit protein eS17	SCZ
NT5C2	11	Cytosolic purine 5'-nucleotidase	SCZ
ETF1	11	Eukaryotic peptide chain release factor subunit 1	SCZ
SMG6	10	SMG6 nonsense mediated mRNA decay factor	SCZ
SLC32A1	10	Solute carrier family 32 (GABA vesicular transporter), member 1	SCZ
SDCCAG8	10	Centrosomal colon cancer autoantigen protein	SCZ

PRPF3	10	U4/U6 small nuclear ribonucleoprotein Prp3	SCZ
KDM4A	10	JmjC domain-containing histone demethylation protein 3A	SCZ
INO80E	10	Coiled-coil domain-containing protein 95	SCZ
F2	10	Coagulation factor II (thrombin)	SCZ
EGR1	10	Nerve growth factor-induced protein A	SCZ
DOC2A	10	Double C2-like domain-containing protein alpha	SCZ
DDB2	10	Damage-specific DNA binding protein 2, 48kDa	SCZ
ATF4	10	Tax-responsive enhancer element-binding protein 67	SCZ
ACTR5	10	ARP5 actin-related protein 5 homolog (yeast)	SCZ
ACTR1A	10	ARP1 actin-related protein 1 homolog A, centractin alpha (yeast)	SCZ
ZNF804A	9	Zinc finger protein 804A	SCZ
YJEFN3	9	YjeF N-terminal domain-containing protein 3	SCZ
XRCC6	9	X-ray repair complementing defective repair in Chinese hamster cells 6	SCZ
TCF4	9	Class B basic helix-loop-helix protein 19	SCZ
SUGP1	9	SURP and G-patch domain-containing protein 1	SCZ
RRAS	9	Related RAS viral (r-ras) oncogene homolog	SCZ
PPP1R3B	9	Hepatic glycogen-targeting protein phosphatase 1 regulatory subunit GL	SCZ
PHF5A	9	PHD finger-like domain-containing protein 5A	SCZ
NCK1	9	SH2/SH3 adaptor protein NCK-alpha	SCZ
MCL1	9	Induced myeloid leukemia cell differentiation protein Mcl-1	SCZ
ITIH3	9	Inter-alpha-trypsin inhibitor heavy chain H3	SCZ
HSPE1	9	10 kDa heat shock protein, mitochondrial	SCZ
HIRIP3	9	HIRA interacting protein 3	SCZ
GABBR2	9	Gamma-aminobutyric acid type B receptor subunit 2	SCZ
FGFR1	9	Basic fibroblast growth factor receptor 1	SCZ

CENPM	9	Proliferation-associated nuclear element protein 1	SCZ
CACNA1D	9	Calcium channel, voltage-dependent, L type, alpha 1D subunit	SCZ
ATPAF2	9	ATP synthase mitochondrial F1 complex assembly factor 2	SCZ
ALMS1	9	Alstrom syndrome protein 1	SCZ
TM6SF2	8	Transmembrane 6 superfamily member 2	SCZ
STAG1	8	Cohesin subunit SA-1	SCZ
SREBF1	8	Sterol regulatory element binding transcription factor 1	SCZ
SETD8	8	SET domain containing (lysine methyltransferase) 8	SCZ
PPP2R3A	8	Serine/threonine-protein phosphatase 2A regulatory subunit B" subunit alpha	SCZ
PLCB2	8	1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase beta-2	SCZ
PDCD11	8	Programmed cell death protein 11	SCZ
NDUFA6	8	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 6, 14kDa	SCZ
LSM1	8	U6 snRNA-associated Sm-like protein LSm1	SCZ
GLT8D1	8	Glycosyltransferase 8 domain containing 1	SCZ
FAM57B	8	Family with sequence similarity 57, member B	SCZ
CNOT1	8	Negative regulator of transcription subunit 1 homolog	SCZ
CDC25C	8	Dual specificity phosphatase Cdc25C	SCZ
ATP2A2	8	Calcium-transporting ATPase sarcoplasmic reticulum type, slow twitch skeletal muscle isoform	SCZ
ARPC3	8	Actin related protein 2/3 complex, subunit 3, 21kDa	SCZ
USMG5	7	Up-regulated during skeletal muscle growth 5 homolog (mouse)	SCZ
TNKS	7	Tankyrase, TRF1-interacting ankyrin-related ADP-ribose polymerase	SCZ
TAOK2	7	Thousand and one amino acid protein kinase 2	SCZ

PTK2B	7	Calcium-regulated non-receptor proline-rich tyrosine kinase	SCZ
PRKCD	7	Tyrosine-protein kinase PRKCD	SCZ
PDE4B	7	cAMP-specific 3',5'-cyclic phosphodiesterase 4B	SCZ
NSUN6	7	NOL1/NOP2/Sun and PUA domain-containing protein 1	SCZ
NDUFAF2	7	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex assembly factor 2	SCZ
NCAN	7	Chondroitin sulfate proteoglycan 3	SCZ
LPAR2	7	Lysophosphatidic acid receptor Edg-4	SCZ
ITIH4	7	Inter-alpha-trypsin inhibitor family heavy chain-related protein	SCZ
GDPD3	7	Glycerophosphodiester phosphodiesterase domain-containing protein 3	SCZ
FTSJ2	7	16S rRNA (uridine(1369)-2'-O)-methyltransferase	SCZ
ERCC8	7	Excision repair cross-complementation group 8	SCZ
EFTUD1	7	Elongation factor Tu GTP-binding domain-containing protein 1	SCZ
DEPDC1B	7	HBV X-transactivated gene 8 protein	SCZ
CSMD1	7	CUB and sushi domain-containing protein 1	SCZ
CENPL	7	Centromere protein L	SCZ
CALB2	7	29 kDa calbindin	SCZ
BCL11B	7	Radiation-induced tumor suppressor gene 1 protein	SCZ
WBP1L	6	Outcome predictor in acute leukemia 1	SCZ
TBX6	6	T-box transcription factor TBX6	SCZ
SETDB1	6	Histone-lysine N-methyltransferase SETDB1	SCZ
RWDD2A	6	RWD domain containing 2A	SCZ
RERE	6	Arginine-glutamic acid dipeptide repeats protein	SCZ
PACSIN3	6	Protein kinase C and casein kinase substrate in neurons protein 3	SCZ

