

THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e. g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Network analysis of the Post Synaptic proteome and its implication for cognition

Grant Robertson



Doctor of Philosophy
Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh
2022

Abstract

The post synaptic proteome is a complex of more than 3,000 proteins that form modular molecular machines that are involved in signal transduction, information processing and learning in the central nervous system. Previous analyses have used techniques from network science such as community detection to show differential enrichment of network modules using over-representation analysis for functions and disorders.

In this thesis I extend these techniques by combining network analysis including community detection with Gene Set Analysis using the full polygenic signal available from summary GWA studies. Using large scale population discovery and replication samples I show that distinct communities in the post synaptic proteome have an enriched association for genes linked to human cognitive ability

No evidence is found for an effect of network centrality on the propensity of genes to have variants associated with differences in intelligence but these are found to be associated with increased loss of function intolerance.

This thesis describes a principled method in which network analysis of protein interactions can be used to understand human population genetic studies and could be extended to neuropsychiatric disorders affecting the synapse

Lay Summary The brain is made up of nerve cells that are connected at special junctions called synapses. The pattern of interconnection of brain cells is responsible for many of the brains remarkable features. The synapse transmits information from one nerve cell to another but is also involved in learning. The synapse comprises thousands of proteins that form molecular machines to transmit information and process it. In recent years we have found that the synapse is much more complex than we previously anticipated and comprises thousands of proteins. The pattern of interconnections is so intricate that we cannot look at a synapse diagram and identify groups of proteins that seem to work together or identify critical proteins. One way to understand complex structures is as a network where components (in this case, proteins) are represented by nodes connected by lines (edges) if they interact. We can use network science to understand these connections and find significant nodes in the network and tightly connected groups of proteins that share a common function using clustering methods.

This thesis uses the most up to date network map of the synapse available developed by the Armstrong and Simpson groups in the School of Informatics to better understand genetic differences in human cognitive ability. Community based clustering methods

prove useful in finding tightly connected groups of proteins within the network of proteins that are particularly associated with differences in human cognitive ability.

These techniques could be extended and may be helpful in finding groups of proteins that are involved in diseases affecting the synapse, such as Alzheimer's disease or depression, and may provide the target for future treatments.

Acknowledgements

I would first like to express my gratitude to the Medical Research Foundation and the Medical Research Council for funding the PsySTAR fellowship that enabled me to carry out the work in this thesis. PsySTAR provides postgraduate training for psychiatrists and is jointly run by the Universities of Aberdeen, Dundee, Edinburgh and Glasgow. All of the materials used in this thesis were funded by PsySTAR. I am grateful to all of the members of the PsySTAR board who provided me with advice and support. I am particularly grateful to Prof Stephen Lawrie, the director of PsySTAR who helped me to refine the objectives and design of the work presented in this thesis.

The PsySTAR fellowship afforded me the opportunity to choose both the subject of my PhD and my supervisors. I have been incredibly fortunate to have had Prof. J. Douglas Armstrong and Prof. Ian Deary as supervisors. Prof Deary is one of the founders of the field of cognitive epidemiology and provided invaluable advice about a complex field and a great deal of practical help including access to resources both in the Centre for Cognitive Epidemiology and Cognitive Ageing and more widely.

I would like to express my particular gratitude to Dr W. David Hill who carried out the GWA studies of intelligence and educational attainment that form the discovery samples in this thesis (section B.5). In addition he was a source of continued advice during the completion of the studies in this thesis particularly on the subject of Gene Set Analysis. Dr Gail Davies provided practical help and advice in the field of the genetics of intelligence as did Dr Riccardo Marioni.

I would like to thank Prof Armstrong for his guidance throughout this thesis. Prof Armstrong and Dr Ian Simpsons group were responsible for the post synaptic proteome network model used throughout this thesis. Dr Oksana Sorokina, Dr David Sterrat, Dr Katharina Heil, Dr Emilia Wysocka and Dr Colin McLean worked on the post synaptic proteome (section B.5) and were a source of advice throughout the completion of this thesis. Dr Xin He gave me advice about Gene Ontology. Dr Colin McLean provided the implementation of spectral community detection used in this thesis and carried out some of the community detection. He was a source of great advice, encouragement and practical help throughout this thesis.

Dr Aysu Okbay was kind enough to prepare and provide the summary GWAS data for the Educational Attianment Replication sample.

Dr J. Mitchell of Napier University read early drafts of this thesis and provided useful advice. Dr Keith Smith of the Usher Institute provided helpful advice.

I'd like to thank my wife, Catriona, for proof-reading this thesis and for her support during this thesis. My mother, Patricia, also was a source of much help, particularly with our son Jamie. I'd like to thank her and my late father Alan for their help over the years.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Grant Robertson*)

Table of Contents

1 Introduction **1**

1.1 The synaptic proteome 1

1.1.1 Synapse introduction 1

1.1.2 Molecular genetic revolution 3

1.2 Synapse structure 4

1.2.1 Presynaptic area structure 5

1.2.2 The postsynaptic area and postsynaptic density 5

1.2.3 Molecular function 7

1.2.4 The synaptomic theory 9

1.2.5 Role in disease 9

1.2.6 Evolution 10

1.2.7 Importance 11

1.2.8 Systems Biology approaches to the synapse 12

1.2.9 Models of the synapse 12

1.2.10 Systems biology of the synapse and networks 14

1.2.11 A note on terminology regarding the Synaptic Proteome network 15

1.3 Intelligence 15

1.3.1 Measures and heritability 15

1.3.2 Cognitive epidemiology 17

1.3.3 The molecular basis of cognition. 18

1.3.4 Genome scale molecular findings 18

1.3.5 Educational attainment 20

1.3.6 Relevance to the systems neuroscience of the synapse 21

1.4 Gene level tests 22

1.4.1 Generating gene level scores from SNP level scores 23

1.4.2 VEGAS 24

1.4.3 MAGMA 25
1.4.4 Pascal 26
1.5 Gene set tests 27
1.5.1 Gene Set Enrichment Analysis 28
1.5.2 GSA with MAGMA 29
1.5.3 GSA with Pascal 29
1.6 Using network topology 30
1.7 Plan 33

2 Network graph and GWAS samples **35**
2.1 Network Graph of the Post Synaptic Proteome 35
2.1.1 Documenting interactions 40
2.1.2 Connected component 43
2.1.3 Gene ontology analysis 46
2.1.4 Gene ontology enrichment methods 47
2.1.5 GTEx profile of PSP genes 52
2.2 Genome Wide Association Studies 55
2.2.1 Population genetic samples 56
2.2.2 Samples 56
2.2.3 MAGMA Gene level statistics: Validation 57
2.2.4 Synaptic genes absent from NCBI 37 58
2.3 MAGMA gene level statistic methods 60
2.3.1 Correlation of MAGMA GWGAS between samples 61
2.3.2 Intelligence Replication GWGAS 62
2.3.3 Education Replication MAGMA GWGAS 64
2.3.4 Intelligence Discovery Phase II GWGAS 65
2.3.5 Education Discovery sample GWGAS 66
2.3.6 Pascal and MAGMA 66
2.4 Gene Ontology Analysis of significant GWGAS genes 70
2.4.1 Intelligence Replication GO analysis 71
2.4.2 Education Replication Gene Ontology analysis 72
2.4.3 PANTHER Gene Ontology analysis Education Replication . . 72
2.4.4 Intelligence Discovery: Gene Ontology enrichment 79
2.4.5 Education Discovery Ontology Over representation analysis . 86
2.5 Synaptic component enrichment MAGMA 105

2.6 Summary 108

3 Do centrality measures in the post synaptic proteome correlate with function? 111

3.1 Properties of 'real world' networks 113
3.1.1 Small world property 113
3.1.2 The configuration model 114
3.2 Centrality measures 115
3.2.1 Degree centrality 117
3.2.2 Degree distribution 118
3.2.3 Scale free 120
3.2.4 Betweenness centrality 121
3.2.5 Random walk betweenness 122
3.2.6 Closeness centrality 122
3.2.7 Eigenvector centrality 122
3.2.8 Transitivity 124
3.2.9 kcores 127
3.3 Assortativity and assortativity of measures 129
3.3.1 Degree assortativity 130
3.4 Centrality and essentialness 130
3.5 Biomedical significance of centrality measures 132
3.6 Methods 134
3.6.1 Calculation of centrality measures 134
3.6.2 Calculating centralities 134
3.6.3 Correlation between centralities and measures of genetic significance 135
3.6.4 Characterising the degree distribution 136
3.6.5 Calculating small world property 137
3.6.6 Calculating assortativity 139
3.6.7 Gene Ontology Analysis of extreme centrality genes 139
3.6.8 Disease term enrichment - DisGenNet 141
3.6.9 Methods Murine LTP and centrality 141
3.6.10 Centrality and Essential Genes 142
3.6.11 Loss of function and evolutionary pressure ExAC and gnomad 145
3.6.12 Cohorts and samples 147

3.7 Results Summary of centrality measures . . . 148
3.7.1 Testing the normality of centrality measures . . . 148
3.8 Correlation of centrality measures . . . 151
3.8.1 Correlation of centrality measures . . . 151
3.9 Global network measures . . . 151
3.9.1 Results Degree distribution . . . 151
3.9.2 Small world property . . . 152
3.9.3 Assortativity . . . 154
3.9.4 Centrality and murine LTP . . . 158
3.10 Characterisation of central genes . . . 165
3.11 Essentialness and centrality . . . 169
3.12 Probability of loss of function intolerance (pLI) . . . 171
3.12.1 Distribution of pLI . . . 171
3.13 Disease enrichment for centrality measures . . . 175
3.13.1 Disease enrichment . . . 175
3.13.2 Distribution of pLI and intelligence . . . 178
3.13.3 Correlation centrality measures and study p . . . 181
3.13.4 Summary . . . 181
3.14 Summary of chapter . . . 183

4 Assessment and selection of clustering algorithms 185
4.1 Network communities . . . 187
4.1.1 Modularity . . . 188
4.2 Community Detection . . . 189
4.2.1 Assessment of community detection . . . 189
4.3 Choice of community detection algorithm . . . 192
4.3.1 Girvan–Newman algorithm . . . 193
4.3.2 Spectral clustering . . . 194
4.3.3 Louvain . . . 194
4.3.4 Infomap . . . 195
4.3.5 Spinglass . . . 195
4.3.6 Walktrap . . . 196
4.3.7 Fast greedy . . . 196
4.3.8 Label propogation . . . 196
4.3.9 Markov Cluster Algorithm . . . 197

4.3.10 Leading eigenvector community detection - LEC 197
4.3.11 Stochastic block model 197
4.4 Lancichinetti–Fortunato–Radicchi (LFR) benchmark 197
4.4.1 Benchmark: Stochastic block model Modularity and Group size 208

5 Topology based Gene Set Analysis 217
5.1 Topology based GSA 217
5.1.1 Study design Topology based GSA 217
5.1.2 Gene ontology enrichment 219
5.1.3 Results 220
5.1.4 Gene ontology analysis 230
5.1.5 Murine and Disease Phenotypes in Community 5 230
5.1.6 Subgroup analysis 231
5.1.7 Genetic Constraint 234
5.1.8 Centrality measures for community 5 240
5.1.9 Louvain MAGMA GSA 240
5.1.10 Summary 245
5.2 Replication in larger samples: Intelligence 2 and EA3 246
5.2.1 Intelligence 246
5.2.2 Education 246

6 Discussion 251
6.1 Principal findings 251
6.1.1 Intelligence and Educational Ability 251
6.1.2 Vertex statistic and intelligence 252
6.2 GSA and Communities 254
6.3 Study design 258
6.3.1 Benchmarks 258

A Abbreviations 259

B Centrality measures - Supplemental 263
B.1 Correlation of centrality measures 263
B.1.1 Supplemental degree distribution 268
B.1.2 Global measures: Relationship of $C(k)$ to k 271
B.1.3 Results Small world 273

B.2 Chapter 2 . . . 280
B.3 Chapter Community Detection . . . 285
B.3.1 Community statistics following Fortunato[134] . . . 285
B.3.2 Supplemental tables . . . 291
B.3.3 Group 5 centrality measures . . . 291
B.4 Supplementary Community Detection . . . 294
B.4.1 Spectral methods using the modularity matrix . . . 294
B.4.2 Community statistics . . . 296
B.5 Generation of the synaptic network graph . . . 299
B.6 Community detection of PSP . . . 299
B.7 Samples . . . 299

C Supplementary tables chapter 3 **301**
C.0.1 Degree . . . 301
C.0.2 Betweenness ToppGene . . . 302
C.0.3 Closeness Centrality ToppGene . . . 309
C.0.4 Eigenvector centrality . . . 314
C.0.5 Transitivity . . . 319
C.0.6 kcoreness . . . 322

D Data sources and code **327**

Chapter 1

Introduction

Psychiatric and substance abuse disorders account for 7.1% of disease burden as measured by disability-adjusted life years[285]. A range of environmental and genetic factors are known to be implicated in their development. Of the genetic factors, most encode molecules that are found in the neuronal synapse, a structure that is responsible for the exchange and processing of information between nerve cells[155].

1.1 The synaptic proteome

The synapse is a complex system composed of thousands of proteins that form higher order structures (the synaptic proteome)[155]. There have been great advances in the understanding of the molecular mechanisms of synaptic function in the last 20 years[158]. The synaptic proteins form a complex system and to understand their role in health and disease it is necessary to understand how it functions as a system. One important way of understanding complex systems is as a network where nodes, in this case proteins, are joined by edges if they interact.

1.1.1 Synapse introduction

A nerve cell is composed of a number of dendrites that receive incoming signals, a cell body or soma and an axon that transmits outgoing signals (figure 1.1). Between the axon and the dendrite are synapses (figure 1.2). Synapses are junctions where information transfer occurs between nerve cells and are the most abundant structure in the brain [155].

Each nerve cell has one outgoing axon but a cortical neuron receives input from

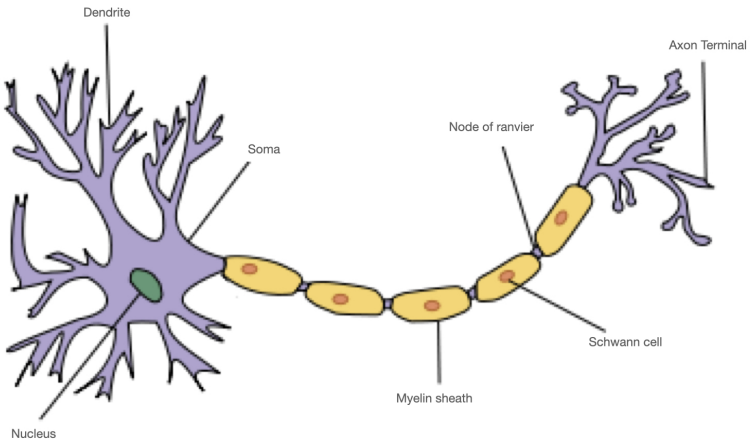


Figure 1.1: Soma and dendrites

around 100,000 synapses, although only around 50 are active at any one time[252][1]. Although synaptic numbers vary over time, the number of synapses in the human neocortex has been estimated to be 1.5×10^{14} (0.15 quadrillion)[319]. A synapse connects a presynaptic neuron and a postsynaptic cell across the synaptic cleft [394]. The synapse is made up of proteins that self organise into macro-molecular machines that are responsible for signal transduction and processing[135]. The nerve cell integrates incoming information from the dendrites into a binary response at the axon[234],[340].

The term synapse, from the Greek "to clasp together" was coined by Sherrington in 1897 writing in Foster's textbook of physiology [379]. Since Cajal showed that there was a gap between nerve cells the accepted model of the nervous system has been of a structure composed of nerve cells built into higher order circuits[349],[157] rather than a reticulum or syncytium. Cognitive complexity was thought to arise from an increasing number of connections (the neuronal hypothesis)[158].

The majority of synapses in the brain belong either to the excitatory glutamate system or the inhibitory gamma-aminobutyric acid (GABA) system. The most common type of synapse and neurotransmitter are the glutamatergic excitatory transmitters which comprise the main information processing system in the brain [391],[302]. Glutamate is also a precursor for GABA.

[1] number has varied see also [236] which has 10,000

1.1.2 Molecular genetic revolution

The neuronal hypothesis holds that the brain's functional unit is the neural circuit and that brain complexity arises from the pattern of nerve cell and circuit connections[158]. Until recently the molecular interactions of the synapse were believed to be relatively simple with a small number of proteins sufficient for signal transfer and the control of synaptic connection strength [159], [251]. This hypothesis is not unreasonable; simple models of neurons can carry out complex calculations requiring only weights at their connections analogous to connection strength at synapses[184]. Synaptic function as signal transmission and modulation of connection strength concurred with the model of learning by increasing synaptic strength formalised by Hebb [171].

The synapse controls the strength of the connection between nerve cells. If the activation of one nerve cell repeatedly activates another their connection will strengthen over time (Hebbian learning)[171]. Hebb's rule [171] requires the arrival of a signal at the presynaptic terminal, and consequent postsynaptic depolarisation leading to a stronger synaptic connection. As neural circuits determine behaviour, these repeated patterns of activation form a basis for learning.

The discovery of long term potentiation (LTP) by Bliss and Lomo [46] ultimately led to the discovery of the importance of glutamate transmission in LTP and that disruption of synaptic proteins reduces LTP and learning. Hebb posited a model where the connection strength between two neurons depends on how often they are activated together. Both short term potentiation and long term potentiation are involved in adaptive neural information processing by controlling the strength of the connections between neurons[438].

Direct molecular interrogation of synaptic proteins became possible with mouse models in 1992 allowing the identification of synaptic proteins and the study of their function using knockdowns [154],[381]. Further advances in molecular genetics revealed the complexity of the synaptic proteome. In 2000 Husi et al. [191] identified 77 proteins in the N-methyl-D-aspartate (NMDA) receptor complex (NRC). Over the next 5 years many more synaptic proteins were identified. In 2006 an analysis considered the network of 186 proteins in the NRC/Membrane Associated Guanylate Kinase (MAGUK) associated signalling complex (NRC/MASC) [340]. The number of known proteins in the post synaptic proteome rose to 1,124 by 2006[81].

The number of proteins and the complexity of their interactions exceeded what might be expected to be necessary to simply maintain connection strength and led to the

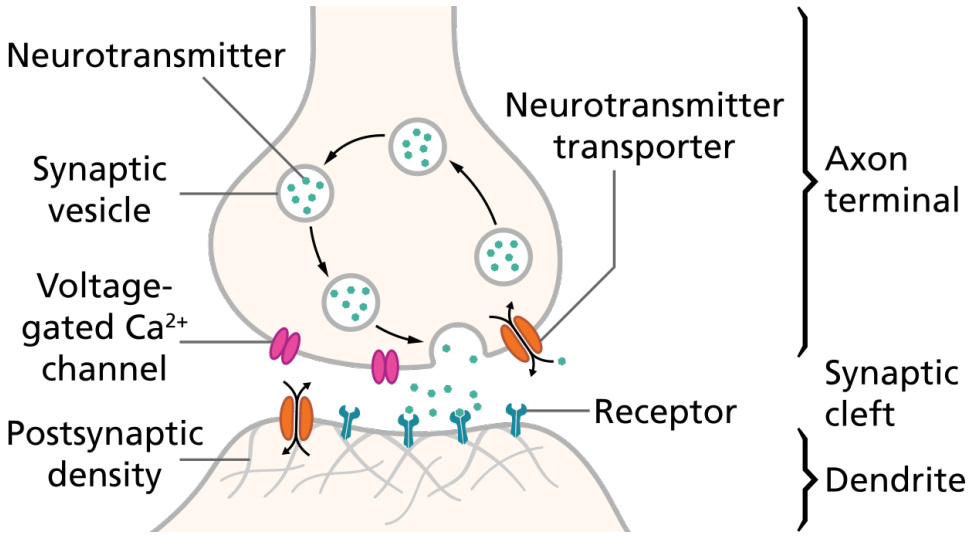


Figure 1.2: Schematic representation of synapse showing presynaptic area (labelled axon terminal) containing vesicles rich in neurotransmitters and post synaptic density. Taken from: Thomas Splettestoesser Creative Commons `https://commons.wikimedia.org/wiki/User:Splette`

recognition that the proteome was vitally involved in signal processing and computation as well as signal transduction[158]. Over 3,000 proteins have now been identified in proteomic studies of the post synaptic density (PSD), although this number appears to be reaching a plateau[173]. The proteins that form the structure of the post synaptic area are modular and form higher order complexes whose structure is related to their function[340],[456],[135]. To understand the synapse we must try to understand the complexity of this system.

1.2 Synapse structure

The synapse is composed of a presynaptic area filled with neurotransmitter vesicles and a postsynaptic area containing neurotransmitter receptors and structural proteins separated by a cleft. Nerve signals arrive at the presynaptic area, cause or inhibit the release of neurotransmitters from synaptic vesicles which diffuse across the cleft and are detected by postsynaptic receptors whose activation may lead to depolarisation of the cell axon.

1.2.1 Presynaptic area structure

The presynaptic area is separated from the post synapse by the synaptic cleft [364], an area approximately 20nm in size [459], spanned by trans-synaptic adhesion molecules[279]. Mitochondria [233] are an integral part of the presynaptic area and were noted in early electron microscopy examinations [160]. The presynaptic area developed from vesicle transport systems in earlier organisms and has developed more recently than the postsynaptic density (PSD)[119].

Neurotransmitters are substances that diffuse across the synaptic cleft and give rise to signal in the postsynaptic receptors. The presynaptic area contains neurotransmitters stored in vesicles docked and fused with the plasma membrane at a specialised active zone. This is an electron-dense area, parallel and juxtaposed to the plasma membrane[371],[394]. A protein complex of 5 evolutionarily conserved proteins: RIM, Munc13, α-liprin, RIM BP and ELKS [394] primes the vesicles. Piccolo and Bassoon proteins are also associated with the active zone in vertebrates[164] and SNARE proteins facilitate vesicular fusion [394]. The active zone is closely associated with a cytoskeletal matrix: the cytoskeletal matrix at the active zone (CAZ) [371]. Postsynaptic activation results in CAZ clustering [149].

Active vesicles docked at the edge of the membrane are released following calcium influx [234]. After neurotransmitter release, the vesicle is recycled by endocytosis[18].

Short term changes in synaptic plasticity (short term potentiation) are mainly caused by changes in the probability of presynaptic vesicle release. In contrast, long term potentiation occurs mainly at the postsynaptic density. A substantial amount of neural computation occurs in the presynaptic area and is dependent on synaptic proteins [394]. In some nerve cells such as photoreceptors, the active zone is ribbon-like and capable of prolonged discharge [18].

1.2.2 The postsynaptic area and postsynaptic density

The postsynaptic density is an electron-dense area parallel to the synaptic plasma membrane and extending $\approx$ 35nm deep to it [169]. Formed from structural proteins that bind to actin in dendritic spines and to the receptor complexes, it contains protein kinases, receptor molecules and protein scaffold complexes such as postsynaptic density protein 95 (PSD-95). There are slight differences in its composition between mouse and man [34], but in general, its component structures are conserved in vertebrates. The components of the postsynaptic area arose in the earliest prokaryotes to detect changes

in chemical gradients with movement and are the most ancient part of the synapse developmentally [157], [119].

Neurotransmitter chemicals typically mediate signal transfer across the synaptic cleft although direct electrical transmission can occur across gap junctions in the retina and some locations in the central nervous system [391]. The neurotransmitter receptors are bound to cytoskeletal proteins and modified by members of the membrane-associated guanylate kinase family (MAGUK) of proteins involved in synaptic plasticity [456]. Glutamate release occurs in areas closely aligned with glutamate receptors in the postsynaptic density (PSD) [169].

The receptors in the postsynaptic density show a geographical variation with the α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptors (AMPAR) and N-methyl D aspartate (NMDA) receptors in close apposition to the presynaptic active zone and the metabotropic glutamate receptors (mGluR) in the perisynaptic zone[370].

The ionotropic receptors have the fastest signalling characteristics and include NMDAR, AMPAR and Kainate receptors (KAR) named after the particular agonists found to provoke them[287],[42]. Kainate acid receptors have a metabotropic G protein-mediated element in their response that has been shown to mediate non NMDA LTP in the hippocampus of the rat [331]. They are less involved in synaptic transmission [86] than the two pure ligand-gated ionotropic receptors NMDAR and AMPAR.

The metabotropic glutamate receptor activates heterotrimeric G proteins and results in both signalling and recruitment of further ionotropic receptors to the postsynaptic area[304],[333].

Postsynaptic proteins show many remarkable features. They are relatively resistant to complete abolition of function [214], [69] and even in knockout animals, some signalling remains.A prime reason is functional redundancy between synaptic protein family members. Proteins can be grouped into families that share a common ancestor[131] and have similar coding sequences[99],[121]. Proteins involved in signalling pathways (such as synaptic proteins) particularly commonly have functioning redundant protein copies [203]. Particular examples in the PSP include APP[377] and MAGUKs[74]. In addition, the dysfunction of many disparate synaptic elements can result in common phenotypes, particularly neuropsychiatric phenotypes. A wide variety of mutations at different proteins and different loci within specific proteins can give rise to common disorders such as Autism Spectrum Disorder (ASD)[402]. By contradistinction point mutations in crucial receptor proteins often give rise to severe phenotypes, as we shall see later, but also give rise to different phenotypes from mutations

at different points in the protein (a case in point here being GRIN2A) [120]. GRIN2A is a *pleiotropic* gene, one that gives rise to multiple effects [318],[385] and diverse neurological and neuropsychological phenotypes including intellectual disability, developmental delay, epilepsy, autism spectrum disorder and ADHD[392], [412],[451]. These features may *in part* be related to the organisation of the PSP into supercomplexes that inherit network based robustness in addition to genetic robustness from paralogue expansion[156].

1.2.3 Molecular function

1.2.3.1 Neurotransmitter receptors

L glutamate, a small amino acid, is the dominant neurotransmitter at excitatory synapses within the brain [304]. Glutamate levels are controlled by the glutamate glutathione pathway and glutamate transporters[373].

Fast excitatory neurotransmission is carried out at the ionotropic glutamate receptors AMPA and NMDA [403]. Ionotropic glutamate receptors consist of four sub-units spanning the extracellular membrane that form a central ion pore. AMPA and Kainate receptors are homo or heterotetrameric, but the NMDA receptor is obligatory heterotetrameric. The AMPA receptor is responsible for most of the fast excitatory transmission in the central nervous system [174]. It is composed of four sub-units GRIA1-4 and is mainly permeable to sodium (Na^+) and potassium (K^+) ions. The glutamate receptor sub-units originate from 18 different gene protein products representing NMDA, AMPA, Kainate and delta groups [403]. The NMDA receptor is also composed of four subunits (a tetrameter) but is permeable to calcium ions. AMPA and KAR are fast-acting, working in less than 1ms[403] while NMDAR depolarise more slowly, in the 10ms range [403],[303].

Magnesium (Mg^{2+}) ions initially block NMDA receptor activity, and AMPA receptor activation is required to displace the Mg^{2+} ions. AMPA is intimately involved in long term potentiation. After Ca^{2+} ion entry from the synaptic cleft, Calcium/calmodulin-dependent protein kinase II (CaMKII) is activated and leads to an increase in AMPA receptors at the postsynaptic membrane. Calcium can enter through NMDA receptors after unblocking of Mg^{2+} ions, and its entry leads to activation of Calcium/Calmodulin dependent protein kinase. The kinase subsequently binds to NMDA receptors and phosphorylates sub-units of the AMPA type glutamate receptor.

Slow synaptic transmission, mediated by G protein second messengers, results

from the action of metabotropic glutamate receptors (mGluR) [304] although mGluR signalling can also occur in time scales closer to ionotropic glutamate receptors (iGluR). There are eight mGluR subtypes. Metabotropic glutamate receptors are family C G-Protein-coupled receptors (GPCRs). Group 1 mGLU promote intracellular calcium release whereas group 2 are localised pre and postsynaptically, and group 3 are localised presynaptically [304]. Metabotropic glutamate receptors exert their effect by increases in protein synthesis. Group II have been related to anxiety, schizophrenia and Alzheimer's disease[396].

1.2.3.2 Long term potentiation

Periods of synaptic activity can increase the strength of synaptic connections by the process of long term potentiation (LTP).

Cajal first proposed that an increase in the strength of the connection between neural cells leads to learning in 1911 [303]. In Hebb's model synaptic modification occurs following coexistent pre and postsynaptic activation (two factors are required)[171]. Later experiments showed that tetanic stimulation of rabbit hippocampal synapses increased the strength of subsequent responses[46], and that NMDAR activity is necessary for LTP[80] but not for neurotransmission. Interfering in the process of LTP in animal models leads to defects in memory and learning [253]). LTP is a complex process with the modular organisation of 70-80nm modules of independently activating AMPA receptors being integral to the variety of potentiation responses[252]. Six different forms of plasticity are observed in the synapse short term plasticity, long term plasticity, late long term plasticity, long term depression, distance-dependent scaling and homeostatic scaling [252]. Short term potentiation (STP) results from weak stimuli and occurs early, and is a promising model of working memory as GluA1 knockdown animals have abnormal STP and working memory[252]. Early LTP is likely mediated by phosphorylation of GluA1 by CaMKII. Late long term potentiation is associated with a change in the size of the PSD. LTP, by definition, begins 30 minutes after a tetanic stimulus. Late LTP is associated with more AMPAR signalling, protein synthesis[323] and an increase in quantal probability. This may require other factors such as dopamine in CA1, postulated as an indicator of salience, or brain-derived neurotrophic factor (BDNF). For this reason (that three factors rather than two are required), the process has been referred to as NeoHebbian[252]. The mechanisms of this late stage include increase in receptors from PSD 95 mediated CaMKII phosphorylation of Stargazin or SynGAP or activation of extracellular signal-regulated (ERK)

kinase through SynGAP. Throughout all of this it is clear that memory and learning appear to be intimately linked to the combination and structure of synaptic proteins.

Computational models of the synapse have been developed that simulate these responses on the assumption that there are 'hotspots' where there are either modules of AMPA receptor present or absent on vesicle release, and plasticity can occur with recruitment to silent hotspots[252].

1.2.4 The synaptomic theory

An alternative model of synaptic learning has been proposed where the complexity of the composition of the proteome is integral to learning[159]. In this model, synaptic proteins form complexes and supercomplexes that are specifically attuned to different temporal neural signals, and synaptic protein organisation acts as a 'zip code' addressing different areas in the brain[158].

Evidence that is inconsistent with synaptic strength as the only mediator of learning[159] include preservation of spatial learning after NMDA signalling and LTP have been blocked[22] and mutations to PSD 95 leading to cognitive impairment but enhanced long term potentiation[305]. It may be that the NMDA receptor both mediates learning and LTP[159]. In this model too the modular structure of the proteome is closely related to its function in cognition.

1.2.5 Role in disease

Bayes et al.[35] identified, in human neocortical biopsies, 1,461 proteins in the human post synaptic density (hPSD) and 748 that were consistent across three replicates. Using the Online Mendelian Inheritance in Man (OMIN) database [168] and limiting the analysis to monogenic diseases the authors showed that 269 diseases resulted from mutations in 199 hPSD genes and that 133 were primary nervous system disorders. The post synaptic density is the micro-molecular structure most implicated in neuropsychiatric disorders[155]. Synapses are the locus of action of many effective neuropsychiatric drugs and are central to the putative mechanisms of many neuro-psychiatric disorders [400], [188]. Psychotropic drugs with effects upon mood and perception frequently act at the synapse [227], as do other less obvious factors influencing cognition and psychiatric health such as obesity[48]. Reviews by Pocklington [339] and Hall [167] document convergent evidence for the importance of synaptic dysfunction in the aetiology of schizophrenia. Evidence of its role in schizophrenia was found

by Föcking et al.[132] using proteomic studies of human brains. Purcell et al.[345] report the role of ARC (Activity-regulated cytoskeleton)-associated protein, which facilitates AMPA receptor re-uptake and calcium transport, in schizophrenia. Glutamatergic signalling is also implicated in Major Depressive Disorder (MDD)[286],[389]. Memantine, a non competitive NMDA receptor antagonist is used in the treatment of Alzheimer's disease[267],[14]. The anaesthetic ketamine, a non competitive antagonist of the ionotropic glutamate NMDA receptor, is reported to be rapidly effective in the treatment of major depressive disorder and its synaptic function has been shown to be integral to its effectiveness[208]. Metabotropic glutamate receptors have been implicated in neuroinflammation and as a potential target for therapies for bipolar affective disorder[128],[219].

1.2.6 Evolution

The PSP has ancient origins, the proteins evolved around the time of the earliest eukaryotes before the neuron[158]. Constituent proteins are found in yeasts, and the new proteins found in the mammalian PSP are the result of paralogue expansion[156]. The PSP originated in the sensory signalling complexes of bacteria. These are too small to detect a difference in chemical gradient along their length and the function of these earliest receptors was to detect differences in chemical composition of their environment over time as they swim. It has therefore been hypothesised that the earliest function of the PSP is in recording differential pattern to the time course of signals arising in the organism. The presynaptic area is derived from endocytosis related mechanisms.

Two rounds of whole-genome duplication occurred approximately 550 million years (550 Ma - mega annum) ago at *"the base of the chordate evolutionary tree"*[305], and is postulated to give rise to increased proteomic complexity by paralogue expansion[156].

PSD proteins are highly conserved and have greater conservation than the rest of the genome in mice and man[35]. The rodent and human synaptic proteome have broadly similar components, with a comparison of their synaptic proteomes revealing 1,556 mouse proteins and 1,461 human ones in 2012[34]. The number of synaptic proteins identified in proteomic experiments has increased since then[173] but it would seem reasonable that the proportions remain similar.

The similarity of mouse and human sequences in orthologue pairs is around 85% [45]. Primates are our closest evolutionary relation, and at the level of amino acid se-

quence differ by about 1% in coding regions, and most of the difference is neutral[284], [274]. Comparison of human and primate proteomes are less exhaustive than murine but reveal differentially expressed genes in primate (human and marmoset) compared with mouse and rat enriched for association with neuroplasticity[225].

Primate cognitive complexity results from non-synaptic factors such as morphological differences controlled by non-PSP genes. Much of the difference between primates and human are transcriptional differences [249]. Other genes have differences in their protein sequence that affect synaptic formation and abundance of proteins in the synapse for example CBLN2 which interacts with NRXN1-3 and GRID 1-2[380]. Differences in expression and protein sequence will not be absent from the synaptic proteome network. Nine percent (9%) of genes show a human-specific developmental trajectory compared to macaque [21] and human-specific features are associated with disease[324]. In non-human primates, a similar pattern of temporal expression is seen in ageing vulnerable local proteomes (e.g. hippocampus) compared with the more resistant occipital[153] that is absent in rodents[256].

Hill combined data from the CHARGE consortium (n=53,949)[95] and UK Biobank (n=36,035)[96] to show, using stratified linkage disequilibrium score regression[130], that areas of the genome enriched for common genetic variants, single nucleotide polymorphisms (SNPs), associated with differences in intelligence were under negative evolutionary pressure, finding these regions contained 2.6% of SNPs but $\sim$40% of SNP based heritability[181]. Linkage disequilibrium refers to the correlation between neighbouring alleles and hence the tendency for neighbouring sections of chromosomes to be inherited together[353][2]. The finding that genetic variants associated with differences in intelligence are under negative selective pressure by Hill et al. [181] raises the question of whether variants associated with intelligence found in the PSP will be central within the proteome network, as centrality has been linked to intolerance of mutation (see 1.2.10) a question that will be tested in this thesis.

1.2.7 Importance

There are many reasons to be interested in synaptic function but I consider five to be of particular importance.

First, synapses are involved in the overwhelming proportion of neuropsychiatric disorders [155]. Second, they are composed of a large number of proteins, many of

[2]Networks as well as reduplication can provide robustness

which have roles outside of the nervous system and are an attractive basis to understand the genetics of complex cognitive traits, involving generalist genes, many of which exhibit significant pleiotropy [378],[336]. Third, the synapse has proved a useful practical model of information processing and simplified models of the synapse have been able to carry out powerful computations [184], [101]. However the synapse is considerably more complex than the simple weight used in these models. Fourth, synapses are the building blocks of higher levels of brain structures [16] and will likely affect their function [101].

Finally they provide a molecular basis for human learning and memory [204],[138].

1.2.8 Systems Biology approaches to the synapse

The function of complex biological systems arise both from the nature of their components and how they are connected[221]. The behaviour of the system can differ significantly from that of component parts and considering system structure, system dynamics, the control method and the design principle can help to understand these differences [221]. Complex systems can be compactly represented as networks (section 1.2.10) and can be analysed using methods derived from fields including graph theory, information theory and statistical physics[289]. Advances in computation [306], algorithms, databases, nomenclatures [195] and ontologies [382] have further improved our ability to understand, model and reason about these systems. Increasingly well powered population genome wide association studies (GWAS) (section 1.4) of intelligence and the related trait of educational attainment have recently become available. These provide an opportunity to study the relationship between the structure of the synaptic proteome and the genetics of a complex cognitive trait (see section 1.3.4.1).