OPCML	6	Opioid binding protein/cell adhesion molecule-like	SCZ
NEK4	6	Serine/threonine-protein kinase NRK2	SCZ
MRPS14	6	Mitochondrial ribosomal protein S14	SCZ
LRRC48	6	Leucine-rich repeat-containing protein 48	SCZ
KDM3B	6	JmjC domain-containing histone demethylation protein 2B	SCZ
IRF3	6	Interferon regulatory factor 3	SCZ
GID4	6	Vacuolar import and degradation protein 24 homolog	SCZ
FES	6	Feline sarcoma/Fujinami avian sarcoma oncogene homolog	SCZ
DRG2	6	Developmentally regulated GTP binding protein 2	SCZ
CUL3	6	Cullin 3	SCZ
CLP1	6	Cleavage and polyadenylation factor I subunit 1	SCZ
CILP2	6	Cartilage intermediate layer protein 2	SCZ
CHRNA5	6	Cholinergic receptor, nicotinic, alpha 5 (neuronal)	SCZ
CHRNA3	6	Cholinergic receptor, nicotinic, alpha 3 (neuronal)	SCZ
CHRM4	6	Muscarinic acetylcholine receptor M4	SCZ
BAG4	6	BAG family molecular chaperone regulator 4	SCZ
B9D1	6	B9 domain-containing protein 1	SCZ
ATP13A1	6	Manganese-transporting ATPase 13A1	SCZ
ASCL1	6	Achaete-scute family bHLH transcription factor 1	SCZ
AS3MT	6	S-adenosyl-L-methionine:arsenic(III) methyltransferase	SCZ
AP3B2	6	Clathrin assembly protein complex 3 beta-2 large chain	SCZ
ALDOA	6	Aldolase A, fructose-bisphosphate	SCZ
ZNF592	5	Zinc finger protein 592	SCZ
WHSC1L1	5	WHSC1-like 1 isoform 9 with methyltransferase activity to lysine	SCZ

WHAMM	5	WAS protein homolog associated with actin, golgi membranes and microtubules	SCZ
TOM1L2	5	Target of myb1-like 2 (chicken)	SCZ
TLR9	5	Toll-like receptor 9	SCZ
TKT	5	Transketolase	SCZ
TCTN1	5	Tectonic family member 1	SCZ
TCF20	5	Stromelysin-1 PDGF-responsive element-binding protein	SCZ
TAC3	5	Tachykinin 3	SCZ
SREBF2	5	Sterol regulatory element binding transcription factor 2	SCZ
SCAF1	5	Serine arginine-rich pre-mRNA splicing factor SR-A1	SCZ
SBNO1	5	Strawberry notch homolog 1	SCZ
SATB2	5	Special AT-rich sequence-binding protein 2	SCZ
RIMS1	5	Regulating synaptic membrane exocytosis protein 1	SCZ
RBFOX1	5	RNA binding protein, fox-1 homolog (C. elegans) 1	SCZ
PRR12	5	Proline-rich protein 12	SCZ
PITPNM2	5	Phosphatidylinositol transfer protein, membrane-associated 2	SCZ
PCGF6	5	Polycomb group RING finger protein 6	SCZ
PBRM1	5	BRG1-associated factor 180	SCZ
NUDT1	5	Nudix (nucleoside diphosphate linked moiety X)-type motif 1	SCZ
NDUFA13	5	Gene associated with retinoic and interferon-induced mortality 19 protein	SCZ
MYO15A	5	Unconventional myosin-15	SCZ
MAP3K11	5	Src-homology 3 domain-containing proline-rich kinase	SCZ
LCAT	5	Phosphatidylcholine-sterol acyltransferase	SCZ
GPN3	5	ATP-binding domain 1 family member C	SCZ
FAM53C	5	Family with sequence similarity 53 member C	SCZ
EDC4	5	Human enhancer of decapping large subunit	SCZ

DPH2	5	S-adenosyl-L-methionine:L-histidine 3-amino-3-carboxypropyltransferase 2	SCZ
CHRNA2	5	Cholinergic receptor, nicotinic, alpha 2 (neuronal)	SCZ
BCL2L12	5	Bcl-2-related proline-rich protein	SCZ
ARL5B	5	ADP-ribosylation factor-like protein 5B	SCZ
ARHGAP1	5	Rho-related small GTPase protein activator	SCZ
ACO2	5	Aconitate hydratase, mitochondrial	SCZ
EP300	45	Protein propionyltransferase p300	ASD-SFARI
CHD8	44	Chromodomain helicase DNA binding protein 8	ASD-SFARI
MECP2	31	Methyl CpG binding protein 2	ASD-SFARI
PTEN	30	Mutated in multiple advanced cancers 1	ASD-SFARI
KMT2C	28	Myeloid/lymphoid or mixed-lineage leukemia protein 3	ASD-SFARI
SIN3A	27	SIN3 transcription regulator family member A	ASD-SFARI
KDM6A	26	Ubiquitously-transcribed X chromosome tetratricopeptide repeat protein	ASD-SFARI
GRIN2B	26	Glutamate receptor, ionotropic, N-methyl D-aspartate 2B	ASD-SFARI
UBE3A	23	Human papillomavirus E6-associated protein	ASD-SFARI
SETD1B	23	Histone-lysine N-methyltransferase SETD1B	ASD-SFARI
NF1	23	Neurofibromatosis-related protein NF-1	ASD-SFARI
YY1	22	Transcriptional repressor protein YY1	ASD-SFARI
SMARCA2	22	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2	ASD-SFARI
HDAC4	22	Histone deacetylase 4	ASD-SFARI

KAT2B	21	Histone acetyltransferase KAT2B	ASD-SFARI
DNMT3A	21	DNA (cytosine-5-)-methyltransferase 3 alpha	ASD-SFARI
MTOR	20	FK506-binding protein 12-rapamycin complex-associated protein 1	ASD-SFARI
KMT2A	20	Myeloid/lymphoid or mixed-lineage leukemia protein 1	ASD-SFARI
FMR1	20	Fragile X mental retardation protein 1	ASD-SFARI
SHANK2	19	SH3 and multiple ankyrin repeat domains protein 2	ASD-SFARI
CNTNAP2	19	Contactin associated protein-like 2	ASD-SFARI
ASH1L	19	Ash1 (absent, small, or homeotic)-like (Drosophila)	ASD-SFARI
SYNGAP1	18	Ras/Rap GTPase-activating protein SynGAP	ASD-SFARI
SMARCC2	18	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 2	ASD-SFARI
PPP2CA	18	Serine/threonine-protein phosphatase 2A catalytic subunit alpha isoform	ASD-SFARI
POGZ	18	Pogo transposable element with ZNF domain	ASD-SFARI
PAX6	18	Aniridia type II protein	ASD-SFARI
NLGN3	18	Gliotactin homolog	ASD-SFARI
CHD7	17	Chromodomain helicase DNA binding protein 7	ASD-SFARI
CACNA1A	17	Calcium channel, voltage-dependent, P/Q type, alpha 1A subunit	ASD-SFARI
ARID1B	17	AT-rich interactive domain-containing protein 1B	ASD-SFARI
NSD1	16	Histone-lysine N-methyltransferase, H3 lysine-36 and H4 lysine-20 specific	ASD-SFARI
CDKL5	16	Serine/threonine-protein kinase 9	ASD-SFARI

CAMK2A	16	Calcium/calmodulin-dependent protein kinase type II subunit alpha	ASD-SFARI
TSC2	15	Tuberous sclerosis 2 protein	ASD-SFARI
STXBP1	15	Syntaxin binding protein 1	ASD-SFARI
CHD3	15	Chromodomain helicase DNA binding protein 3	ASD-SFARI
RELN	14	Reelin	ASD-SFARI
PTPN11	14	Protein tyrosine phosphatase, non-receptor type 11	ASD-SFARI
MBD5	14	Methyl-CpG-binding domain protein 5	ASD-SFARI
HUWE1	14	HECT, UBA and WWE domain containing 1, E3 ubiquitin protein ligase	ASD-SFARI
SMC3	13	Basement membrane-associated chondroitin proteoglycan	ASD-SFARI
SMC1A	13	Structural maintenance of chromosomes protein 1A	ASD-SFARI
SETD5	13	SET domain-containing protein 5	ASD-SFARI
NTRK2	13	Neurotrophic tyrosine kinase, receptor, type 2	ASD-SFARI
NIPBL	13	Nipped-B homolog (Drosophila)	ASD-SFARI
MED13L	13	Thyroid hormone receptor-associated protein complex 240 kDa component-like	ASD-SFARI
CUL3	13	Cullin 3	ASD-SFARI
SYT1	12	Synaptotagmin I	ASD-SFARI
SRCAP	12	Snf2-related CREBBP activator protein	ASD-SFARI
PHF21A	12	BRAF35-HDAC complex protein BHC80	ASD-SFARI
KAT6A	12	Runt-related transcription factor-binding protein 2	ASD-SFARI

EHMT1	12	Euchromatic histone-lysine N-methyltransferase 1	ASD-SFARI
BRAF	12	v-Raf murine sarcoma viral oncogene homolog B1	ASD-SFARI
TBL1XR1	11	Nuclear receptor corepressor/HDAC3 complex subunit TBLR1	ASD-SFARI
SLC6A1	11	Solute carrier family 6 (neurotransmitter transporter), member 1	ASD-SFARI
RLIM	11	Ring finger protein, LIM domain interacting	ASD-SFARI
RIMS1	11	Regulating synaptic membrane exocytosis protein 1	ASD-SFARI
LEO1	11	Leo1, Paf1/RNA polymerase II complex component, homolog (S. cerevisiae)	ASD-SFARI
CAMK2B	11	Calcium/calmodulin-dependent protein kinase type II subunit beta	ASD-SFARI
CACNA1D	11	Calcium channel, voltage-dependent, L type, alpha 1D subunit	ASD-SFARI
SUV420H1	10	Suppressor of variegation 4-20 homolog 1 (Drosophila)	ASD-SFARI
SLC1A2	10	Solute carrier family 1 (glial high affinity glutamate transporter), member 2	ASD-SFARI
SATB2	10	Special AT-rich sequence-binding protein 2	ASD-SFARI
PPP2R5D	10	Serine/threonine-protein phosphatase 2A 56 kDa regulatory subunit delta isoform	ASD-SFARI
TSC1	9	Tuberous sclerosis 1 protein	ASD-SFARI
MET	9	MET proto-oncogene, receptor tyrosine kinase	ASD-SFARI
HERC2	9	HECT and RLD domain containing E3 ubiquitin protein ligase 2	ASD-SFARI
FOXP1	9	Mac-1-regulated forkhead	ASD-SFARI
TNRC6B	8	Trinucleotide repeat-containing gene 6B protein	ASD-SFARI
RHEB	8	Ras homolog enriched in brain	ASD-SFARI