1.2.9 Models of the synapse

In a network model of the synaptic proteome the nodes are genes that encode a synaptic protein and an edge shows that the protein products of two genes (nodes) interact. A network can be analysed at three levels: the network's properties as a whole, the properties of individual nodes, or the presence of structures or organisation within the network[294].

Initial yeast network models of the synapse were of poor quality so data mining of the literature for interactions was required to develop effective and comprehensive

static models (section 1.2.9.1) [340],[16]. Databases include high quality experimental evidence of protein interactions and use standardised ontologies to organise their information [57], [176], [212]. Databases contain evidence on interactions that varies in reliability from direct experimental records to inferred interactions (e.g. literature data mining and co-expression). In order to develop a useful model of synaptic interactions, a significant amount of data cleaning, organisation and transformation of these databases is still required. The synaptic proteome network in this thesis uses only interactions inferred from experimental evidence of direct interaction in mouse or human (section 2.1). Synaptic components are purified using centrifugation and protein components examined by mass spectrometry. Proteomic experiments provide a parts list for the nodes of the synaptic network and data integration of interaction databases provides edges. Synaptic proteomics are reviewed in depth by Laßek et al. [233] and Armstrong and Sorokina (2012) provide a comprehensive review of the development of systems biology models of the synapse [16]. Network models of the synapse can show simply the interactions between proteins (static models) or can model the rate of flow within the network (dynamic models)

1.2.9.1 Static methods

Protein-protein interactions in the synapse can be represented as an interaction graph (network) with dots (vertices) representing the proteins and an edge connecting the vertices if the proteins interact. Alternatively the gene that encodes the protein product may be used. This is helpful if we wish either to annotate the structure with results from the biomedical or experimental literature, or to use the graph structure to better understand results from population molecular genetics.

1.2.9.2 Dynamic models

Synapses can be modelled as dynamic entities with interactions between components and specific location, stoichiometric and kinetic constraints. Two common types are those based on systems of partial differential equations or agent based models [387],[386], [421].

Although this thesis will not address dynamic models the interaction logic of the static model is essential for the generation of dynamic models. In addition structures within the network that have an enriched association with animal or population genetic models of cognition will be fruitful areas for future dynamic simulation .

1.2.10 Systems biology of the synapse and networks

Many complex networks such as the internet, academic collaboration networks, and protein interaction networks share common properties at the level of the entire network. These properties include the small world phenomena or scale-free degree structures[24],[429], and provide evidence of how these networks form, how information passes through them and how resilient they are to disruption [362],[6],[43]. Recently the 'small world property' has been posited as a potential mechanism for an omnigenic model of genetic influence on complex traits [429],[54]. The synaptic proteome appears well suited to network analysis as synaptic proteins form modular structures [340] whose structure appears 'optimally tuned' for normal function [155]. Complex networks also possess meso-scale structure the most studied being communities. Identifying and studying these large scale structures in the network may help in understanding the role of synaptic components in complex traits and neuropsychiatric disease. In addition there are many features of the synapse that may be understood in terms of network structure in addition to other genetic explanations (for example not all knockdowns lead to complete loss of function [214], [69]). I will introduce the relevant background to graph measures, vertex measures (section 3.2) and structures within the synaptosome (section 4.1) in more detail in the relevant later sections.

Critical genes in protein-protein interaction networks are reported to be more likely to be essential to life[197] and an outline of the network properties of disease genes[26] has been reported. This has led to numerous hypotheses about the effects of network structure on biological phenotypes, for example that central nodes will be more important or that dense sub graphs (communities) may be associated with functions. Analyses have been reported of the network properties of earlier models of the synaptic proteome [338],[266] that support the importance of communities; however previous studies have been limited to useful but heterogeneous gene-disease records. We now have available high quality population genetic studies (GWAS) and can use these to investigate the utility of network analysis of the proteome and learn practical lessons on their integration in genomic examinations. Neuropsychiatric nosology is complex and controversial, and for this preliminary investigation we would like to use a complex trait or disorder that can be accurately measured, that is heritable, that shows evidence of an association with synaptic variation, and that has the best powered studies in neuropsychiatric complex trait genetics. Intelligence is the ideal trait for this.

1.2.11 A note on terminology regarding the Synaptic Proteome network

This thesis is an interdisciplinary work combining network science with molecular genetics. In this paragraph I would like to address as plainly as I can a possible source of confusion.

The network referred to in the thesis is a network of genes, the vertices represent genes but it is a protein interaction network (PIN). The actual synaptic proteome is made up of connected proteins. In living organisms proteins are made from instructions found in genes, so there is a one to one correspondence between proteins and genes. I could have presented the network as a graph where the vertices (nodes) are proteins and there is an edge between the nodes if they interact. Because this thesis deals with the association between changes in genes and cognitive ability I follow the practice of previous authors and use genes as the vertices[338],[173],[266]. To recap the vertices are genes but they are representing the proteins that the gene makes. There is an edge between the vertices if the proteins (made by the gene) interact. This network is different from genetic interaction networks which have an edge between genes if their expression is correlated. This is explained in more depth in section 1.6 which includes a brief discussion of genetic interaction networks.

1.3 Intelligence

People differ in their cognitive abilities and their ability to solve mental problems. The approaches to testing intelligence or cognitive abilities have been informed by function or lesion studies in the neuroscience community and on correlation between disparate tests revealing latent factors in the differential psychology literature [106].

1.3.1 Measures and heritability

Given a number of different tests of cognitive ability, individuals who do well on one test will tend to do well on others, and hence there will be a correlation between these results. This correlation between performance over several categories is the general intelligence factor, *g* or 'general intelligence', described by Spearman in 1904 [104],[106]. This construct accounts for about half the variance in individual cognitive performances [108]. The g factor can be discovered from a variety of disparate cognitive tests using principal component or factor analysis or by using cognitive tests that

'load highly on the general cognitive factor' [200], [108].

As well as general cognitive ability (g), there is a positive correlation in performance on similar tests that relate to factors representing cognitive domains such as memory, processing speed, vocabulary, reasoning and spatial ability [105]. In addition crystallised intelligence (gc) and fluid intelligence (gf), identified as factors of g by Cattell and Horn, change at different rates over the life course. Crystallised intelligence is composed of features such as knowledge of facts, vocabulary and certain number skills and does not tend to deteriorated until old age in contrast with fluid intelligence which requires *'on the spot processing'*[105],[95]. Fluid intelligence has been defined as 'novel problem-solving capacity' and declines with age (see [211]) for a review. Ritchie and Tucker-Drob[361] differentiate between fluid measures such as *'reasoning, memory, processing speed, and other tasks that could be completed without outside knowledge from the world'* and '*crystallized tests (tests that assessed skills such as vocabulary and general knowledge)*' in categorising studies of intelligence and educational attainment.

The heritability of cognitive abilities has been recognised since Galton (1865) [103],[139]. The genetic influence on the variance of human intelligence has been estimated at between 30 and 80% and the heritability of the g factor increased from 30% in youth to 50% in adulthood to older age[108].

Intelligence measures are relatively stable over the lifespan, estimates of the correlation between measures of performance in youth and age vary between about 0.5 and 0.7 (0.51-0.67) (see for example Gow[152] comparing results on identical tests taken at 11 years old and again at either 79 or 87). Twin studies show greater correlation in changes in cognitive ability over a lifetime for monozygotic than dizygotic twins[108]. Those of higher intelligence do not have better cognitive ageing but are more likely to have habits such as exercise and education that predispose to successful cognitive ageing[106] and have higher intelligence due to the stability of the measure[152].Genetic differences may have different effects across the lifespan such as variations in apolipoprotein E which affects intelligence at older ages[109].

1.3.1.1 Intelligence and synaptic proteome models

Cognitive ability is an attractive complex cognitive trait to investigate using a structural model of the synaptic proteome. It is highly heritable and we know that a large proportion of its heritability can be marked by single nucleotide polymorphisms. In addition it can be reliably measured quantitatively, has a positive as well as a negative extreme,

and is normally distributed [336]. Finally we have evidence that synaptic components are involved in variations in human cognitive abilities[178].

The fact that this trait is normally distributed with a positive as well as negative extreme is attractive as it allows better discrimination of the importance of genes specifically for cognition. Some genetic mutations may cause problems with cognition as part of a more general effect on vitality. Hubs, genes or proteins in particular with large numbers of connections to other nodes are disproportionately involved in severe phenotypes but less so in complex and polygenic traits [26] [35]. Previous network analyses have found communities in portions of the synaptic network to be enriched for terms related to intellectual disability [340]. It is not clear however whether this is because they are associated with intellectual disability through a direct effect on cognition or whether these arise through mutations of important genes with pleotropic effects including intellectual disability. Pocklington et al. (2006) found that intellectual disability, unlike other neuropsychiatric disorders associated with cognition was widely spread in the genome and in the clusters that he identified in the MASC complex[340]. Human population genetic studies of intelligence (variations in cognitive ability) using network methods will allow us to see whether there are locations that are enriched for differences in human cognitive ability, whether it is particularly central genes that have these effects, whether there are specific regions associated with intellectual disability and cognitive ability, and if there are specific regions, whether these are the same.

1.3.2 Cognitive epidemiology

Cognitive epidemiology is the study of how intelligence as measured by psychometric tests – is associated with mortality, illness and health [105]. The effects of differences in cognitive function in health outcomes have been recognised since the 1930s [107].

A systematic review of nine cohort studies found that IQ in childhood and early adulthood was associated with lower mortality in middle and late adulthood [32]. Higher cognitive ability is associated with important outcomes such as educational achievement and occupational success.

It is possible that the effect on lifespan might be due to social factors that have a common effect on intelligence, but studies including an analysis of three twin registries supported a genetic effect to explain correlation between lifespan and intelligence (for example the correlation is stronger for monozygotic than dizygotic twins) [15]. Gen-

eral cognitive function has been reported to be correlated with differences in the incidence of psychiatric disorders and physical disorders across he life course[435],[179]. Hill et al. showed, using linkage disequilibrium regression, a common effect of genetic variants on cognitive ability and mental disorders with different effects at different points in the life course[179].

1.3.3 The molecular basis of cognition.

Two fundamental approaches have been made to the molecular basis of cognition. One exemplified by the Genes to Cognition group[88] is a pathway based approach of examining variations in MASC and other synaptic genes which are linked to changes in memory in experimental animals[108].

Molecular genetic approaches need to determine whether there is an effect of genetics on the environment, i.e. genetics mediating environmental change, or whether genetic effects alter individuals' susceptibility to the effects of environment [108], [432]. The other approach is to determine the association of common genetic variants in the population with differences in intelligence (GWAS).

1.3.4 Genome scale molecular findings

Genome wide association studies (GWAS) test the association of common genetic variants (Single Nucleotide Polymorphisms - SNPs) with a trait[416]. SNPs (section 1.4.1 tend to be inherited in blocks due to a correlation in their frequency between nearby SNPs (linkage disequilibrium)[353]. In the last few years there has been a dramatic increase in the size of GWA studies of intelligence and hence their power and the information contained in them. In 2014 the largest GWA study of intelligence in children (n=18,000) found no variants of genome wide significance [39]. Measuring educational attainment, which has a moderate correlation with intelligence (section 1.3.5), allowed studies to increase their sample size which led to genome wide significant findings but the initial effect sizes were small (section 1.3.5)[358]. Recently cohorts with larger sample sizes such as EA2 [309] and UK Biobank[180] have permitted meta-analyses to identify variants with significant association with cognitive ability at the genome wide level.

Genome wide complex trait analysis (GCTA) [445] of GWAS studies had shown that common SNPs could account for about half the heritability found from twin studies. Davies et al (2011)[94], used GCTA with five cohorts in the Cognitive Ageing

Genetics in England and Scotland (CAGE) consortium to estimate the proportion of crystallised and fluid cognitive ability accounted for by linkage disequilibrium (LD) between SNPs and unknown causative loci. The estimates were 51% for fluid intelligence and 40% for crystallised intelligence. Using data from the Cohorts for Ageing Research in Genomic Epidemiology (CHARGE) consortia [343], Davies et al. (2015)[95] estimated the lower bound of the narrow sense heritability for variation in cognitive ability to be 0.29 and 0.28 from the two largest cohorts in the meta-analysis[3]. They found 13 SNPs of genome wide significance for cognitive ability in a meta-analysis of data from the CHARGE consortium (n= 53949). Using VEGAS (section 1.4.2) to generate gene level statistics, one gene was found of genome wide significance: HMGN1. The genomic inflation factor ($\lambda = 1.14$) suggests a polygenic effect on the phenotype.

Davies et al.[96] in 2016 reported a genome wide association study of 112,151 individuals from UK Biobank [65],[118]. The association was tested between genotype and four measures (educational attainment and three cognitive tests: reaction time, memory, numerical reasoning). Gene based statistics were calculated using MAGMA without a gene window [239](section 1.4.3). 149 SNPs at three independent regions were significantly associated with verbal numerical reasoning. Gene based analysis identified 17 significant genes across seven genomic regions. These included genes associated with mitochondrial function (NDUFA6), septin 3 which is associated with Alzheimer's disease and other genes associated with neurobiological pathways (e.g. ATXN3L). The proportion of variance explained by common genetic variants calculated by GCTA was 31%. For reaction time there were 36 SNPs of genome wide significance spanning two regions. 23 genes across 9 regions were found to have significant associations with reaction time.

There were no genome wide significant SNPs for memory. 115 SNPs were identified that were associated with educational achievement. Gene based analysis identified 95 genes across 28 regions associated with educational attainment.

1.3.4.1 Recent GWAS

Sniekers et al.[383] in 2017 reported a meta-analysis of 78,308 individuals of European ancestry from 13 cohorts including UK Biobank (*n*=54,119). 336 SNPs were discovered in 18 genetic loci. Twenty two genes were implicated by gene location, and

[3]Atherosclerosis Risk in Communities Study (N=6,617) and the Health and Retirement Study (N=5,976).

MAGMA genome-wide gene association analysis (GWGAS) (section 1.4.3) identified 47 genes at a genome wide level of significance. Seventeen of these were also found by SNP genetic loci. Only one gene set was found to be significant after correction for multiple comparisons GO: regulation of cell development using MAGMA GSA ($n = 809$, adjusted p=0.03, p=3.5×10^{-6})[239],[383]. The authors note that the four next most significant gene sets (although not significant after correcting an α level of 0.05 for multiple comparisons) are related to neural function. The authors report that 14 of 44 genes which had data from the GTEx consortium were differentially expressed in the brain [84]. Further analysis by combining this data with high intelligence cohorts is described [79].

Hill et al. [180] used multi trait analysis of genome wide association studies using summary statistics (MTAG)[407] to increase sample size to 248,482 from 199,242. They included summary statistics from Sniekers et al.(n=78,308)[383] and the Educational Attainment 2 study (EA2) of educational attainment reported by Okbay et al. (n=329,417)[309]. These were augmented by 120,934 new UK Biobank participants (UK Biobank Participants in the other studies contributing to the analysis being excluded). The new UK Biobank participants from this study will be used as a population sample in this thesis and referred to as the $\text{Intelligence}_{\text{Discovery}}$ sample (section 2.2.1).

They found 187 genetic loci associated with intelligence involving 538 genes. Gene set analysis using MAGMA (section 1.5.2) revealed enrichment for seven pathways including neurogenesis, regulation of nervous system development, regulation of cell development, neuron projection, central nervous system neuron differentiation, synapse , neuron differentiation and oligodendrocyte differentiation.

Savage et al. [367] reported in 2018 a meta analysis of 14 different studies of general intelligence. This was again dominated by UK Biobank (n=195,653) with a significant contribution from COGENT. They reported 205 genomic loci and 1,016 involved genes of which 939 were novel. For an in depth review of these studies see Plomin and Stumm[337].

1.3.5 Educational attainment

The use of educational attainment as a proxy for intelligence has allowed a great expansion in sample size in GWA studies [337], as educational attainment is routinely reported in a number of social science and other GWA studies. The correlation between educational attainment and intelligence is phenotypically 0.5 and genotypically

0.65 [337], [359]. In their report of a meta-analytic GWAS of intelligence, Sniekers et al. used LD score regression to calculate the correlation of their sample with 14 traits[383],[60]. The strongest assocation was with educational attainment (r_g=0.70).

Rietveld et al (2013)[358] measured educational attainment as a proxy for cognitive ability in a meta-analytic GWA study examining relationship between both college attendance (as a binary variable) and the number of years of education against the genotype. The discovery sample was 101,069 individuals, and the replication cohort 24,490 individuals. A total of 126,599 individuals underwent genotyping and summary SNP data has been made available on the Social Science Genetic Association Consortium (educational attainment) website [4]. The author found three SNPs of genome wide significance, and all measured SNPs accounted for 2% in the variance of outcome[358].

Okbay et al. (2016) [309] reported 74 loci associated with educational attainment in an expansion of the cohort used by Rietveld et al.[358]. The discovery cohort was now 293,723 individuals in size and the replication cohort carried out on UK Biobank (n=101,069). Gene set analysis showed enrichment for synaptic components, particularly SHANK2.

The largest educational attainment study to date is of 1.1 million individuals[237]. The study was conducted using years of education as a phenotype in individuals of European ancestry. The authors identified 1,271 independent genome wide SNPs and were able to explain 11% of the variance in educational attainment using a polygenic risk score [237]. Using DEPICT (Data-driven Expression Prioritized Integration for Complex Traits) they were able to show enrichment for pathways involved in the nervous system[329]. The authors performed an analysis to support their DEPICT findings using gene set analysis (GSA) using MAGMA.

1.3.6 Relevance to the systems neuroscience of the synapse

Intelligence as the manifestation of the integrated function of neural systems is 'central' to systems neuroscience approaches [336].

The genetic architecture of differences in cognitive ability, as of all complex traits to the extent that we can extrapolate from current studies, is the combination of many genes of small effect[336]. There is reason therefore to believe that studying their interactions may be profitable. Many of these genes are pleotropic, shown by their association with health outcomes [102] and psychiatric morbidity[179] but there is

[4] http://ssgac.org/documents/SSGAC_Rietveld2013.zip

some evidence that genes encoding synaptic proteins may be particularly associated with cognitive ability and educational attainment [178],[309].

Network studies have implicated subsets of synaptic genes in intellectual disability and disorders with a cognitive phenotype [340],[266] but have been limited to over representation analysis (ORA) of synaptic topological components (see section 1.5 for ORA).

1.4 Gene level tests

Genome-wide association studies (GWASs) test the association between genetic variants tagged by single nucleotide polymorphisms (SNP) and a phenotype, disease or trait in a population[416]. Testing the hypothesis that structures (*communities* - section 1.2.10) in the synaptic proteome network are associated with a particular phenotype requires a measure of the significance of the association between the group of genes encoding the structures' constituent proteins and the trait. Determining whether central proteins in the network are associated with a phenotype similarly requires a measure of the significance of the association of the gene encoding each protein with the phenotype.

GWA studies test the association between SNPs and phenotype. In generating a gene-level statistic, SNPs must first be assigned to genes and their aggregate significance calculated[330]. This generates a gene level test of association with phenotype. An example is the Genome Wide *Gene Association* Study (GW*GA*S) produced by MAGMA[239].

By aggregating genetic variation from SNPs to genes, the number of statistical tests can be reduced from over 500,000 to approximately 20,000[5], reducing the penalty for multiple testing and increasing power. Genes are a natural biological unit; gene-based tests are easier to interpret and provide the basis for further analyses such as gene set (pathway) analysis (section 1.5).

Many gene-level methods exist, but at present, a few are widely used in the secondary analysis of GWA studies (see table 1.1). Ideal gene-based methods should be computationally tractable, reliable and reproducible, particularly given documented issues with reproducibility in computational approaches to network and pathway studies (see also section 1.6)[280].

To ensure, as far as possible, that results in this thesis are generalisable and not

[5] approximately the number of protein-encoding genes

Implementation	test	year	*n* cit.	n(2016)	n(2019)
MAGMA [239]	mean χ^2 sampling distribution F test with genotype data	2015	802	753	486
FUMA [428]	MAGMA	2017	567	541	453
Pascal [228]	Fisher's method (modified)	2016	190	187	102
VEGAS [254]	Monte Carlo	2010	772	357	93
VEGAS 2 [278]	Monte Carlo	2015	162	156	60
HYST	implemented in KGGv[245]	2012	104	65	14
FAST	GATES	2013	38	34	9
KGG (version 4.2) [245]	GATES	2010	51	26	11

Table 1.1: Implementations of gene level association statistics for Genome Wide Association studies. The total number of citations (n cit.) and number since 2016 and 2019 are shown calculated using Google Scholar. For FUMA published in 2017 the number of citations on or after 2016 is less than total. Test column refers to the test statistic used to generate the gene level p value. Data accessed 28.07.20

dependent upon the accuracy of a particular software implementation, two of the three most widely cited packages (MAGMA[239] and Pascal[228]) are used for critical gene-level tests and GSA as described below.

1.4.1 Generating gene level scores from SNP level scores

The SNPs contained in a particular gene can be identified by comparing the SNPs genetic coordinates to the gene boundaries. SNPs are identified by reference SNP ID numbers (rs ID numbers) assigned by the National Centre for Biotechnology Information (NCBI) to a group of SNPs mapping to an identical location. These are recorded in the Single Nucleotide Polymorphism Database (dbSNP) maintained by the NCBI[223]. A "window" of up to about 50 kilobases(kB) can be included around the gene to try to include regulatory elements. If a SNP is less than the length of the window from either gene boundary it is included in the gene. The optimal size of window is not clearly established [330] (see also section 2.3.0.1 and section 6.2.0.0.1) and approximately 50% of SNPs remain inter-genic despite a window, and are excluded from analysis (section 2.3.0.1). Alternate approaches exist, including using linkage disequilibrium, but are not used in this thesis and will not be discussed (see [430] for an account).

Once mapped to a gene, the significance of the association of each SNP with the trait has to be combined to produce a gene level test statistic. The simplest approach, using the p value of the top scoring SNPs in a gene[422],[185], can be biased by gene size as larger genes will contain more SNPs [422],[117],[330]. Fisher's combination test[90] and the truncated p value test[442] have been used but are sensitive to the assumption of independence between SNP p values which is not valid in the presence of LD[246]. When new gene-level tests are introduced[254],[239],[228] the "gold standard" they are compared to is typically the set based test in PLINK [344], which permutes phenotype labels for a set of SNPs to achieve a null distribution. To do this for all SNPs in all genes is prohibitively computationally expensive and most methods use an approximation or asymptotic distribution to improve computational efficiency (for example Monte Carlo sampling in VEGAS or adaptive permutation in MAGMA. In addition the set based test requires genotyped data rather than summary statistics which may not be available. The analyses in this thesis use MAGMA [239] as the primary gene level test, and as an alternative, Pascal[228](section 1.4.4). VEGAS, the third highly cited method was used in a pilot study (not shown)[254] but not for subsequent analysis, for the reasons below.

1.4.2 VEGAS

VEGAS (Versatile Gene Based Association Study) is a widely used method for calculating gene level statistics. Offline and web based implementations have been available[6]. It was used to generate the gene level statistics for a study reporting Gene Set Enrichment Analysis [393] of synaptic components for differences in cognitive ability[178]. VEGAS uses Monte Carlo sampling as an alternative to a permutation strategy.

For all of the SNPs in a gene their p values are transformed to an upper tail chi square with one degree of freedom. The test statistic is the sum of these chi square values. Simulated p values are generated by drawing from a multivariate normal distribution with covariance matrix equal to the pairwise LD values calculated using PLINK[344] with HapMap phase 2 as a reference population [85]. This is done by generating a vector of independent standard normal values and multiplying them by the Cholesky decomposition of the LD matrix. The package does not require genotyped data and can use summary data.

[6]only vegas2 available now `https://vegas2.qimrberghofer.edu.au/`

It is simple to use, with limited options including a fixed 50kB window, and VEGAS results are available in the supplementary material for several studies of intelligence and educational ability [94], [178] [358]. While its running time is fast in comparison to the set test in PLINK, fast can mean days. The VEGAS 2 website suggests 24 hours as a typical running time for a contemporary GWAS, although I have often found it to be considerably longer. The number of simulations and hence running time limits the lower limit of calculated p values[228].

VEGAS 2 [278] uses 1000 Genomes phase 1 version 3 [91] rather than HapMap as the default reference population, allows the choice of gene windows and has a pathway enrichment package. It otherwise implements the same methods and has a comparable running time to VEGAS which rendered it difficult to use with the larger number of SNPs found in recent GWA studies, particularly over a remote connection, and for these reasons I found Pascal and MAGMA superior alternatives.

1.4.3 MAGMA

MAGMA (Multi-marker Analysis of GenoMic Annotation) is a standalone C++ application that calculates gene based statistics using a multiple regression rather than a Monte Carlo model [239]. The application can also be accessed through a web interface as part of the FUMA (Functional Mapping and Annotation of Genome Wide Association studies) package [428]. MAGMA can use either raw genotype data or summary data. Using summary gene data MAGMA uses either the mean χ^2 statistic or top χ^2 statistic to generate a p value. The mean method is the default and uses the properties of the sampling distribution[59]. The Top χ^2 method scores are calculated using an adaptive permutation method. Linkage disequilibrium is accounted for by using data from a reference population (here 1000 Genomes [83]). SNP locations in the reference population are those of human genome build 37 and subpopulations of European, African, South Asian, East Asian and Middle/South American populations are available at the project website[7]. Synonym files based on dbSNP 151 are provided to identify duplicate SNPs.

MAGMA has been widely used in the primary GWA studies of cognitive ability and educational attainment (see table 1.2 and section 1.4.4.1) and is the primary analysis method used in this thesis for gene-level tests and GSA. The popularity of MAGMA allows comparison of results generated from summary data with published results for

[7] https://ctg.cncr.nl/software/magma

benchmarking. The results may vary slightly as the methods for calculating gene values are slightly different when raw genotype data is available and there have been at least six versions (v1.0 to 1.7) of the software.

1.4.4 Pascal

Although not as widely cited as MAGMA, Pascal[228] (Pathway scoring algorithm) has been used in a number of important GWAS analyses [255],[265],[201], on occasion in conjunction with MAGMA[93]. Pascal processes summary GWAS statistics to generate gene level scores and perform gene set analysis. The authors report close concordance with VEGAS for genes with P values above 10^{-6} combined with a 100 fold reduction in speed. It is fast (29 hours on a test set for VEGAS, 30 minutes for Pascal), simple to use, written in Java and tested for Linux and MacOS. It can discriminate p values as low as 10^{-15} compared to VEGAS which, even with Monte Carlo runs increased to 10^8, could not estimate p below 10^{-6} with precision (at this level the VEGAS running time is 9 days). The z scores of the SNPs in a gene region are assumed to follow a multivariate normal distribution and the test statistics are the sum of the squared z score of gene SNPs (SOC) or the maximum value of the z score chi square (MOC). For the max score (MOC) method gene scores are calculated using Monte Carlo draws and if the p value is too low to be calculated in this way rectangular integration over a multivariate normal is used[142],[228][8]. The distribution function of the SOC statistic can be calculated numerically[97]. Linkage disequilibrium is accounted for using a reference panel (1000 Genomes) and in addition information on LD is informs the Gene Set Analysis score (see section 1.5.3)[9].

1.4.4.1 Gene level results available for GWA studies of intelligence or educational attainment

MAGMA [239] (see section 1.4.3) and VEGAS have both been used in a number of studies of intelligence and cognitive ability as gene level tests and gene set analysis (see table 1.2). The availability of all or portions of the gene level results in the supplementary material to these papers, allows validation and benchmarking for the gene level analyses used for the analyses in this thesis (see section 2.2.3).

[8] p 12

[9] a 50kb window was used in benchmarking

Phenotype	Study	Method	Year
Educational attainment	Rietveld et al.[358]	VEGAS	2013
Intelligence	Hill et al.[178]	VEGAS	2014
Intelligence	Sniekers et al.[383]	MAGMA	2017
Educational attainment	Lee et al.[237]	MAGMA	2018
Intelligence	Savage et al.[367]	MAGMA	2018
Intelligence	Davies et al. [96]	MAGMA	2016

Table 1.2: Gene level tests reported in secondary analysis of GWA studies of intelligence or educational ability

1.5 Gene set tests

Gene-set analysis, also known as pathway analysis, tests the association of a *set* of genes with a particular phenotype. The set of genes represents any natural grouping of genes often curated using prior biological knowledge. For example, gene sets are often made up of the components of metabolic pathways or the members of gene ontologies. In this thesis I will use communities in the PSP network discovered by community detection algorithms as gene sets. The combination of the use of community detection with gene set analysis that takes into account the full polygenic signal from GWAS is, to my knowledge, novel. In addition it provides a method for generating gene sets "based on external unbiased large-scale molecular data" as recommended by Lamparter et al.[228].

Gene set analysis methods were developed for the analysis of gene expression experiments[399],[283] and providing a biological explanation of the noisy differential expression of a large number of genes.

Testing the association of a set of genes increases power[423], allowing large numbers of genes of smaller aggregate effect to generate a detectable signal and allows experiments to be understood in terms of the actions of biological pathways[393]. A variety of approaches have been described and the statistical properties of those used to analyse GWA studies are reviewed by de Leeuw [100].

Gene set analysis (GSA) is now a well established part of the secondary analysis of genome wide association studies [199] in conditions such as schizophrenia[198] and multiple sclerosis [27] and in complex traits such as cognitive ability [178],[383],[180]

and body mass index[390].

GSA tests can be classified as self contained tests, which test the hypothesis that the set is associated with a trait more than would be expected by chance, or competitive tests which test whether the set is more associated with the phenotype than the rest of the genome [100]. Examples of competitive tests include the test provided as part of the MAGMA suite[239] and Gene Set Enrichment Analysis [393],[263].

An alternative classification differentiates between over representation analysis (ORA), Functional Class Scoring (FCS) and third generation methods [216][281]. Third generation methods incorporating network topology are addressed in section 1.6. ORA treats genes as discrete elements of a set of interest if they exceed a cut-off level of significance or some other property and tests whether this set of genes are found in a pathway more or less than expected by chance. Test statistics include Fisher's exact test based on the hypergeometric distribution, the chi square test or binomial test of proportion. The group of genes used to derive the expected number of genes in a pathway can be either the genome or a subset of possible genes such as those on a probe or those expressed in a particular location. This background set of genes is sometimes informally referred to as the "gene universe" in the bioinformatics literature [217] and is an important aspect of the analysis[357]. Over representation analysis has been widely used, both as a bioinformatics tool to investigate the properties of a group of genes and as an outcome measure for network biology approaches [144],[266],[26],[357]. Functional class scoring was developed [283] to overcome the limitation of discretisation by ranking genes rather than using a cutoff[460].

1.5.1 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis is a Functional Class Scoring method [393]) and the most widely cited pathway analysis method[460]. It has been used to analyse GWA studies and in particular to investigate enrichment of synaptic components for human cognitive differences[178]. It is provided as a desktop Java application or web interface by the Broad Institute along with a variety of gene sets[10].

The gene set enrichment analysis technique described by Subramanian et al. [393] tests the difference in the rank ordering of a set of genes between two phenotypes using the Kolmogorov Smirnov test. This is a statistical test that compares the similarity of two probability distributions. The test statistic is the supremum of the difference

[10] https://www.gsea-msigdb.org/gsea/index.jsp

between two cumulative probability distributions, in this case where gene set elements occur in a ranked gene list representing a phenotype and control. A significance value is calculated using phenotype permutation. In the case of the analysis of GWAS data the genes are permuted rather than the phenotypes[460]. An excellent intuitive introduction is supplied by Clark and Ma'ayan [75].

1.5.2 GSA with MAGMA

MAGMA also includes a gene set test which tests the self contained hypothesis that the genes in the set have a greater association with the phenotype than all other genes. The gene p values are transformed into z values having a roughly normal distribution and reflecting the strength of association of the gene with the phenotype.

The gene set test uses a generalised least squares model where errors are the product of the variance and the gene gene correlation matrix R which is calculated from reference data. Gene size, density and sample number are corrected for by treating them as covariates.

The gene sets used are often from the Molecular Signature Database (MSigDB)[248] used for GSEA containing 10,891 pathways (the default for the FUMA package [428], see for example [93]).

1.5.3 GSA with Pascal

Like MAGMA, Pascal splits gene scoring and GSA function and allows use of custom generated gene sets[228], and so is a suitable method for Gene Set Analysis of network communities. An additional attraction is that the source code is available [11].

The gene set analysis, in common with MAGMA, avoids the requirement for a cut-off and so can include contributions from low scoring genes. The pathway approach also accounts for the fact that genes in a pathway may be in LD with one another and treats multiple genes in LD with one another as a single entity for pathway scoring (Gene Fusion). Pathway scores are computed from gene scores using either the chi square method or the empirical method. The chi square method transforms gene p scores to a χ^2_1 and the sum of m gene scores in a pathway compared to the χ^2_m distribution. In the empirical method a pathway score is obtained from a Monte Carlo estimate of the sum of transformed gene scores compared with the score for randomly selected gene sets of the same size as the test set.

[11] https://github.com/dlampart/Pascal

1.6 Using network topology

Gene Set Analysis provides a method to test the hypothesis that genetic variation in certain synaptic proteome structures (communities) are associated with intelligence or educational attainment. This approach generates gene sets that respect the topology of pathways and provide an automatic method to generate gene sets from primary proteomic experiments (see further research in [228]). Next generation or third generation pathway analysis methods are those that incorporate network topology [216],[460] and I will briefly review these and some other popular network based approaches to genomic analysis below.

Most next generation or topology based methods have analysed differentially expressed genes within metabolic pathways or gene co-expression networks.

WGCNA (Weighted Gene Co-expression Network Analysis) examines networks of genes connected by co-expression [232]. Widely used, with solid foundations in network analysis it is less well suited for the analysis of protein protein interaction networks (PPIN) which are based on direct physical interaction. A number of other approaches based on gene expression are included in reviews of network based gene set analysis. For the same reason they are less appropriate for analysis of the synaptic proteome but, as they form the majority of the literature on network or topology based GSA, they are discussed in brief below.

Two recent comprehensive reviews cover topology based methods[301] and the related methods of subgraph identification[300]. These reviews also discuss the useability and reproducibility of these analysis packages which I will mention in brief and return to in the discussion. These reviews are explicitly *network* based and hence topology based avoiding the confusion between pathway analysis, gene set analysis and topology based analyses (see [281]). The review by T Nguyen et al. "Network-Based Approaches for pathway level analysis" [301] extents the extensive 2013 review by one of the co-authors [281] in which they note [281] that in an influential review of pathway analysis [216] which classified methods as over representation analysis, functional class scoring, and pathway topology or third generation, methods identified as topology based analysis simply used portions of pathways (such as KEGG[205] - which have a topology) as lists and did not incorporate any topological information. Mitrea[281] and Nguyen's reviews [301] are confined to methods where sets of candidate genes were scored or analysed or generated using network topology.

Nguyen et al.[301] reviewed the use of topological pathway GSA across 36 pack-

ages. The focus of the methods are on analyses of gene expression although methods addressing PPI are reviewed. These approaches in general take a list of differentially expressed genes (DEGs) and provide a list of potential networks involving these genes using a combination of gene expression and network features. 26 of 34 packages analyse directed networks. Directed networks better allow the analysis of small scale structures such as motifs or detecting pathway equivalence and are commonly used to represent gene regulatory networks or metabolic pathways such as KEGG[205]. Protein interaction networks are typically undirected [289]. They are characterised by widely ramifying structural interconnections that determine system function and many nodes are highly interconnected. These graph analysis tools serve a different purpose and typically require gene expression data. Of the undirected network packages two (Score PAGE and TAPPA) are specific to metabolites). The remaining six are PWEA, TopoGSA, TopologyGSA, DEGraph, GANRA and Enrichnet. Only TopoGSA and Enrichnet would be suitable to analyse a set of genes found in a GWAS study, however they would not be able to use the full genetic signal across these genes but would have to use a subset defined by a cutoff value.

TopoGSA [147] for a given gene set calculates network parameters such as degree, local clustering coefficient, shortest path length, node betweenness and eigenvector centrality along with a Fisher's exact test for pathways found in KEGG, Interpr, BioCarta or Gene Ontology and with network data provided by querying the STRING database[397]. An overall similarity score is provided based on the sum of ranks. For an explanation of network metrics see chapter 3.

It has an easy to use web interface but the user must decide on a cut off significance level and no use is made of the ranks of the genes in the set.

The authors do not record if they use direct interactions, or interactions based on techniques such as literature data mining, which are available in STRING[12][397]. Finally the web based interface makes it impractical for large scale analysis of GWAS data.