PTCHD1	8	Patched domain-containing protein 1	ASD-SFARI
PHF3	8	PHD finger protein 3	ASD-SFARI
PCDH19	8	Protocadherin 19	ASD-SFARI
FBXO11	8	Protein arginine N-methyltransferase 9	ASD-SFARI
CUX1	8	Homeobox protein cut-like 1	ASD-SFARI
CACNA2D3	8	Calcium channel, voltage-dependent, alpha 2/delta subunit 3	ASD-SFARI
ADNP	8	Activity-dependent neuroprotector homeobox protein	ASD-SFARI
UPF3B	7	UPF3 regulator of nonsense transcripts homolog B (yeast)	ASD-SFARI
TCF4	7	Class B basic helix-loop-helix protein 19	ASD-SFARI
TCF20	7	Stromelysin-1 PDGF-responsive element-binding protein	ASD-SFARI
RPS6KA3	7	Ribosomal protein S6 kinase, 90kDa, polypeptide 3	ASD-SFARI
PSMD12	7	Proteasome (prosome, macropain) 26S subunit, non-ATPase, 12	ASD-SFARI
NTNG1	7	Laminet-1	ASD-SFARI
MED13	7	Thyroid hormone receptor-associated protein complex 240 kDa component	ASD-SFARI
KATNAL2	7	Katanin p60 ATPase-containing subunit A-like 2	ASD-SFARI
TRAF7	6	TNF receptor-associated factor 7, E3 ubiquitin protein ligase	ASD-SFARI
STAG1	6	Cohesin subunit SA-1	ASD-SFARI
SLC9A6	6	Solute carrier family 9, subfamily A (NHE6, cation proton antiporter 6), member 6	ASD-SFARI
RERE	6	Arginine-glutamic acid dipeptide repeats protein	ASD-SFARI

HNRNPU	6	Heterogeneous nuclear ribonucleoprotein U (scaffold attachment factor A)	ASD-SFARI
DDX3X	6	DEAD (Asp-Glu-Ala-Asp) box helicase 3, X-linked	ASD-SFARI
WAC	5	WW domain-containing adapter protein with coiled-coil	ASD-SFARI
KCNB1	5	Potassium voltage-gated channel, Shab-related subfamily, member 1	ASD-SFARI
IQSEC2	5	IQ motif and SEC7 domain-containing protein 2	ASD-SFARI
GABBR2	5	Gamma-aminobutyric acid type B receptor subunit 2	ASD-SFARI
DEPDC5	5	DEP domain-containing protein 5	ASD-SFARI
CTNND2	5	Catenin (cadherin-associated protein), delta 2	ASD-SFARI
CNTN4	5	Brain-derived immunoglobulin superfamily protein 2	ASD-SFARI
CNKSR2	5	Connector enhancer of kinase suppressor of Ras 2	ASD-SFARI
CACNA1H	5	Calcium channel, voltage-dependent, T type, alpha 1H subunit	ASD-SFARI
BCL11A	5	Ecotropic viral integration site 9 protein homolog	ASD-SFARI
ARX	5	Aristaless related homeobox	ASD-SFARI
AHI1	5	Abelson helper integration site 1 protein homolog	ASD-SFARI
WDFY3	4	WD repeat and FYVE domain-containing protein 3	ASD-SFARI
UNC13A	4	Unc-13 homolog A (C. elegans)	ASD-SFARI
RALA	4	V-ral simian leukemia viral oncogene homolog A (ras related)	ASD-SFARI
PIK3R2	4	Phosphatidylinositol 3-kinase 85 kDa regulatory subunit beta	ASD-SFARI
NCKAP1	4	Membrane-associated protein HEM-2	ASD-SFARI

MEIS2	4	Meis1-related protein 1	ASD-SFARI
KIF5C	4	Kinesin heavy chain neuron-specific 2	ASD-SFARI
HOXA1	4	Homeobox protein Hox-1F	ASD-SFARI
CHD2	4	Chromodomain helicase DNA binding protein 2	ASD-SFARI
CUX2	4	Homeobox protein cut-like 2	ASD-SFARI
CNOT3	4	CCR4-NOT transcription complex, subunit 3	ASD-SFARI
CEP290	4	Bardet-Biedl syndrome 14 protein	ASD-SFARI
CDK13	4	Cholinesterase-related cell division controller	ASD-SFARI
ATP1A1	4	Sodium/potassium-transporting ATPase subunit alpha-1	ASD-SFARI
AP1S2	4	Clathrin assembly protein complex 1 sigma-1B small chain	ASD-SFARI

Appendix G: Network Control Analysis Results

Table G.1: Descriptive statistics of permutation analysis for protein-protein interactions networks. Using permutation analysis, we designate hub proteins as nodes with a degree of the average level of connectivity of the corresponding control datasets size plus 1. Brain disorder protein networks are significant if they are beyond the minimum and maximum range of the control datasets (n=20 per dataset).