The Enrichnet package (2012) by the same author [148] makes greater use of network topology, calculating a distance metric between the gene sets and the STRING database, using a random walk with restart (Xd distance). This is compared to the result of Fisher's exact test. Again the approach identifies pathways by a re-weighting of Fisher's exact test using network topology. I found it hard to interpret the significance of its results and would not be able to use it to test a hypothesis about the synaptic

[12]Search Tool for Recurring Instances of Neighbouring Genes - originally a genomic context database

proteome.

It is relatively widely used with 87 citations. The use of random walk to make a sample of a particular network incorporate information about link density similar to community structure.

H Nguyen et al. [300] provide a comprehensive survey of identification of active subnetwork modules. The authors discuss topological modules citing Girvan and Newman (2002) [146] and their initial description is similar to modularity defined communities. Most of the methods however seek to find a group of genes closely related in the network to a set of differentially expressed genes (an induced sub graph). Of the 22 methods reviewed only DIAMOnD [144] and Enrichnet [148] are not constrained to use differential expression levels of genes in gene expression studies. The method I will describe is designed for the secondary analysis of GWA studies but can easily be adapted to look at differentially expressed genes but the converse is not true. The majority of these approaches try to find an induced subgraph of the network which optimises a cost or utility function based on gene expression or association with disease. The approach in this thesis is to find structures based on network theory and test if they are associated with a specific phenotype.

1.6.0.1 General points from these reviews

Mitrea[281] and Nguyen [301] draw attention to the fact that the results of methods that actually use network topology are difficult to assess and compare (see the discussion above of Enrichnet). The review points out that it is often impossible to reproduce results even with the cooperation of the authors. A large number of the packages are web based and inflexible in the networks that they use.

1.6.0.2 Other approaches

ToppGene [73] provides a guilt by association network function (ToppGenNet). A set of seed genes is drawn and a test set generated using measures of network centrality: Page rank, Hits with priors and k-step markov prioritisation from a neighbourhood distance selected by the user [72]. The advantage is it actually uses recognised network methods and a PPI database (although based on literature and primary proteomics (BIND, BioGRID and HPRD)). The disadvantage is that when the prioritised genes are tested by seeing if they have similar ontology, disease or other term enrichment to the disease (seed) genes. Its guilt by association approach means that for a PPI network

such as the synapse, there are a lot of associates and the analysis therefore reduces to the subsequent ORA. A random walk will also be affected by community structure as the walk will spend longer in communities. This is also the case with the enrichnet package[148]. WebGestalt [247] also offer network based analysis but this consists of over representation of portions of a variety of networks including PPI.

1.6.0.3 Other approaches

Barabasi [26] reviewed a number of methods that used network based statistics such as clustering coefficient to augment the analysis of gene expression and GWAS studies. Of particular interest is the use of a 'disease community' which I will return to in chapter 4. The difference with these approaches is they are based on heterogeneous lists of disease (see section 3.5 and 4.2.1)),

Pandey et al. report using eigenvector centrality of SNPs within a pathway to prioritise genes for pathway enrichment[321]. The edges within the network however refer to gene - gene interaction strength rather than network topology.

It is interesting to note that Mooney and Wilmot [282] suggest using gene sets derived from network analyses. In the course of describing the DiaMOND algorithm Ghiassian et al.[144] report comparing topological communities generated using the Louvain multilevel method[47] (section 4.3.3 and Markov Clustering method [411] (section 4.3.9) to a list of disease genes. They concluded that this approach was not useful and hence proposed the DIAMOnD, algorithm. They do not use a tissue specific network and use a variety of definitions of disease genes but do include GWA data from the GWA catalog [262] but use them as items in an over-representation analysis. It is the most similar approach in the literature to the methods in this thesis that I am aware of.

1.7 Plan

A network model of the post synaptic proteome can be created using proteomics experiments to identify its component parts and its interactions taken from protein-protein interaction databases. The network can then be clustered into cohesive units using community detection methods. These communities in turn can be used as gene sets to analyse GWA studies using Gene Set Analysis.

To my knowledge a systematic analysis using the full polygenic signal of genome

wide analysis of a complex trait using a network model of the proteome has not been conducted.

Differences in intelligence and educational attainment have been shown to be associated with variations in post synaptic genes. Previous synaptic network studies have suggested that particular communities are associated with different functions and disorders but have not used population genetic GWAS data.

My first hypothesis is that particular communities may be associated with differences in intelligence or educational attainment. The second is that genes encoding proteins that are key network nodes are associated with differences in intelligence or educational attainment.

To test if communities are associated with intelligence or educational attainment I will identify communities of interest using Gene Set Analysis of summary GWAS data using competitive Gene Set Analysis. I will then test whether the interesting communities are associated with the phenotype in an independent replication sample.

To do this I will obtain GWAS summary data of independent population samples for both intelligence and educational attainment.

To test whether key network genes are particularly associated with intelligence or educational attainment I will use the GWA summary data to calculate gene-level significance scores for the components of the network. I will then test whether six common measures of the importance of nodes in a network are correlated with these measures of significance.

Chapter 2

Network graph and GWAS samples

The previous chapter described how an analysis of the post synaptic proteome network might be used to augment the analysis of GWA studies of traits or diseases that involve the synapse. This in turn could show if network methods are useful in this context. Important nodes are tested for genetic variants associated with the phenotype (differences in intelligence or educational attainment) and GSA tests the association of groups of genes encoding proteins forming meso-scale network structures.

This chapter describes the generation of the post synaptic proteome network and how the discovery and replication GWAS samples were constructed for intelligence and educational attainment.

Analysis of genome wide gene association study (GWGAS) results from studies of intelligence and educational attainment are reported along with other tests of the integrity of the data analysis pipeline. Genes associated with intellectual disability are found to be enriched in studies of educational attainment but not of intelligence.

2.1 Network Graph of the Post Synaptic Proteome

The post-synaptic proteome (PSP) network was developed as part of an ongoing project to provide comprehensive maps of the synaptic proteome, synaptosome and presynaptic proteome. Generation of the network required the integration of the synaptic proteomic literature and protein-protein interaction databases and was the work over several years of members of the Armstrong groups and Simpson groups at the University of Edinburgh(see acknowledgements B.5). The network presented here represents a “freeze” of the evolving network data, albeit at a point where the rate of discovery of new network elements had significantly slowed [173].

An undirected, unweighted interaction network graph was used to represent the interactions of the post synaptic density and closely associated proteins (post synaptic proteome PSP). The nodes (vertices) in the network represent genes encoding proteins and were derived from curation of recent proteomic studies of the PSP. An edge occurs between genes if there is direct experimental evidence, recorded in protein-protein interaction databases, that their protein products interact. Figure 2.1 shows the network of the NRC-MASC complex from an earlier paper[338],[266],

Protein-protein interaction networks are built using evidence of molecular interaction curated in protein-protein interaction databases. Human protein-protein interaction databases include those that map to primary proteomic experiments: the Dictionary of Interacting Proteins (DIP) [365], MIntAct[314] and BioGRID *(Bio*logical*G*eneral *R*epository for *I*nteraction *D*atasets)[70]; those that contain computational predictions of interactions; and those that contain both, such as STRING[397].

Alternatively one can distinguish between primary databases reporting the results of proteomic experiments and meta databases that collate primary databases and may also include computational predictions[112]. Experimental evidence of protein-protein interaction comes from high throughput or low throughput studies. High throughput methods include affinity purification mass spectrometry and two-hybrid screening[112]. High throughput studies constitute the minority of publications curated in databases but account for most of the interactions. For example, in the IntAct database[175], 78% of publications report ten or fewer interactions, but these constitute only 4% of the interactions in the database [112]. Conversely, only 2% of studies report over 100 interactions, but these supply 77% of the interactions in the database [112].

To permit efficient data sharing a common database schema was developed by the Molecular Interactions workgroup (MI) of the Human Proteome Organisation (HUPO) Proteomics Standard Initiative (HUPO-PSI-MI) in 2002. Protein-protein interactions are documented using a controlled vocabulary and indexed by accession numbers[312]. Both a tab-delimited text format (MITAB) and stable XML format (PSI-MI XML) are available allowing a variety of tools to parse information from these databases [312]. Protein-protein interaction databases differ due to both differences in detection technology and in how experimental data is recorded (curation). The International Molecular Exchange (IMEx) consortium seeks to establish common curation rules and avoid redundant curation of interactions by coordination across members [312], [313].

Despite the use of a standard, machine-readable vocabulary, there is only moderate

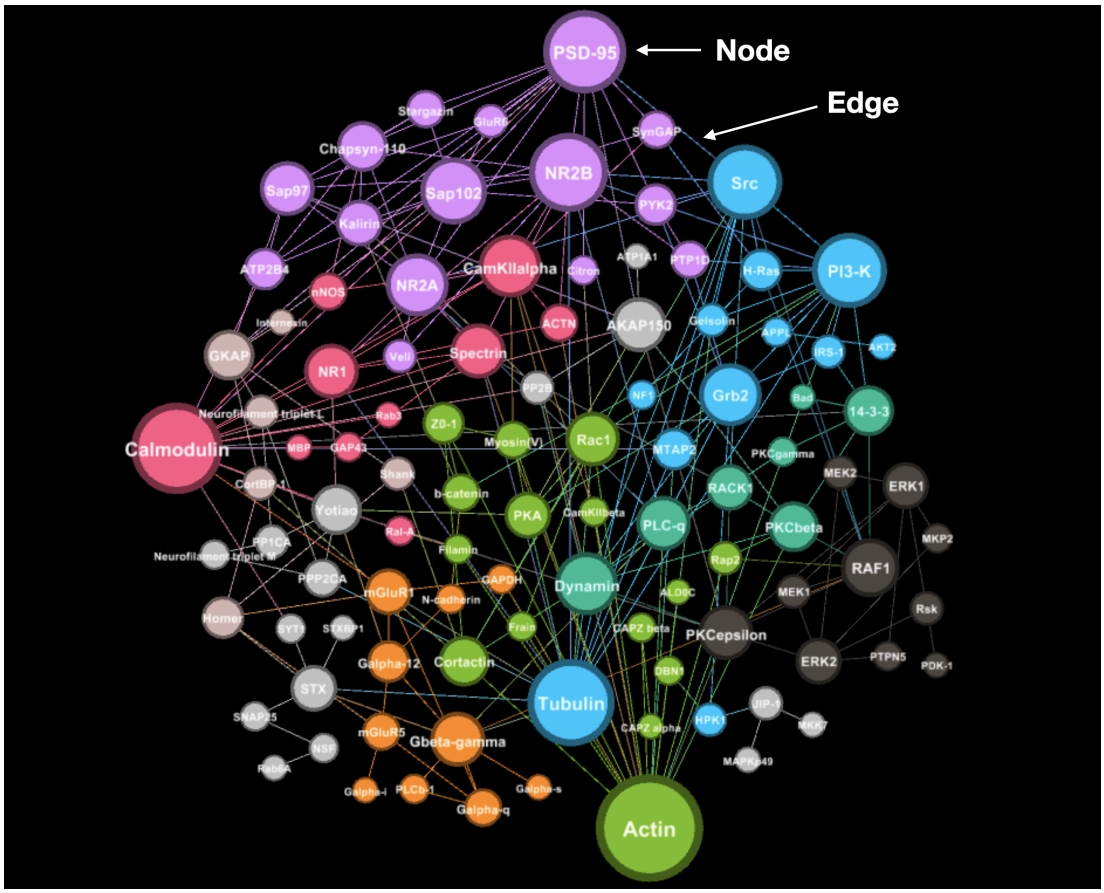


Figure 2.1: Network graph of the NRC MASC complex. Arrows show a node or vertex which correspond to a gene encoding NRC MASC protein and an edge or link between two nodes which shows that the proteins interact. The network is that described by Pocklington et al. [340] used in the paper by McLean[266] and the gml file is provided by the first author[1]. Graph visualisation in Gephi[30] with colours corresponding to community membership. Node size is proportional to degree (the number of edges connected to a node)

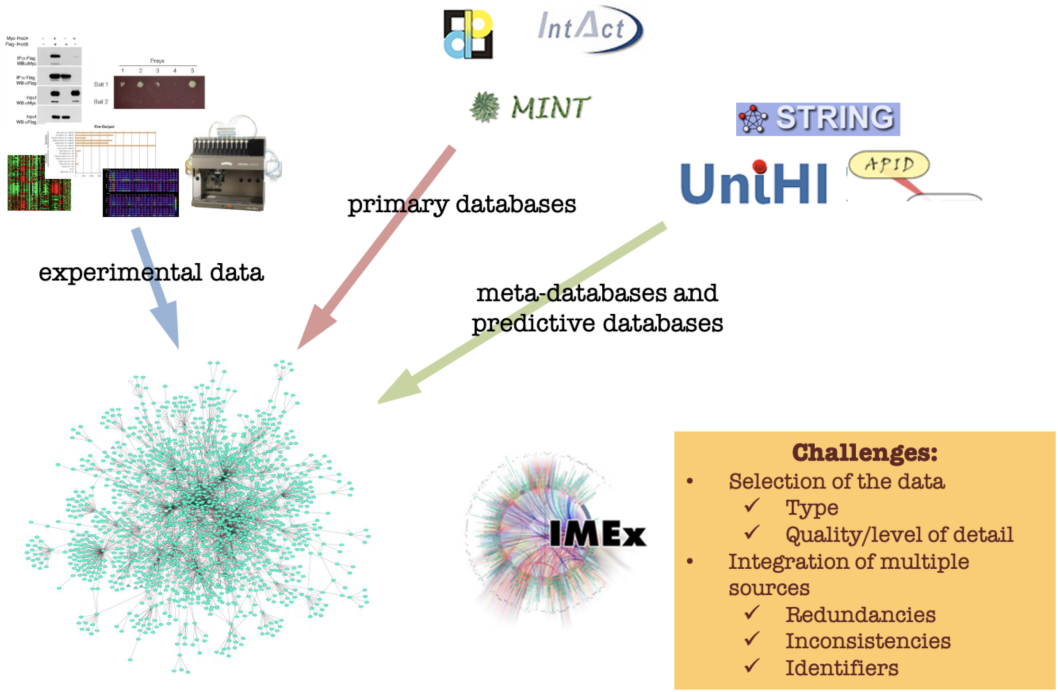


Figure 2.2: Sources of data for PPI network.Image credit creative commons https://www.ebi.ac.uk/training/online/courses/network-analysis-of-protein-interaction-data-an-introduction/building-and-analysing-ppins/sources-of-ppi-data/

	BioGRID	DIP	IntAct	MINT
Interactions	506,234	81,731	659,284	125,464
Species	62	834	980	611
BioGRID		13%	41%	18%
DIP	19%		32%	20%
IntAct	32%	10%		23%
MINT	48%	26%	100 %	

Table 2.1: Overlap in records between protein protein interaction databases. The table includes all databases used in the final PSP network with the addition of MINT for purposes of illustration. The table is asymmetric and should be read down the columns as the amount of the *row name* (for example *MINT*) contained in the *column name* (for this example *IntAct*; 100%). MINT has been included as an example and is a subset of IntACT. All of MINTs interactions are in IntAct so the row MINT, column IntAct; is 100% but the row IntAct, column MINT is 23% as the contents of MINT account for this amount of IntAct. Modified from figure 3 in Dong and Provart (2018) [112]. The date of collection of this data (July 2017) is within 6 months of the generation of the PSP graph.

overlap between protein-protein interaction databases[112][2](see table 2.1). Generating a comprehensive network model of a particular proteome therefore requires harmonising data from multiple databases[11]. Although meta databases combine information from primary databases, they may also include data from computational predictions and other less reliable data sources. Further data curation is necessary to construct a network with reliable, consistent interactions.

This is a significant task due to well-documented problems including the use of multiple identifiers, incomplete overlap in experiments, differences between matrix and spokes models and different confidence thresholds [240]. An overview of network curation is shown in figure 2.2 and is reviewed in depth by Dong and Provart [112].

[2]although some of this is due to non-redundant curation

2.1.1 Documenting interactions

Protein-protein interactions for the PSP network were obtained from the mining and harmonisation of data from the Database of Interacting Proteins (DIP) [437], IntAct database [314], BioGRID [70], and HIPPIE[368]. HPRD[213] was excluded as it did not comply to PSI-MI TAB standards at the time of compiling the network. HIPPIE was ultimately also excluded from further analysis due to a lack of specific interaction information (see Heil(2018)[173] chapter 4 for full details).

The mapping of identifiers and accession numbers across different databases is a key bioinformatics task[189]. These can be different identifiers for a particular object (such as HGNC Gene Symbols or Entrez Gene ID) or mapping identifiers between related objects (e.g. between a protein accession identifier and that of the encoding gene). The mapping of identifiers from UniprotKB to Entrez Gene ID was the most commonly used task in a protein identifier mapping service[189].

A bidirectional map was created between protein identifications (Uniprot accession numbers[408]), and gene identifiers (the GeneID provided by the Entrez Gene portal at the National Centre for Biotechnology Information - NCBI). NCBI Entrez Gene GeneIDs are unique integer identifiers and were used as the main identifier for all subsequent data tables and analyses [264]. For brevity I use Entrez GeneID synonymously with NCBI Entrez Gene GeneID. The widespread use of Entrez Gene identifiers in genomic studies and genomics software [189] makes them a suitable primary identifier for network nodes.

Protein-protein interactions were included if there was high quality evidence for direct physical interaction in humans. The combined interaction list was filtered to ensure all proteins had a human taxonomy ID of 9606. The details of interactions were recorded in PSI-MI TAB 2.5 format (MITAB25)[3][194],[173]. Where mirrored duplicates occurred (e.g. A-B and B-A) only one was retained. The resulting data contained 396,837 rows and 237,789 unique PPI. The interaction list was further filtered to generate the protein interactions represented by edges in the network graph. Protein-protein interactions were excluded if there was no direct experimental evidence of interaction documented by Proteomics Standard Initiative - Molecular Interaction ontology terms in the databases. PSI-MI is a structured ontology used to describe the evidence for protein-protein interactions.[194] PPIs based on “genetic interactions”

[3]MITAB25 is part of PSI-MI 2.5 `https://github.com/HUPO-PSI/miTab/blob/master/PSI-MITAB25Format.md`

(MI:0208) , “genetic interference” (MI:0254), “genetic interaction” (MI:0208)[4] and “co-localisation” were excluded (MI:0403). Sixty six interaction MI accession codes were identified representing evidence of direct interaction including one obsolete term, ‘physical interaction’ (MI:0218). The final filtered list contained 352,325 unique rows representing 205,461 binary interactions.

2.1.1.1 Proteomics data

Proteomic studies conducted between 2000-2017 of the human, mouse and rat synapse ($n = 39$) were curated into a database of 6,500 potential synaptic genes mapped to human Entrez Gene GeneIDs[173],[264]. Of these studies, seventeen examined the post synaptic density (PSD) and are shown in table 2.2. The human proteins and human orthologues of murine proteins found in these studies were used as a “parts list” for the PSD and associated interacting proteins expressed in the post synaptic area (PSP).

The nodes in the network used in this thesis are genes encoding proteins discovered in the proteomic studies in table 2.2 which have high quality experimental direct interaction data in humans.

The largest connected component of the resulting graph was used for community detection and in calculating network statistics (3,457 nodes, 30,498 edges). The details of the small number of excluded vertices not part of the largest connected component ($n = 16$) are in section 2.1.2 below.

The network was stored as a named edge list consisting of two columns. Each row represents an interacting protein pair and each protein is identified using the Entrez Gene Id of its encoding gene. Network metadata and interactions were stored as a graph modelling language (GML) file generated from this edge list[182]. GML is an ASCII based, hierarchical, lightweight format supported by igraph [89], Gephi [30] , NetworkX [165] and graph-tool [326]. A master copy of the original file supplied to me was stored on the University of Edinburgh informatics server. The original `gml` file was generated by Dr Colin McLean and also contains community detection data and is included in the supplementary material.

2.1.1.2 Value of tissue specific networks

The post synaptic proteome network is tissue specific; it is the network of interacting proteins found in the post synaptic area. I will, as part of overall quality control confirm

[4] which includes suppression (MI:0796 and synthetic MI:0794)

	First author	Year	Reference	Species	*n*
1	WALIKONIS	2000	Walikonis et al. (2000)[420]	rat	29
2	PENG	2004	Peng et al. (2004) [327]	rat	325
3	SATOH	2002	Satoh et al. (2002)[366]	mouse	45
4	YOSHIMURA	2004	Yoshimura et al. (2004)[448]	rat	435
5	FARR	2004	Farr et al. (2004)[127]	rat	71
6	TRINIDAD	2005	Trinidad et al. (2005) [404]	mouse	234
7	COLLINS	2006	Collins et al. (2006)[81]	mouse	717
8	DOSEMESI	2007	Dosemeci et al. (2007)[115]	rat	548
9	TRINIDAD	2008	Trinidad et al. (2008) [405]	mouse	2150
10	SELIMI	2009	Selimi et al. (2009)[375]	mouse	61
11	FERNANDEZ	2009	Fernández et al. (2009)	mouse	292
12	BAYES	2011	Bayés et al. (2011)[35]	human	1441
13	BAYES	2012	Bayés et al. (2012)[34]	mouse	1545
14	SCHWENK	2012	Schwenk et al. (2012)[372]	unknown	34
15	DISTLER PSD2*	2014	Distler et al. (2014)[110]	mouse	2092*
16	BAYES	2014	Bayés et al. (2014) [36]	human	1141
17	FOCKING	2016	Föcking et al. (2016)[133]	human	2026

Table 2.2: Primary proteomic studies contributing to the PSP Clean Published network adapted from Heil 2018[173] PhD p84 table 5.1.

in section 2.1.5 that these genes are differentially expressed in the CNS in humans given the expansion in the number of genes in the PSP network compared to previous models[266].

A tissue specific network is desirable as different tissues contain different proteins and therefore different nodes and edges in their proteome network. Some tissues are known to have very distinct gene expression patterns (e.g solid tissues and blood, CNS [269]) and their proteome network will also differ. In addition, it has been reported that disease genes tend to be expressed in specific tissues (and to be on the periphery of cellular networks)[26],[222]. Finally, the synaptic proteome *in particular* has a modular structure arising from the interaction of its protein components and giving rise to higher order structures and functions [338],[155] that are particular to the synapse.

Two examples help illustrate the issues. Ghiassian et al. (2015)[144] performed an analysis similar to the one presented here (see section 1.6.0.3) before going on to describe a method for identifying disease modules. They concluded that topological units were not associated with disease, but that disease results from loss of function of an area of a *global* interactome where a group of disease associated genes reside. The network used was general and composed of heterogenous elements in addition to protein-protein interactions.

By contrast a review of systems biology [322] gives the use of a tissue specific cardiac PPI to investigate long QT syndrome (LQTS) by Lundby et al.[259] as an example of the value of tissue specificity. The authors used a mouse cardiac protein-protein interaction network derived from High Performance Liquid Chromatography Tandem Mass Spectrometry (HPLC-MS/MS) of five seed proteins encoded by genes identified in a GWAS of LQTS to generate candidate genes later confirmed in zebra fish. These genes could not be identified using more general networks and significantly the affected genes were those in the *neighbourhood* of receptors and ion channels.

2.1.2 Connected component

All of the nodes in the parts list other than sixteen are connected. In the simplest random generative models of networks (Erdős - Rényi[122]) where nodes are connected with a uniform probability p, isolated nodes spontaneously aggregate to form a large connected component (connected in that all of its nodes can be reached via 'hops' along its edges) containing the majority of nodes above a critical value of p[289]. In real world networks, the proportion (S) of nodes in the largest component varies but

one component always dominates the others[289].

2.1.2.0.1 Methods The network containing all nodes, including those not in the largest connected component, was generated by Dr Colin McLean (section B.5) and stored as a gml file. Components were identified using the `components` command for igraph version 1.2.5[89] for R and the components not part of the largest component recorded[5].

igraph for R version 1.2.5 was used for the majority of graph processing tasks in this thesis[89]. R version 3.6.3[193] is used throughout and all statistical analyses are carried out in R unless otherwise stated. On occasion igraph for Python (v0.8; Python 3.6) was used, when methods from other Python packages were required. This is noted in the corresponding methods sections and associated scripts.

Nodes not in the connected component are excluded from further analysis, as community detection algorithms and most vertex statistics are not valid when there are unconnected components.

The composition of the excluded components was analysed to assess whether they contained a plurality of elements that might bias the subsequent analyses (see section 2.1.2.1).

2.1.2.0.2 Results In addition to the large connected component (n=3,457 nodes), 15 smaller components were found within the PSP graph; these were all single isolated components other than one component of size two. The proteins encoded by these genes appear in the "parts list" derived from proteomic studies but lack experimental evidence of direct interaction with the protein products of the genes in the large component of the PSP.

The sixteen genes not in the giant connected component are shown in table 2.3. The only component with two genes contained Oligodendrocyte Myelin Glycoprotein (OMG) and Reticulon 4 Receptor (RTN4R)[6].

These constitute 0.46% (S= 0.995; see[289] p305 for a comparison with other networks) of the genes identified in proteomic experiments (section 2.1.1.1) with high quality experimental interaction data. This is comparable to that reported for undirected metabolic network ($S = 0.996$) and high compared to earlier protein-protein interaction networks (n=2115,S=0.689)[289][7].

[5] `RProjects/network_graph/R/notLCC_gmt/genes_not_inLCC.R-GIT`

[6] member of growth cone pathway and implicated in schizophrenia

[7] p 305

2.1.2.1 Analysis of genes not in largest connected component

Gene ontology enrichment analysis of the genes ($n = 16$) not in the largest connected component was carried out to ensure they did not belong to a common system or component that would bias the analysis. Gene ontology analysis with ToppGene did not show any significant enrichment for terms compared to the rest of the PSP. Gene ontology enrichment analysis is widely used in the rest of this thesis and the methods introduced below.

GeneID	Gene Symbol	Gene Name	n
4974	OMG	oligodendrocyte myelin glycoprotein	2
65078	RTN4R	reticulon 4 receptor	2
98	ACYP2	acylphosphatase 2	1
166785	MMAA	metabolism of cobalamin associated A	1
549	AUH	AU RNA binding methylglutaconyl-CoA hydratase	1
55847	CISD1	CDGSH iron sulfur domain 1	1
23305	ACSL6	acyl-CoA synthetase long chain family member 6	1
26027	ACOT11	acyl-CoA thioesterase 11	1
7915	ALDH5A1	aldehyde dehydrogenase 5 family member A1	1
4685	NCAM2	neural cell adhesion molecule 2	1
5649	RELN	reelin	1
2741	GLRA1	glycine receptor alpha 1	1
152404	IGSF11	immunoglobulin superfamily member 11	1
27000	DNAJC2	DnaJ heat shock protein family (Hsp40) member C2	1
2686	GGT7	gamma-glutamyltransferase 7	1
55262	MAP11	microtubule associated protein 11	1

Table 2.3: Genes with cognate proteins in the PSP as defined by their presence in proteomic studies but that are not part of the giant connected component. These encode proteins that lack direct experimental evidence of interaction with the giant connected component. GeneID is Entrez Gene ID. n is the number of genes in the component.

2.1.3 Gene ontology analysis

A gene ontology provides a controlled, maintained vocabulary to describe sets of genes according to their Biological Function, Cellular Component and Molecular Function. It is widely used in high throughput genetics to describe the function of a set of genes[17]. Gene ontology terms are structured forming a directed acyclic graph that shows a hierarchical relationship with more specific child terms arising from parent terms (see figure 2.6).

Gene ontology over-representation analysis tests whether a set of genes are more commonly annotated to a specific term than would be expected by chance. Some difference in results is reported across different platforms and therefore more than one method is reported[357],[215].

2.1.3.0.1 Multiple testing Gene ontology terms overlap, particularly at different levels in the hierarchy, invalidating the assumption of independence made by multiple testing correction methods. The issue of multiple testing in Gene Ontology enrichment analysis is complex and beyond the scope of this thesis. Rhee et al. (2008) [357] provide a comprehensive summary and a more recent, technical analysis is found in Meijer et al (2016)[268]. A brief description will introduce the methods in this chapter and highlight potential issues common to pathway enrichment methods that use GO terms.

Rhee et al. (2008) [357] recommend the following approach to multiple testing:

'*An unbiased search for significant GO associations can be done with a bottom up approach as follows: for every leaf term, calculate p-values with the genes directly associated to it. If any term is significant do not propagate its genes above. This would provide the most specific node that is significant in that particular branch. If a term is not significant, propagate the annotations to its parent and recalculate with the parent term. The genes will propagate upwards until a significant node is found or until the root is reached. A careful analysis is still necessary to properly correct for multiple comparisons*.[357]'

To my knowledge there is no widely available implementation of this method designed for genomic analysis. The closest approximation being the methods described by Alexa et al.[10] and implemented in topGO[9]. Rhee et al.[357] argue that the Bonferroni correction is too conservative and that the Holm and FDR are better alternatives. A different method for correction is implemented in g:Profiler[352] which uses the sgs method to increase power using the interdependence of GO terms.

Rhee et al. recommend minimising the number of tests wherever possible[357]. One way to reduce the number of tests is to use a restricted or clipped Ontology. This can be challenging to implement and although a neural-immunology clipped ontology is available[141], I am not aware of any current application using it for gene ontology enrichment. The SynGO project[226] uses an ontology restricted to the synapse, with curated gene-term relationships supported by high quality experimental evidence.

2.1.3.0.1.1 elim method The *elim* and *weight* methods described by Alexa et al. (2006) [10] and implemented in the popular topGO R package[9] have the effect of decreasing the significance of very general GO terms (increase p) in comparison to more specific child terms. These methods start testing at child nodes and only move genes for testing to higher levels if they have not been scored.

This decreases the significance of very general terms and makes specific terms more significant. The overall effect is to increase raw p values but to do this most for general terms. It does not correspond exactly to multiple testing correction but does prioritise general terms and reduce the significance of general ones. The authors state that most multiple comparison methods are too strict given the connected structure of GO and they prefer to regard the *elim* results as already corrected [8].

2.1.3.0.1.2 Background set The number of terms that are expected to be found in a group of genes depends on the background set and this plays an essential role in calculating over representation. Some packages allow the use of custom background sets (ToppGene has very recently added this feature - section 2.1.4.2). If not indicated the background set is the default for that enrichment package usually the protein encoding genes in the genome.

2.1.4 Gene ontology enrichment methods

Some methods use direct data entry to a web interface, others an API, and some methods can be locally installed. The main advantage of web based ORA is regular updating of the ontologies and a common computational platform. To reduce transcription errors, whenever text entry of a list of genes is used on a web interface a script is used to write the list of genes directly to the clipboard.

2.1.4.1 PANTHER methods

Gene ontology ORA was carried out using the PANTHER classification system (protein annotation through evolutionary relationship) version 15.0[272] via the web-based user interface hosted at the Gene Ontology Consortium (GOC) website[8].

PANTHER provides information on gene function at a genomic scale, and as a member of the Gene Ontology Consortium, provides accurate annotation with regular updates. Results can be displayed showing their relation in the gene ontology hierarchy[272]. The authors recommend the use of the False Discovery Rate to correct for multiple comparisons because of the correlation between tests involving related annotations [272]. It can be accessed using an API or web application. It was initially used for tests where authoritative, accurate testing using an up to date ontology was required and where custom background sets for enrichment were required. The hierarchy view was also informative, however problems with the interface and gene registration caused me to switch to ToppGene once it became possible to use background sets in this application.

PANTHER 15.0 (released 2020/02/14)[271] was used to test sets of genes that were significant at MAGMA GWGAS for over representation of terms from each of the three Gene Ontology clades (Cellular Component, Biological Process and Molecular Function). The PANTHER reporting details are as follows: The PANTHER Overrepresentation Test used implementing Fisher's Exact test (released 2020/07/28). The Gene ontology database was DOI: 10.5281/zenodo.3873405 Released 2020-06-01. The Benjamini-Hochberg procedure [38] was used to calculate the False Discovery Rate (FDR) to account for multiple comparisons, and corrections were applied for each clade.

2.1.4.1.1 Annotations Despite using Entrez Gene ID integer values with the recommended string prefix "GeneID:" (PANTHER interprets integers as HGNC identifiers by default), several significant genes from the population samples were not identified using Ensembl and Entrez Gene ID.

Using PANTHER, there were identifiers that were not recognised using Ensembl ID, Entrez ID and GeneID. The most robust appeared to be gene symbol. I used translation from Entrez ID to Gene Symbol including synonyms until all genes were identified. In addition to the code documenting the data transformation [9], the gene mapping

[8] http://geneontology.org/

[9] RProjects/paper_xls_output/R/chapter_2/FUMA/edit_gene_table/EA2_

compatible with PANTHER is included in the supplementary materials.

PANTHER displays the hierarchy of the ontology terms that can be useful. Although the FDR values are only nominally significant the very specific term related to neurogenesis: positive regulation of collateral sprouting (FDR p 0.041) is prominent in comparison to the same data ordered by FDR shown in table 2.5.

2.1.4.2 ToppGene

Toppgene is a web based gene ontology enrichment service that also supports murine and human phenotypes and disease enrichment[73]. It has a regularly updated database (see [401] for importance of regular updates). ToppGene uses Entrez ID as its default identifier which made it well suited for analysis of the PSP network.

The ToppGene API was used for some translation of identifiers and where this has been done it is noted in the manuscript. Custom background gene sets can only be used with the web interface so this was used when required in chapters 3 and 4.The output tables for significant ToppGene analyses including terms searched are included as tab separated text files in the supplementary materials.Database update 28-07-20. Gene Ontology Biological Process 16,465 annotations 20,148 genes. Cellular component 2,043 annotations, 20,584 genes, molecular function 4,745 Genes 20,584 with GO ontology from PANTHER DB.

2.1.4.3 FUMA Gene2Func package

FUMA (Functional Mapping and Annotation of Genome-Wide Association Studies)[10] is a comprehensive web based annotation and analysis package[428] that has been widely used in the secondary analysis of GWA studies of intelligence[180],[367]. It comprises two modules: SNP2Gene takes summary SNP data as an input and implements gene level GWGAS using MAGMA, and annotates significant genes, and the Gene2Func module which takes a list of significant genes and provides over representation analysis.

I have used the GTEx expression enrichment analysis from the Gene2Func module to check that the PSP genes are expressed in the CNS (FUMA v1.3.6 was used throughout). This differs from the results and methods presented in section 2.1.5.2 in using a predetermined set of differentially expressed genes and testing if a set of genes

`FUMA2clipboard.R`

[10] `fuma.ctglab.nl`

are over-represented in these using the hypergeometric test and Bonferroni correction for multiple comparison.

This provides a quick, graphical representation of the differences in tissue expression of genes in the PSP and genome wide significant genes at GWGAS in the discovery and replication samples.

I have also used the Gene Set Analysis in the Gene2Func module which uses hypergeometric tests to assess over representation of terms. Data sets are provided, including gene ontology terms, Wiki Pathways and GWAS catalogue with Bonferroni correction per category. I have used the GWAS catalogue set enrichment to ensure that genes excluded from analysis are not potential sources of bias (ie reported in intelligence GWA or other neuro-cognitive phenotypes).

2.1.4.4 g:Profiler

g:Profiler, is a Gene Ontology Consortium (GOC) backed Gene Ontology functional enrichment package. It has a recently updated web interface and an R API. Over representation of ontology terms is tested using the cumulative hypergeometric test[352]. The g:SCS (set counts and sizes) method, which takes into account the overlapping between gene sets based on GO terms[356], is used to correct for multiple comparisons. A Manhattan like plot of over representation is provided (figure 2.17), and a comprehensive translation service that can be accessed from the R API. Although it includes other data from KEGG and wiki pathways, I have limited the analysis to GO enrichment terms (these are tested using ToppGene).

Functional enrichment was carried out using g:Profiler version e100_eg47_p14_7733820 (database updated on 07/07/2020). Multiple testing was controlled using g:SCS method applying a significance threshold of 0.05[11] and not displaying under-representation.

2.1.4.4.1 Results ORA of genes not in connected component There is no significant enrichment of pathways or ontology terms for the genes that are not in the connected component using ToppGene for ORA analysis and the PSP genes as background.

There is no evidence of any enrichment of an association between genetic variants in the set of 16 genes not in the largest connected component, and differences in intelligence or educational attainment in any of the discovery or replication samples (see

[11] recommended citation format from website

Sample	β	p
$\text{Intelligence}_{\text{Replication}}$	0.30	0.10
$\text{Education}_{\text{Replication}}$	0.11	0.32
$\text{Intelligence}_{\text{Discovery}}$	-0.09	0.64
$\text{Education}_{\text{Discovery}}$	0.21	0.21

Table 2.4: MAGMA GSA enrichment of the set of genes not in LCC n = 16. Testing the competitive hypothesis that genetic variants mapped to these genes are more associated the Intelligence or Educational Attainment phenotypes than the rest of the genome

RProjects/chapter2/R/LCC/write_not_lcc_gmt_out.R

section 2.2.2) using MAGMA Gene Set Analysis (see table 2.4; details on methods for MAGMA GSA in section 2.5).