Permutation Data						**BEFORE MATRIX-GC**				**AFTER MATRIX-GC**			
Number of genes	Mean of avg, degree	Min	Max	Avg. - min (lower)	Max-Avg. (upper)	Avg. Degree	#Genes	Disorder (#genes)	Sig.	Avg. Degree	#Genes	Disorder (#genes)	Sig.
10	0.33	0	1.5	0.33	1.18	17.62	21	AD (21)	$p \leq 0.05$	8.8	10	AD (10)	$p \leq 0.05$
20	0.6	0	1	0.6	0.4	0.95	21	Epilepsy (21)	$p \leq 0.05$	0.17	12	Epilepsy (12)	$p \geq 0.05$
50	1.25	1	2	0.246	0.75	0.93	29	ASD (29)	$p \geq 0.05$	0.42	20	ASD (20)	$p \geq 0.05$
100	1.49	1.14	2.14	0.346	0.65	1.31	30	BIP (30)	$p \geq 0.05$	0.95	21	BIP (21)	$p \geq 0.05$
200	2.08	1.47	2.63	0.612	0.55	5.74	52	MS (52)	$p \leq 0.05$	3.54	27	MS (27)	$p \leq 0.05$
500	3.71	2.99	4.21	0.722	0.5	2.21	69	PD (69)	$p \geq 0.05$	1.44	43	PD (43)	$p \geq 0.05$
750	4.89	4.14	5.46	0.746	0.57	8	105	ADHD (105)	$p \leq 0.05$	5.74	71	ADHD (71)	$p \leq 0.05$
						11.05	235	SFARI (235)	$p \leq 0.05$	8.56	183	SFARI (183)	$p \leq 0.05$
						2.6	268	MDD (268)	$p \geq 0.05$	1.65	167	MDD (167)	$p \geq 0.05$
						5.63	731	SCZ (731)	$p \leq 0.05$	4.13	455	SCZ (455)	$p \geq 0.05$

Appendix H: Tissue Expression Results

Table H.1: Significant results of genes expressed in different tissues. Using the Fisher's exact test, we report whether there is a statistically significant difference in the proportion of expressed genes that are bound by Matrix-GC and those that are expressed and not bound by Matrix-GC.

Disorder	Tissue	Matrix-GC	Expressed Matrix-GC	Non-Matrix-GC	Expressed Non-Matrix-GC	p-value
SCZ	brain	421	366	208	136	8.60E-10
SCZ	skin	421	327	208	120	3.71E-07
SCZ	digestive	421	372	208	149	5.01E-07
SCZ	fat	421	324	208	123	6.36E-06
SCZ	urinary	421	347	208	139	1.92E-05
SCZ	immune	421	370	208	155	3.59E-05
SCZ	lung	421	332	208	133	1.04E-04
SCZ	reproductive	421	386	208	170	4.97E-04
SCZ	heart	421	304	208	121	5.61E-04
MDD	brain	159	142	87	62	6.16E-04
Epilepsy	skin	12	10	9	2	9.20E-03
MDD	immune	159	140	87	65	1.15E-02
ASD-SFARI	immune	180	169	51	42	1.99E-02
BIP	reproductive	20	20	9	6	2.30E-02

Bibliography

1. Goldberg, A. D., Allis, C. D. & Bernstein, E. Epigenetics: A Landscape Takes Shape. *Cell* **128**, 635–638 (2007).
2. Ausió, J., de Paz, A. M. artíne. & Esteller, M. MeCP2: the long trip from a chromatin protein to neurological disorders. *Trends Mol. Med.* **20**, 487–498 (2014).
3. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
4. Iacono, W. G. Endophenotypes in psychiatric disease: Prospects and challenges. *Genome Med.* **10**, 10–12 (2018).
5. Anttila, V. *et al.* Analysis of shared heritability in common disorders of the brain. *Science* **360**, 289–313 (2018).
6. Churko, J. M., Mantalas, G. L., Snyder, M. P. & Wu, J. C. Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases. *Circulation Research* vol. 112 1613–1623 (2013).
7. Craig Venter, J. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
8. Chen, G., Ning, B. & Shi, T. Single-cell RNA-seq technologies and related computational data analysis. *Front. Genet.* **10**, (2019).
9. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
10. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
11. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).
12. Kuehner, J. N., Bruggeman, E. C., Wen, Z. & Yao, B. Epigenetic regulations in neuropsychiatric disorders. *Front. Genet.* **10**, 268 (2019).
13. Draghici, S. *et al.* A systems biology approach for pathway level analysis. *Genome Res.* **17**, 1537–1545 (2007).
14. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform. Biol. Insights* **14**, 7–9 (2020).

15. Horgan, R. P. & Kenny, L. C. 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. *Obstet. Gynaecol.* **13**, 189–195 (2011).
16. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2002).
17. Joyner, M. J. & Pedersen, B. K. Ten questions about systems biology. *J. Physiol.* **589**, 1017–1030 (2011).
18. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–504 (2003).
19. Joshi-Tope, G. *et al.* Reactome: A knowledgebase of biological pathways. *Nucleic Acids Res.* **33**, D428–D432 (2005).
20. Pico, A. R. *et al.* WikiPathways: Pathway Editing for the People. *PLoS Biol.* **6**, 1403–1407 (2008).
21. Pavlopoulos, G. A. *et al.* Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future. *Gigascience* **4**, 38 (2015).
22. Le Novere, N. Quantitative and logic modelling of molecular and gene networks. *Nat. Rev. Genet.* **16**, 146–158 (2015).
23. Miller, R. A. *et al.* Beyond Pathway Analysis: Identification of Active Subnetworks in Rett Syndrome. *Front. Genet.* **10**, 1–10 (2019).
24. Roeder, R. G. Transcriptional regulation and the role of diverse coactivators in animal cells. *FEBS Lett.* **579**, 909–915 (2005).
25. Guy, J., Cheval, H., Selfridge, J. & Bird, A. The Role of MeCP2 in the Brain. *Annu. Rev. Cell Dev. Biol.* **27**, 631–652 (2011).
26. Hertz, G. Z. & Stormo, G. D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563–577 (1999).
27. Vihinen, M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics* **13**, S2 (2012).
28. Makedonski, K., Abuhatzira, L., Kaufman, Y., Razin, A. & Shemer, R. MeCP2 deficiency in Rett syndrome causes epigenetic aberrations at the PWS/AS imprinting center that affects UBE3A expression. *Hum. Mol. Genet.* **14**, 1049–1058 (2005).

29. Coffee, B., Zhang, F., Ceman, S., Warren, S. T. & Reines, D. Histone Modifications Depict an Aberrantly Heterochromatinized FMR1 Gene in Fragile X Syndrome. *Am. J. Hum. Genet.* **71**, 923–932 (2002).
30. Wang, Y. *et al.* Repression of TSC1/TSC2 mediated by MeCP2 regulates human embryo lung fibroblast cell differentiation and proliferation. *Int. J. Biol. Macromol.* **96**, 578–588 (2017).
31. Li, H., Yamagata, T., Mori, M., Yasuhara, A. & Momoi, M. Y. Mutation analysis of methyl-CpG binding protein family genes in autistic patients. *Brain Dev.* **27**, 321–325 (2005).
32. Suter, B., Treadwell-Deering, D., Zoghbi, H. Y., Glaze, D. G. & Neul, J. L. Brief report: MECP2 Mutations in People without Rett Syndrome. *J. Autism Dev. Disord.* **44**, 703–711 (2014).
33. Chen, W. G. *et al.* Derepression of BDNF Transcription Involves Calcium-Dependent Phosphorylation of MeCP2. *Science* **302**, 885–889 (2003).
34. Srivastav, S., Walitza, S. & Grünblatt, E. Emerging role of miRNA in attention deficit hyperactivity disorder: a systematic review. *ADHD Atten. Deficit Hyperact. Disord.* **10**, 49–63 (2018).
35. Autry, A. E. & Monteggia, L. M. Brain-Derived Neurotrophic Factor and Neuropsychiatric Disorders. *Pharmacol. Rev.* **64**, 238–258 (2012).
36. Sullivan, P. F., Daly, M. J. & O'Donovan, M. Genetic architectures of psychiatric disorders: The emerging picture and its implications. *Nat. Rev. Genet.* **13**, 537–551 (2012).
37. Stahl, E. A. *et al.* Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat. Genet.* **51**, 793–803 (2019).
38. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
39. Abrahams, B. S. *et al.* SFARI Gene 2.0: A community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol. Autism* **4**, 10.1186/2040-2392-4–36 (2013).
40. Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).
41. Hayman, V. & Fernandez, T. V. Genetic insights into ADHD biology. *Front. Psychiatry* **9**, 10.3389/fpsyt.2018.00251 (2018).

42. Howard, D. M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat. Neurosci.* **22**, 343–352 (2019).
43. Duncan, L. *et al.* Significant Locus and Metabolic Genetic Correlations Revealed in Genome-Wide Association Study of Anorexia Nervosa. *Am. J. Psychiatry* **174**, 850–858 (2017).
44. Chang, D. *et al.* A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat. Genet.* **49**, 1511–1516 (2017).
45. Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
46. Moss, D. J. H. *et al.* Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. *Lancet Neurol.* **16**, 701–711 (2017).
47. Bashinskaya, V. V., Kulakova, O. G., Boyko, A. N., Favorov, A. V. & Favorova, O. O. A review of genome-wide association studies for multiple sclerosis: classical and hypothesis-driven approaches. *Hum. Genet.* **134**, 1143–1162 (2015).
48. Rheenen, W. van *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1043–1048 (2016).
49. Abou-Khalil, B., Auce, P., Avbersek, A., Bahlo, M. & Balding, D. J. Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. *Nat. Commun.* **9**, 10.1038/s41467-018-07524-z (2018).
50. Bach, S. *et al.* Methyl - CpG - binding protein 2 mediates overlapping mechanisms across brain disorders. *Sci. Rep.* **10**, 1–13 (2020).
51. Klose, R. J. *et al.* DNA binding selectivity of MeCP2 due to a requirement for A/T sequences adjacent to methyl-CpG. *Mol. Cell* **19**, 667–678 (2005).
52. Stormo, G. D., Schneider, T. D., Gold, L. & Ehrenfeucht, A. Use of the 'Perceptron' algorithm to distinguish transational initiation sites in E. coli. *Nucleic Acids Res.* **10**, 1–15 (1982).
53. Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the

identification of regulatory elements. *Nat. Rev. Genet.* **5**, 276–287 (2004).