2.1.5 GTEx profile of PSP genes

2.1.5.1 Introduction to GTEx

The gene tissue expression project (GTEx) is a repository of gene expression derived from RNA seq in over 1,000 individuals from 54 tissue areas[241]. I have used this to confirm that the PSP genes are highly expressed in the brain, if there are significant regional differences in expression that might affect the analysis, and that the expression of PSP genes is as expected significantly greater than of those not in the PSP.

2.1.5.2 GTEx Methods

Gene expression data (GTEx version 8 - dbGaP Accession phs000424.v8.p2) were obtained from the GTEx Portal on 07/03/2020. Median gene expression counts normalised by Transcripts Per Million (TPM) were downloaded from the GTEx Portal website[12],[84],[82]. The median gene expression values were calculated by the GTEx authors from the file `GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz`.

Two identifiers are provided for each gene, Ensembl Gene ID (ENSG) and HUGO gene nomenclature committee (HGNC) Gene Symbol[161]. The ToppGene annotation service [73] was used to provide Entrez gene GeneIDs from Gene Symbol to allow the results to be combined with the synaptic network and cohort study results[13],[14].

Fifty-six thousand two hundred rows were downloaded from the GTEx Portal indexed by Ensembl id, which included a version suffix. Each Ensembl ID, including version number was unique; 56,156 unique ID had only one Ensembl version.

One hundred ninety-nine duplicated gene names were identified in 1,608 duplicated GTEx entries. One hundred fourteen had a single duplicate (a total of Two Ensembl Id to one Gene Name). Seven hundred thirty-seven duplicates were identified as Y RNA (45.8%); small non-coding RNA [328]. 166 were metazoa signal recognition particles (SRP)[209] (10.3%).

54,592 unique Gene Symbols could be associated with unique Ensembl id (97.1%)[15]. This is a lot more than the number of protein-encoding genes, and to exclude non-coding elements and reduce the data size and I excluded those not amongst the 18,390

[12] `https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz`

[13] `https://toppgene.cchmc.org/API/lookup`

[14] `RProjects/gtex/R/exploratory_gtex/load_toppgene_annotations.R`

[15] `RProjects/gtex/R/exploratory_gtex/first_load_gtex.R`

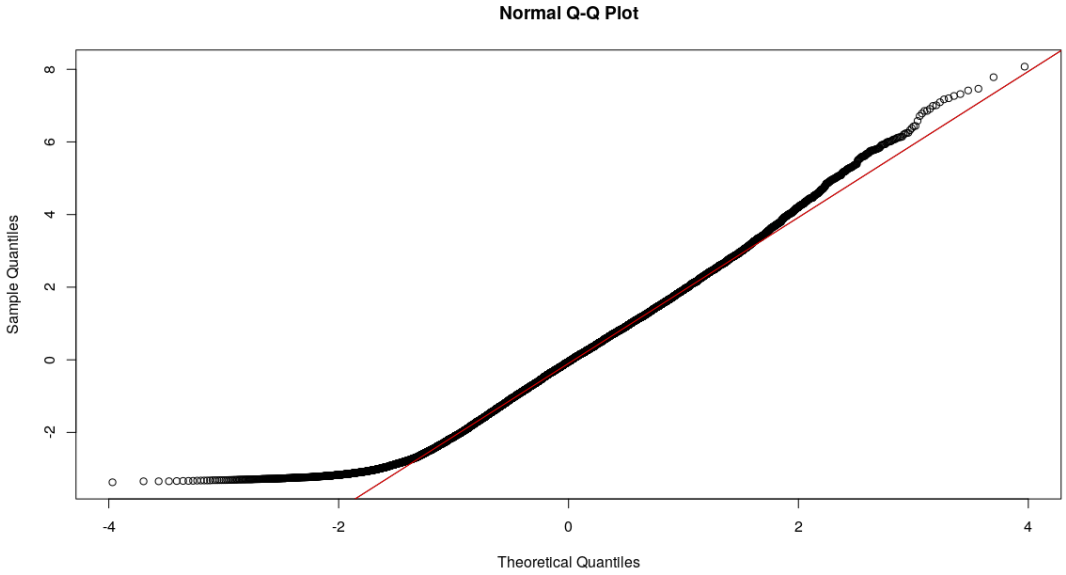


Figure 2.3: QQ plot log2 transformed and mean centred gene expression fronal cortex all genes passing quality control 13,787 genes

genes that have gene-level results in at least one of four population samples calculated in MAGMA. 17,584 unique Entrez ID were identified from the remaining elements and used for subsequent analysis.

Entries were retained if the TPM count was greater than 0.1 in approximately 80% (79.6%) of samples. TPM counts were then log_2 transformed with a pseudocount of 1 and zero centred [383],[398],[269], see [453][16].

The QQ plot of log2 frontal cortex expression is shown for illustration in figure 2.3. The plot does not seem to deviate significantly from normal so Pearson's product moment correlation was calculated across brain regions. In the absence of checks of all brain regions, I carried out a non-parametric test (Wilcoxon rank sum test) to see if expression was higher in the PSP than other regions (multiple testing correction Bonferroni using R - see figure 2.4).

2.1.5.3 Results Brain Expression of PSP genes

3,055 genes with GTEx data passing the above quality control were available for the PSP and 10,732 non synaptic genes.

Expression is significantly higher in the PSP genes across all 12 brain regions ($p < 2.2 \times 10^{-16}$ Wilcoxon rank sum test;spinal cervical1 has been removed as is known

[16] `RProjects/gtex/R/exploratory_gtex/df_with_toppgene_log_normalise_plots.R`

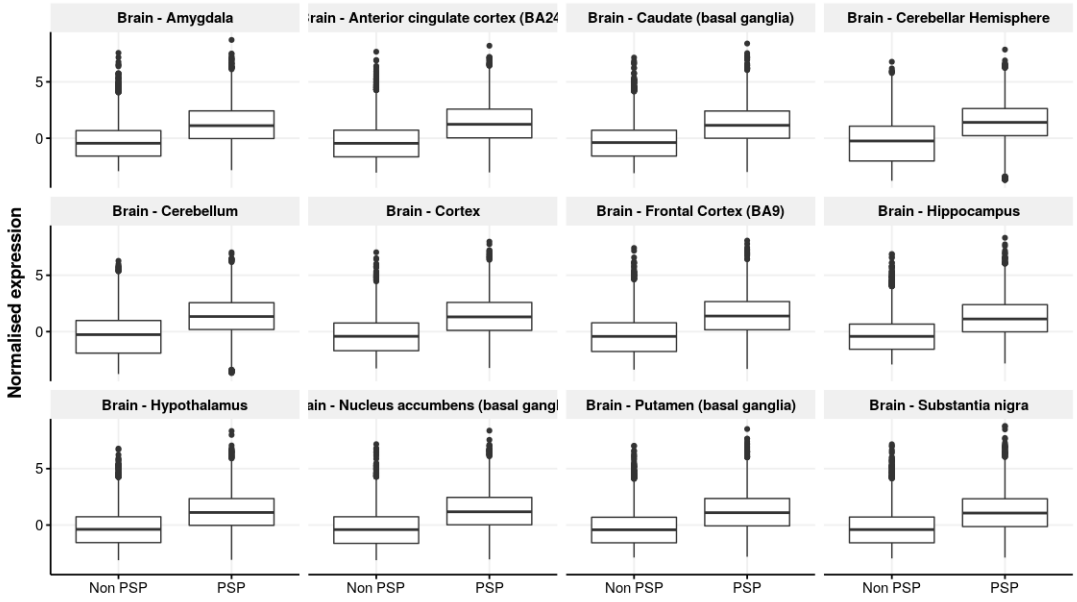


Figure 2.4: Box plot of log2 transformed and mean centred gene expression of PSP genes versus non PSP genes across 12 brain regions. p $< 2 \times 10^{-16}$ all comparisons. Wilcoxon test.

RProjects/gtex/R/graph_gtex/multi_boxplot.R

to be an outlier; see figure 2.4). There is strong correlation between gene expression counts in all brain regions(table 2.5). Expression across brain regions was closely correlated as reported in [84]. The results above indicate that the PSP genes have higher expression levels than other genes and that the expanded PSP continues to show increased CNS expression.

An alternate way to calculate brain expression of these genes is implemented in FUMA using GTEx v 8 data. This is to create a list of differentially expressed genes and see if a test set is over represented amongst these. Results for the complete PSP are shown in figure 2.19, the PSP and Consensus PSD are compared in figure 2.20, and the expression in the PSP, Consensus PSD, Non PSP genes and those genes found in the PSP but not Consensus PSD are shown in figure 2.23.

2.1.5.4 Correlation of GTEx expression across brain

Table 2.5 shows the correlation of expression of PSP genes between brain regions (Pearson product moment correlation). Anterior cingulate cortex (mean correlation 0.94), was the most correlated brain region and cerebellum the least (mean correlation r=0.85). The minimum pairwise correlation was found between the cerebellar hemi-

sphere and substantia nigra (0.792)[17]. The highest correlation is between caudate and putamen (r =0.996), which are both in the basal ganglia. The overall correlation was high (median 0.948, Mean 0.914).

PSP gene expression is high in the brain and correlated with the lowest correlations found in the cerebellar hemisphere as reported by Mele et al.[269], suggesting that the PSP group is not markedly different from CNS expressed genes overall.

Expression for genes in the PSP but not in the consensus PSD used by Hill et al[178] (n=1323) is increased compared to all other genes (which includes the consensus genes). The consensus genes are more highly expressed than the PSP genes not in the consensus set but the effect is much more modest (figure 2.20)[18].

	Amy	ACC	Cau	Ce H	Cer	Cor	FC	Hip	Hyp	NA	Put	SN
Amygdala	1.00	0.98	0.97	0.80	0.81	0.96	0.95	0.99	0.97	0.97	0.97	0.95
Anterior cingulate cortex	0.98	1.00	0.96	0.83	0.83	0.99	0.99	0.97	0.96	0.97	0.95	0.91
Caudate*	0.97	0.96	1.00	0.80	0.81	0.95	0.94	0.97	0.95	0.99	1.00	0.93
Cerebellar Hemisphere	0.80	0.83	0.80	1.00	0.99	0.84	0.84	0.81	0.83	0.81	0.79	0.78
Cerebellum	0.81	0.83	0.81	0.99	1.00	0.86	0.84	0.82	0.84	0.82	0.80	0.79
Cortex	0.96	0.99	0.95	0.84	0.86	1.00	0.99	0.96	0.93	0.95	0.94	0.88
Frontal Cortex	0.95	0.99	0.94	0.84	0.84	0.99	1.00	0.95	0.94	0.95	0.93	0.88
Hippocampus	0.99	0.97	0.97	0.81	0.82	0.96	0.95	1.00	0.97	0.96	0.97	0.95
Hypothalamus	0.97	0.96	0.95	0.83	0.84	0.93	0.94	0.97	1.00	0.95	0.95	0.97
Nucleus accumbens*	0.97	0.97	0.99	0.81	0.82	0.95	0.95	0.96	0.95	1.00	0.98	0.91
Putamen*	0.97	0.95	1.00	0.79	0.80	0.94	0.93	0.97	0.95	0.98	1.00	0.94
Substantia nigra	0.95	0.91	0.93	0.78	0.79	0.88	0.88	0.95	0.97	0.91	0.94	1.00

Table 2.5: Correlation of PSP gene expression across brain regions. Expression values log_2 transformed with pseudocount of 1. n=3055 genes passing quality control. *=basal ganglia. Frontal cortex = Brain area 9, Anterior cingulate cortex = 24.

RProjects/gtex/R/graph_gtex/multi_boxplot.R

2.2 Genome Wide Association Studies

Previous network analyses have shown that synaptic topological models are differentially enriched for disease and gene ontology terms. There is also evidence that central nodes may be more likely to be disease nodes. These analyses have not however made

[17] RProjects/gtex/R/graph_gtex/gtex_correlation_table.R
[18] RProjects/gtex/R/graph_gtex/gtex_correlation_table_hPSD_too.R

use of the complex polygenetic signal available in genome wide association studies. Chapter 1 introduced the reasons for using intelligence as a phenotype.

To test the association between synaptic proteome network properties and structures and population genetic variation associated with intelligence, we will need to calculate a statistic measuring the association of genetic variants with the phenotype for each gene that is represented by a node in the network. These values will be calculated using MAGMA in four samples, two of intelligence and two of educational attainment[239].

2.2.1 Population genetic samples

To investigate whether topological communities are enriched with genetic variants associated with cognitive ability and educational attainment, and to reduce type II error, I will use a discovery and replication cohort design (see section 5.1.1 and figure 5.1).

For consistency I will refer throughout the text to the GWA study of genetic variants associated with intelligence in UK Biobank[65] as $\text{Intelligence}_{\text{Discovery}}$. and the Phase II Educational attainment GWA study conducted using UK Biobank as $\text{Education}_{\text{Discovery}}$.

We refer to the meta-analytic GWAS of intelligence carried out by Sniekers et al. from the Complex Trait Genetics (CTG) lab as $\text{Intelligence}_{\text{Replication}}$ [383]. The meta-analysis of GWA studies of educational attainment excluding UK Biobank and 23andMe by Okbay et al.[309] (often called EA2 - Educational Attainment 2) is referred to as $\text{Education}_{\text{Replication}}$. Details of the generation and composition of these samples are found in section 2.2.2.

Two further samples became available after the completion of the primary analysis. These are the largest intelligence and education cohorts to date [367], [237] but overlap with the samples above. They will be used to check the findings from earlier sections in chapter 5.2.

2.2.2 Samples

The information in this section was prepared by Dr Hill (section B.5) for a manuscript we are preparing. Dr Hill calculated the χ^2 estimates of heritability. I have repeated these calculations using LD hub,[19] a web interface for LD score regression [455].

[19]LD hub `http://ldsc.broadinstitute.org/ldhub/` code to prepare files for LD hub at `RProjects/format_hub/R/format_summary_hub.R`

The $\text{Intelligence}_{\text{Discovery}}$ sample was formed using the "fluid" cognitive test from UK Biobank; verbal-numerical reasoning (VNR). The VNR test contains 13 multiple choice items, six of which are verbal with the remaining seven being numerical in nature. Each individual's VNR score was the sum of correct responses attained within a two-minute time period. This test has a large genetic correlation with intelligence scores derived from psycho-metrically validated cognitive tests [180]. A total of 120,934 participants (mean χ^2=1.52) were available for analysis who did not contribute data to the $\text{Intelligence}_{\text{Replication}}$[383] data set. A full description of how data from these participants was analysed can be found in Hill et al. [180].

UK Biobank was also used as the $\text{Education}_{\text{Discovery}}$ sample. Summary statistics were taken from Hill et al. on 273,274 participants (mean χ^2= 1.67)[180]. Education was defined as a binary choice of whether or not the individual had a University/College degree. A full description of how these data were analysed can be found in Hill et al. [180].

The $\text{Intelligence}_{\text{Replication}}$ data set was formed using summary statistics obtained from the Sniekers et al. GWAS meta-analysis of intelligence (n = 78,308, mean χ^2= 1.30). [383] An $\text{Education}_{\text{Replication}}$ data set was formed using summary GWAS data based on the the GWAS meta-analysis by Okbay et al (n = 217,569, mean χ^2= 1.43) with UK Biobank participants excluded. [309]. The summary statistics with UK Biobank removed were kindly provided by Dr Okbay and colleagues who generated the new summary statistics. (See tables 2.9 and 2.11). For genotype, meta-analysis methods and association analyses see the publications reporting the original GWAS[383],[309],[180].

2.2.3 MAGMA Gene level statistics: Validation

To ensure the accuracy of the analysis of the gene level significance statistic I compared the gene level scores obtained using MAGMA to those reported in available studies[239]. Potential sources of variation or bias include variations in command line options for software such as MAGMA, different calculation of gene scores for raw genotype and summary data (see section 1.4.3) and different software versions, different reference populations to account for LD and variation in low level libraries across platforms[218]. There were no published papers available containing a MAGMA gene-level analysis of the $\text{Intelligence}_{\text{Discovery}}$ sample or $\text{Education}_{\text{Discovery}}$ as they were generated for this analysis.

The study by Okbay et al. (EA2)[309], providing the $\text{Education}_{\text{Replication}}$ sample did not report MAGMA gene scores, and differs from the $\text{Education}_{\text{Replication}}$ sample by including UK Biobank Educational attainment data. The only published reports of MAGMA gene scores from any of the exact discovery and replication samples was, therefore, the $\text{Intelligence}_{\text{Discovery}}$ sample [383][20].

Summary data has recently been made available from the current largest GWAS of intelligence to date by Savage et al. [367] ($n = 269,867$) this reports MAGMA Gene level analysis results although only for significant genes[21].

The pairwise correlation between the gene scores reported in these studies and the results I calculated using MAGMA and their summary statistics are shown in table 2.6. Magma 1.06b for MacOS was used with 1000 Genomes Phase 3 data to account for linkage disequilibrium. NCBI 37 was used for gene boundaries. Summary GWAS statistics were used to calculate gene scores using default settings. No gene window was used.

Study	n	t	d.f.	p	Pearson r	CI
Savage (2018) [367]	269,687	122.26	352	$< 2.2 \times 10^{-16}$	0.99	0.986-0.991
Sniekers (2017)[383]	78,308	466.41	18246	$< 2.2 \times 10^{-16}$	0.96	0.959-0.962

Table 2.6: Correlation between MAGMA gene scores obtained from summary statistics and those published in supplementary material to papers. CI = Confidence interval

2.2.3.1 Results

The correlation coefficient varied in the range 0.96-0.99 despite slight differences in software version, gene windows and the potential use of genotyped versus summary data. I conclude that apart from small variants that there is no systematic inaccuracy in the calculation of MAGMA gene scores using the plaform and pipeline.

2.2.4 Synaptic genes absent from NCBI 37

All of the primary GWA studies from which summary statistics were derived used NCBI 37 to provide genetic co-ordinates. Seventeen PSP genes were not in the refer-

[20] supplementary table 8 at https://static-content.springer.com/esm/art%3A10.1038%2Fng.3869/MediaObjects/41588_2017_BFng3869_MOESM2_ESM.xlsx)

[21] table S15 in supplementary materials at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6411041/bin/NIHMS1006639-supplement-Supplementary_Tables.xlsx)

ence NCBI 37 boundaries text file used by MAGMA to assign SNPs to genes. Files with NCBI 36,37 and 38 gene boundaries are supplied on the MAGMA project website, five of the genes not in the NCBI 37 file are found in NCBI 38 (table 2.7[22]).