54. Xia, X. PositionWeightMatrix, Gibbs Sampler, and the Associated Signifcance Tests in Motif Characterization and Prediction. *Scientifica* **2012**, 1–15 (2012).
55. Maunakea, A. K., Chepelev, I., Cui, K. & Zhao, K. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Res.* **23**, 1256–1269 (2013).
56. Wang, S. *et al.* Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nat. Protoc.* **8**, 2502–2515 (2013).
57. Rube, H. T. *et al.* Sequence features accurately predict genome-wide MeCP2 binding in vivo. *Nat. Commun.* **7**, 10.1038/ncomms11025 (2016).
58. Kim, H. G. *et al.* Disruption of Neurexin 1 Associated with Autism Spectrum Disorder. *Am. J. Hum. Genet.* **82**, 199–207 (2008).
59. Singh, T. *et al.* Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat. Neurosci.* **19**, 571–577 (2016).
60. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362–D368 (2017).
61. Doncheva, N. T., Morris, J. H., Gorodkin, J. & Jensen, L. J. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. *J. Proteome Res.* **18**, 623–632 (2019).
62. Lips, E. S. *et al.* Functional gene group analysis identifies synaptic gene groups as risk factor for schizophrenia. *Mol. Psychiatry* **17**, 996–1006 (2012).
63. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. ClusterProfiler: An R package for comparing biological themes among gene clusters. *Omi. A J. Integr. Biol.* **16**, 284–287 (2012).
64. Yu, G. & He, Q. Y. ReactomePA: An R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.* **12**, 477–479 (2016).
65. Fisher, R. A. *Statistical Methods for Research Workers*. (Oliver and Boyd, 1925). doi:10.2307/2528855.
66. Fisher, R. A. On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. *J. R. Stat. Soc.* **85**, 87–94 (1922).

67. Sanfeliu, A., Hokamp, K., Gill, M. & Tropea, D. Transcriptomic analysis of Mecp2 mutant mice reveals differentially expressed genes and altered mechanisms in both blood and brain. *Front. Psychiatry* **10**, 10.3389/fpsyt.2019.00278 (2019).
68. Rodrigues, D. C. *et al.* Shifts in Ribosome Engagement Impact Key Gene Sets in Neurodevelopment and Ubiquitination in Rett Syndrome. *Cell Rep.* **30**, 4179–4196 (2020).
69. Chahrour, M. *et al.* MeCP2, a key contributor to neurological disease, activates and represses transcription. *Science* **320**, 1224–1229 (2008).
70. Urdinguio, R. G. *et al.* Mecp2-null mice provide new neuronal targets for rett syndrome. *PLoS One* **3**, (2008).
71. Zweier, M. *et al.* Mutations in MEF2C from the 5q14.3q15 microdeletion syndrome region are a frequent cause of severe mental retardation and diminish MECP2 and CDKL5 expression. *Hum. Mutat.* **31**, 722–733 (2010).
72. Chao, H. T., Zoghbi, H. Y. & Rosenmund, C. MeCP2 Controls Excitatory Synaptic Strength by Regulating Glutamatergic Synapse Number. *Neuron* **56**, 58–65 (2007).
73. Na, E. S., Nelson, E. D., Kavalali, E. T. & Monteggia, L. M. The impact of MeCP2 loss- or gain-of-function on synaptic plasticity. *Neuropsychopharmacology* **38**, 212–9 (2013).
74. Shovlin, S. & Tropea, D. Transcriptome level analysis in Rett syndrome using human samples from different tissues. *Orphanet J. Rare Dis.* **13**, (2018).
75. Liu, Z. *et al.* Autism-like behaviours and germline transmission in transgenic monkeys overexpressing MeCP2. *Nature* **530**, 98–102 (2016).
76. Peters, S. U. *et al.* The behavioral phenotype in MECP2 duplication syndrome: A comparison with idiopathic autism. *Autism Res.* **6**, 42–50 (2013).
77. Wong, D. F. *et al.* Are dopamine receptor and transporter changes in Rett syndrome reflected in Mecp2-deficient mice? *Exp. Neurol.* **307**, 74–81 (2018).
78. Soreq, H. & Wolf, Y. NeurimmiRs: MicroRNAs in the neuroimmune interface. *Trends Mol. Med.* **17**, 548–555 (2011).

79. Visvanathan, J., Lee, S., Lee, B., Lee, J. W. & Lee, S. K. The microRNA miR-124 antagonizes the anti-neural REST/SCP1 pathway during embryonic CNS development. *Genes Dev.* **21**, 744–749 (2007).
80. Yin, J. *et al.* MiR-137: A new player in schizophrenia. *Int. J. Mol. Sci.* **15**, 3262–3271 (2014).
81. Zoghbi, H. Y. *et al.* Cerebrospinal Fluid Biogenic Amines and Biopterin in Rett Syndrome. *Ann. Neurol.* **25**, 56–60 (1989).
82. Lekman, A. *et al.* Rett syndrome: Biogenic amines and metabolites in postmortem brain. *Pediatr. Neurol.* **5**, 357–362 (1989).
83. Wolfe, N. *et al.* Neuropsychological profile linked to low dopamine: In Alzheimer's disease, major depression, and Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry* **53**, 915–917 (1990).
84. Dobryakova, E., Genova, H. M., DeLuca, J. & Wylie, G. R. The dopamine imbalance hypothesis of fatigue in multiple sclerosis and other neurological disorders. *Front. Neurol.* **6**, 10.3389/fneur.2015.00052 (2015).
85. Volkow, N. D. *et al.* Evaluating dopamine reward pathway in ADHD: Clinical Implications. *J. Am. Med. Assoc.* **302**, 1084–1091 (2009).
86. Money, K. M. & Stanwood, G. D. Developmental origins of brain disorders: roles for dopamine. *Front. Cell. Neurosci.* **7**, 10.3389/fncel.2013.00260 (2013).
87. Lappalainen, R. & Riikonen, R. S. High levels of cerebrospinal fluid glutamate in Rett syndrome. *Pediatr. Neurol.* **15**, 213–216 (1996).
88. Maezawa, I. & Jin, L.-W. W. Rett Syndrome Microglia Damage Dendrites and Synapses by the Elevated Release of Glutamate. *J. Neurosci.* **30**, 5346–5356 (2010).
89. Meng, X. *et al.* Manipulations of MeCP2 in glutamatergic neurons highlight their contributions to Rett and other neurological disorders. *Elife* **5**, 10.7554/eLife.14199 (2016).
90. Kerr, D., Krishnan, C., Pucak, M. L. & Carmen, J. The immune system and neuropsychiatric diseases. *Int. Rev. Psychiatry* **17**, 443–449 (2005).
91. Khandaker, G. M. *et al.* Inflammation and immunity in schizophrenia: Implications for pathophysiology and treatment. *The Lancet Psychiatry* **2**, 258–270 (2015).

92. Delgado, I. J., Kim, D. S., Thatcher, K. N., LaSalle, J. M. & Van den Veyver, I. B. Expression profiling of clonal lymphocyte cell cultures from Rett syndrome patients. *BMC Med. Genet.* **7**, 1–13 (2006).
93. Kinney, J. W. *et al.* Inflammation as a central mechanism in Alzheimer's disease. *Alzheimer's Dement. Transl. Res. Clin. Interv.* **4**, 575–590 (2018).
94. De Felice, C. *et al.* Rett syndrome: An autoimmune disease? *Autoimmun. Rev.* **15**, 411–416 (2016).
95. Baker, D. *et al.* Endocannabinoids control spasticity in a multiple sclerosis model. *FASEB J.* **15**, 300–302 (2001).
96. Vigli, D. *et al.* Chronic treatment with the phytocannabinoid Cannabidivarin (CBDV) rescues behavioural alterations and brain atrophy in a mouse model of Rett syndrome. *Neuropharmacology* **140**, 121–129 (2018).
97. Wen, A. Y., Sakamoto, K. M. & Miller, L. S. The Role of the Transcription Factor CREB in Immune Function. *J. Immunol.* **185**, 6413–6419 (2010).
98. Ghisletti, S. *et al.* Cooperative NCoR/SMRT interactions establish a eorepressor-based strategy for integration of inflammatory ana anti-inflammatory signaling pathways. *Genes Dev.* **23**, 681–693 (2009).
99. Lengauer, C. & Issa, J.-P. The role of epigenetics in cancer. *Mol. Med. Today* **4**, 102–103 (1998).
100. Paciorkowski, A. R. *et al.* MEF2C Haploinsufficiency features consistent hyperkinesis, variable epilepsy, and has a role in dorsal and ventral neuronal developmental pathways. *Neurogenetics* **14**, 99–111 (2013).
101. Ghosh, A., Carnahan, J. & Greenberg, M. E. Requirement for BDNF in Activity-Dependent Survival of Cortical Neurons. *Science* **263**, 1618–23 (1994).
102. Cardoza, B. *et al.* Epilepsy in Rett syndrome: Association between phenotype and genotype, and implications for practice. *Seizure* **20**, 646–649 (2011).
103. Operto, F. F., Mazza, R., Pastorino, G. M. G., Verrotti, A. & Coppola, G. Epilepsy and genetic in Rett syndrome: A review. *Brain Behav.* **9**, 10.1002/brb3.1250 (2019).
104. McCarthy, S. E. *et al.* De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and

intellectual disability. *Mol. Psychiatry* **19**, 652–658 (2014).

105. Nagarajan, R. P., Hogart, A. R., Gwye, Y., Martin, M. R. & LaSalle, J. M. Reduced MeCP2 expression is frequent in autism frontal cortex and correlates with aberrant MECP2 promoter methylation. *Epigenetics* **1**, 172–182 (2006).
106. Xu, Y. *et al.* Multiple epigenetic factors predict the attention deficit/hyperactivity disorder among the Chinese Han children. *J. Psychiatr. Res.* **64**, 40–50 (2015).
107. Fornes, O. *et al.* JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).
108. Wingender, E., Dietze, P., Karas, H. & Knüppel, R. TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* **24**, 238–241 (1996).
109. Crawford, G. E. *et al.* Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* **16**, 123–131 (2006).
110. Jones, P. L. *et al.* Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat. Genet.* **19**, 187–191 (2002).
111. Placzek, W. J. *et al.* Identification of a novel Mcl-1 protein binding motif. *J. Biol. Chem.* **286**, 39829–39835 (2011).
112. Fantini, J. & Barrantes, F. J. How cholesterol interacts with membrane proteins: An exploration of cholesterol-binding sites including CRAC, CARC, and tilted domains. *Front. Physiol.* **4**, (2013).
113. Glass, K. & Girvan, M. Annotation enrichment analysis: An alternative method for evaluating the functional properties of gene sets. *Sci. Rep.* **4**, (2014).
114. Haynes, W. A., Tomczak, A. & Khatri, P. Gene annotation bias impedes biomedical research. *Sci. Rep.* **8**, (2018).
115. Gillis, J. & Pavlidis, P. Assessing identity, redundancy and confounds in Gene Ontology annotations over time. *Bioinformatics* **29**, 476–482 (2013).
116. Barriot, R., Sherman, D. J. & Dutour, I. How to decide which are the most pertinent overly-represented features during gene set enrichment analysis.

BMC Bioinformatics **8**, 332 (2007).

117. Stoney, R. A., Schwartz, J. M., Robertson, D. L. & Nenadic, G. Using set theory to reduce redundancy in pathway sets. *BMC Bioinformatics* **19**, 1–11 (2018).
118. Wang, G., Oh, D. H. & Dassanayake, M. GOMCL: A toolkit to cluster, evaluate, and extract non-redundant associations of Gene Ontology-based functions. *BMC Bioinformatics* **21**, 1–9 (2020).
119. Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics* **13**, 397–406 (2014).
120. Feder, M. E. & Walser, J.-C. C. The biological limitations of transcriptomics in elucidating stress and stress responses. *J. Evol. Biol.* **18**, 901–910 (2005).
121. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).