Gene ID	Symbol	Chr	Gene type	Full name from nomenclature authority
1804	DPP6	7	protein-coding	dipeptidyl peptidase like 6
3635	INPP5D	2	protein-coding	inositol polyphosphate-5-phosphatase D
3800	KIF5C	2	protein-coding	kinesin family member 5C
9554	SEC22B	1	protein-coding	SEC22 homolog B, vesicle trafficking protein
23380	SRGAP2	1	protein-coding	SLIT-ROBO Rho GTPase activating protein 2

Table 2.7: Synaptic PSP genes found in NCBI 38 gene boundary file supplied with MAGMA but not in NCBI 37 file. Gene ID = Entrez Gene Gene ID, chr = chromosome number.

Ten PSP genes are missing from both NCBI 37 and NCBI 38 gene boundary files but are annotated in the *Homo sapiens* gene information at Entrez Gene[23] (table 2.8).

Gene ID	Symbol	Chr	Gene type	Full name from nomenclature authority
293	SLC25A6	X|Y	protein-coding	solute carrier family 25 member 6
3493	IGHA1	14	other	immunoglobulin heavy constant alpha 1
3500	IGHG1	14	other	immunoglobulin heavy constant gamma 1
3502	IGHG3	14	other	immunoglobulin heavy constant gamma 3
3503	IGHG4	14	other	immunoglobulin heavy constant gamma 4
3507	IGHM	14	other	immunoglobulin heavy constant mu
3514	IGKC	2	other	immunoglobulin kappa constant
4509	ATP8	MT	protein-coding	mitochondrially encoded ATP synthase 8
4513	COX2	MT	protein-coding	mitochondrially encoded cytochrome c oxidase II
4538	ND4	MT	protein-coding	mitochondrially encoded NADH dehydrogenase 4

Table 2.8: PSP genes ($n = 10$) not in MAGMA n NCBI gene boundary files (NCBI36,37,38) annotated in Entrez Gene *Homo sapiens* gene info files. GeneID= NCBI Entrez Gene GeneID, Chr = Chromosome number.

https://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz

[22] RProjects/annotations_phd/R/missing_NCBI_genes.R

[23] https://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz

Two further pseudo genes were identified using the ToppGene API[73]: RAB1C, Entrez id: 441400, member RAS oncogene family pseudogene, UniProt Q92928, Protein transport. Probably involved in vesicular traffic (By similarity); POTEKP[24], POTE ankyrin domain family member K, Entrez id:440915, pseudogene[40], Uniprot Putative beta-actin-like protein 3, Q9BYX7.

Gene ontology enrichment analysis using ToppFun (part of the ToppGene suite) (section 2.1.4.2) did not find any phenotype or ontology suggestive of being associated with cognition (GO:0034987, immunoglobulin receptor binding, p Bonferroni 6.224×10^{-9}, six terms; GO:0097478, leaflet of membrane bilayer, Bonferroni p=1.296×10^{-3}, six terms).

The potentially concerning gene is SRGAP2, Entrez 23380,SLIT-ROBO Rho GTPase activating protein 2 which is associated with the cellular component dendritic spine head GO:0044327 and associated with neural morphogenesis[162]. The gene list was also analysed using the Gene2Func interface in FUMA [428] (see section 2.1.4.3) and no enriched gene sets or hits in GWAS catalog[431] were found.

While it would be possible to make a custom annotation file for these genes this would complicate our analysis and cause it to differ methodologically from the rest of the literature using MAGMA and addressing the genetics of intelligence and educational attainment.

2.3 MAGMA gene level statistic methods

MAGMA version 1.06 for MacOS was used to calculate gene level statistics from summary data and 1000 Genome European Ancestry Phase 3 data[25] was used to account for linkage disequilibrium. Gene boundaries were defined by the Genome Reference Consortium Human Build 37 (NCBI 37.3) data file provided on the MAGMA website[26].

Output logs and output files are included in the supplemental material. Command line arguments were supplied using a bash shell script and default settings were used for summary GWAS statistic data (SNPwise sum)[27]. No gene window was used around gene boundaries (see section 2.3.0.1). Where the number of individuals tested per SNP was available in the summary data this was supplied as a command line argument -

[24] POTE expressed in Prostate ovary testis and placenta

[25] `https://ctg.cncr.nl/software/MAGMA/ref_data/g1000_eur.zip`

[26] `https://ctg.cncr.nl/software/MAGMA/aux_files/NCBI37.3.zip`

[27] `Programs/copy_paperMagma_draft-master/magma_script.sh`

otherwise the overall sample size was used. SNPs that were analysed in less than 50 individuals were discarded (default setting).

2.3.0.1 Gene Windows

In order to include gene regulatory elements in gene level analysis a number of "windows" around the gene boundary have been used. VEGAS [254] used a set 50kB boundary and this is incorporated in studies that used this method (e.g Hill et al. (2014)[178])

Reported window sizes vary. Zhao and Nyholt [453] in a study of gene level statistics across psychiatric disorders using the GATES method [246] used a 15kb window reasoning that 90% [332] of eQTL are found in this region.

The predominant practice at present appears to be either not use a gene window or only to use a small one (for example 5kb as reported in Savage et al (2018. [367]). In the absence of a widely agreed upon window size I have not used one believing this to be the simplest and most conservative option.

2.3.1 Correlation of MAGMA GWGAS between samples

Educational attainment and Intelligence are genetically correlated[337] (see table 2.11). It is likely that gene level scores in samples between and within these phenotypes are also correlated although this should be less than the overall SNP correlation because of inter-genic SNPs and a loss of information in the aggregation of SNPs to genes. It is worthwhile checking that the value is not extreme or unexpected (for example higher between low powered samples with different phenotypes) as part of overall quality control.

Correlation was calculated between the -log10 transform of gene level p values between samples using Pearsons product moment correlation test in R. Where a gene did not have a p value in a study this pairwise comparison was omitted. The correlation is greatest between the two educational attainment samples (0.506; CI 0.495-0.516, t = 78.473, df = 17928, p-value $< 2.2 \times 10^{-16}$) (table 2.10). For each sample other than $\text{Intelligence}_{\text{Discovery}}$ the correlation is highest with the matching phenotype.

The two discovery samples are similar in composition and the $\text{Intelligence}_{\text{Replication}}$ sample is likely to have least power (lowest sample size, meta-analysis and lowest χ^2 on testing heritability using LD score regression (section 2.2.2). $\text{Intelligence}_{\text{Replication}}$ is most highly correlated with $\text{Intelligence}_{\text{Discovery}}$ (0.378) despite its generally lower

Phenotype	sample	n	mean χ^2	LD intercept (SE)	Reference
Intelligence	Discovery	120,934	1.53	1.02 (0.01)	NA
	Replication	78,308	1.30	1.02 (0.01)	Sniekers et al.(2017)[383]
Education	Discovery	273,274	1.67	1.042 (0.01)	NA
	Replication	217,569	1.43	1.02 (0.01)	Okbay et al.(2016)[309]

Table 2.9: References and heritability of each data set. Table adapted from table generated by W.D Hill for publication. Calculations and arrangement are work of Dr W.D. Hill

	Replication		*Discovery*	
	Intelligence	Education	Intelligence	Education
$\text{Intelligence}_{\text{Replication}}$		0.249	0.378	0.308
$\text{Education}_{\text{Replication}}$	0.249		0.325	0.506
$\text{Intelligence}_{\text{Discovery}}$	0.378	0.325		0.446
$\text{Education}_{\text{Discovery}}$	0.308	0.506	0.446	

Table 2.10: Correlation of MAGMA GWGAS scores between samples. Values are pearson product moment correlation ($p < 2.2 \times 10^{-16}$ all). Correlation is between -$\log_{10}$(p) transformation of gene level p values. Cases where a gene score is present in one study are excluded

`RProjects/paper_xls_output/R/chapter_2/MAGMA/pairwise_gene_comparison.R`

correlation levels.

2.3.2 Intelligence Replication GWGAS

2.3.2.1 Methods

The $\text{Intelligence}_{\text{Replication}}$ sample is the 2017 "Intelligence wave 1" meta-analysis study by Sniekers et al.[383] (table 2.9).

GWAS summary data were downloaded from the authors Complex Trait Genomics (CTG) lab[28]. The meta-analysis of 78,308 individuals combined data from 13 cohorts with 29,610 individuals reported for the first time. UK Biobank participants took ei-

[28] `https://ctg.cncr.nl/documents/p1651/sumstats.txt.gz`

		Intelligence		*Education*	
Phenotype	Samples	Intelligence		Education	
		Discovery	Replication	Okbay	Hill
Intelligence	Hill	21.19%(0.91%)			
	Sniekers	0.99(0.02)	19.45%(1.09)		
Education	Okbay	0.62(0.03)	0.75(0.02)	9.78(0.44%)	
	Hill	0.66(0.02)	0.72(0.02)	0.96(0.02)	11.67%(0.91%)

Table 2.11: Using LD score regression heritability within, and genetic correlations between, each of the data sets used is shown. Heritability on the observed scale is shown on the diagonal and genetic correlations are shown on the off diagonal with standard error in brackets. Table contents and caption from Dr W.D. Hill B.5.

ther a web based ($n = 17,852$, raw genotype data used, reported for first time) or touch screen test ($n = 35,257$). The Childhood Intelligence Consortium (CHIC) consortium supplied 12,441 participants[39]. Five additional cohorts contributed $n = 11,749$ participants.

The individuals from this meta-analysis do not contribute to any of the other three samples used in this thesis: individuals from UK Biobank were excluded from the $\text{Intelligence}_{\text{Discovery}}$ sample summary statistics provided by Dr Hill[29]. The authors report finding 47 genome wide significant genes using MAGMA to perform a Genome Wide Gene Association (GWGAS)[383],[239].

Gene level scores were calculated using MAGMA v1.06b (Mac OS)(section 2.3) [30]. The number of individuals tested for each SNP was available and used for sample size. The European panel of the 1000 Genomes Phase 3 was used to account for linkage disequilibrium (updated 19/09/2018) and downloaded from CTG MAGMA website[31]. SNPs were assigned to genes using the NCBI37.3 Gene boundary file from the CTG website and no window was used around gene boundaries. Default options for summary statistics were used (SNPwise). All subsequent GWGAS performed using MAGMA use identical methods unless otherwise stated.

[29] see contributors/acknowledgments sec B.5

[30] `/RProjects/paper_xls_output/R/0_write_MAGMA_gene_level_results_PhD.R`

[31] `https://ctg.cncr.nl/software/magma`

2.3.2.2 Results

MAGMA v1.06 calculated gene level scores for 18,262 genes using the supplied summary statistics. The Bonferroni correction for multiple comparisons was used to calculate a genome wide significance level ($\alpha = 0.05/18262$; p=2.74×10^{-6}). Forty seven genes achieved genome wide significance using the adjusted α level.

This is identical to the number reported by Sniekers et al. [383]. Sixteen significant genes were found in the PSP (34.0% of significant genes) (table 2.12)[32].

Significant synaptic genes are shown in table 2.13. All significant genes (n=47) are listed in the supplemental material. Due to the number of significant genes in subsequent sections their details are recorded in the supplementary material.

Sample	n significant PSP	n significant genes	percent (%) in PSP
$\text{Intelligence}_{\text{Replication}}$[383]	16 (7)	47	34.0
$\text{Education}_{\text{Replication}}$[309]	25 (12)	99	25.2
$\text{Intelligence}_{\text{Discovery}}$[180]	51 (17)	196	26.0
$\text{Education}_{\text{Discovery}}$[180]	58 (27)	235	24.7

Table 2.12: Number of genome wide significant genes (using Bonferroni correction of 0.05/n genes) using MAGMA Genome Wide Gene Association Analysis Study (GWGAS) for Educational attainment and Intelligence discovery and replication samples. The references in the study column refer to the publication from which the majority of the sample are taken. Number in parenthesis are number of significant genes in 1312 gene consensus hPSD.

RProjects/paper_xls_output/R/chapter_2/exploratory_data_analysis_core_PSD.R

2.3.3 Education Replication MAGMA GWGAS

2.3.3.1 Methods

The $\text{Education}_{\text{Replication}}$ sample is a subset of individuals from the EA2 (Educational Attainment 2) meta-analytic GWA study of educational attainment[309]. Educational attainment is measured as the number of years of schooling completed (EduYears). The study comprised 293,723 individuals with 111,349 independent samples from UK Biobank used for out of sample replication [395]. A combined meta-analysis was

[32]Code for Gene level statistics at `RProjects/paper_xls_output/R/0_write_MAGMA_gene_level_results.R`

carried out using the discovery and replication cohorts (n = 405,702) allowing an increase in the number of genome wide loci reported from 74 to 162. Due to Institutional Review Board (IRB) restrictions, which did not allow reporting of over 10,000 SNPs in the summary statistics for participants from 23andme, the summary statistics provided on the SSGAC website are of n=328,917 individuals excluding 76,155 individuals from 23andme[33].

The authors kindly made summary statistics available to us (see acknowledgements B.5) excluding both 23andme and UK Biobank creating a new sample that does not overlap with $\text{Education}_{\text{Discovery}}$; n= 217,568 (293,723 - 76,155 = 217,568)[309][34]. The summary statistics used were unsorted by sex. The modified summary statistics excluding UK Biobank and 23andMe are deposited with the University of Edinburgh Research Data Service [35]. The number of individuals tested for each SNP was available and used. Twenty duplicate SNPs were identified.

2.3.3.2 Gene level analysis with MAGMA

Using SNPs from the summary data MAGMA 1.06b calculated a gene level significance scores for 17,942 genes. The Bonferroni correction for multiple comparisons was used to calculate a genome wide significance level (α = 0.05/17942; p = 2.79×10^{-06}). At this α level 99 genes were identified of genome wide significance. Twenty five (25.3%) of these were found in the PSP (see table 2.12). All significant and synaptic significant genes are recorded in the supplementary material.

2.3.4 Intelligence Discovery Phase II GWGAS

2.3.4.1 Methods

The $\text{Intelligence}_{\text{Discovery}}$ sample comprise UK Biobank Phase II participants from an analysis reported by Hill et al.[180]. The authors used MTAG (Multi-Trait Analysis of GWAS) a technique that allows meta-analyses to increase their effective sample size using genetically related traits [407]. The published meta-analysis[180] combines the $\text{Intelligence}_{\text{Replication}}$ sample data [383] and UK Biobank Intelligence Phase II ($\text{Intelligence}_{\text{Discovery}}$ sample) combined with the genetically related educational attainment phenotype (the SSGAC EA2 study)[309]. Summary GWAS statistics including

[33] `http://ssgac.org/documents/EduYears_Main.txt.gz`

[34] See Okbay et al. supplemental information `http://ssgac.org/documents/SI_74_loci_educational_attainment.pdf` p 16

[35] https://www.ed.ac.uk/information-services/research-support/research-data-service

only participants in UK Biobank Intelligence Phase II (n = 120,934) were obtained from the Centre for Cognitive Ageing and Cognitive Epidemiology (CCACE) from the author Dr W.D. Hill (section B.5). A column is provided with the summary data of numbers tested per SNP but is constant and equal to the sample size and was used as the input for the GWGAS. The participants had taken the VNR test in UK Biobank which is closely related to intelligence[180]. For further details on the cohorts, genotyping and testing see Hill et al.[180].

2.3.4.2 MAGMA Gene level analysis

18,287 genes were assigned gene level statistics using MAGMA v1.06 to provide GWGAS. With a Bonferroni α level of 2.73×10^{-06}, 196 genome wide significant genes were found at GWGAS using MAGMA. Of these 51 genes were in the PSP synaptic set (26.0% of total; table 2.12).

2.3.5 Education Discovery sample GWGAS

2.3.5.1 Methods

The GWAS summary statistics for the $\text{Education}_{\text{Discovery}}$ Phase II are taken from UK Biobank Phase II. The data were obtained from the Centre for Cognitive Ageing and Cognitive Epidemiology (CCACE) from Dr W.D Hill (see acknowledgements B.5). The sample size is n=273,274. The outcome measure is the achievement of a college or university degree which has been shown to correlated closely with years of education[358].

2.3.5.2 MAGMA Gene level analysis

18,287 genes contained SNPs. Numbers of individuals tested for each SNP were available and used in the command line argument. At a Bonferroni adjusted α level of 2.73×10^{-06}, 235 significant genes were identified. Of these, 58 were members of the PSP (24.7%).

Significant genes are recorded in the supplemental material.

2.3.6 Pascal and MAGMA

Due to the importance of gene level p values to the network analysis I have used two different popular methods to calculate gene level p values for summary GWA data.

Pascal[36] uses methods similar to the popular VEGAS package and is discussed in more detail in section 1.4.4 [228]. There is a strong linear correlation between MAGMA gene level statistics and those calculated using PASCAL see table 2.14.

2.3.6.1 Pascal software

Pascal was used using the default options and no gene boundary. The empirical method was used for calculating p values. No window was used around the gene boundaries defined using NCBI 37. 1000 Genomes Phase 3 was used to account for linkage disequilibrium. The sole release version Alpha 0.1 was used and instructions were recorded as a shell script.

[36] version 0.1 15 April 2016

GeneID	Symbol	chr	n SNPs	z-score	MAGMA p
1434	CSE1L	20	176	5.96	1.26×10^{-9}
60412	EXOC4	7	2184	5.88	1.99×10^{-9}
55964	SEPT3	22	69	5.75	4.35×10^{-9}
6138	RPL15	3	19	5.40	3.35×10^{-8}
23109	DDN	12	8	5.35	4.41×10^{-8}
85358	SHANK3	22	199	5.31	5.49×10^{-8}
487	ATP2A1	16	47	5.23	8.37×10^{-8}
11273	ATXN2L	16	36	5.17	1.17×10^{-7}
2870	GRK6	5	39	5.07	2.03×10^{-7}
28512	NKIRAS1	3	73	5.04	2.29×10^{-7}
4700	NDUFA6	22	16	5.04	2.32×10^{-7}
4733	DRG1	22	99	4.83	6.91×10^{-7}
10564	ARFGEF2	20	406	4.80	7.83×10^{-7}
7284	TUFM	16	7	4.79	8.49×10^{-7}
553115	PEF1	1	30	4.77	9.17×10^{-7}
8411	EEA1	12	475	4.61	2.03×10^{-6}

Table 2.13: MAGMA GWGAS $\text{Intelligence}_{\text{Replication}}$ PSP genes meeting genome wide significance at GWGAS using MAGMA for the $\text{Intelligence}_{\text{Replication}}$ sample. chr = chromosome number. Sixteen of genes are found in the PSP (34.0%, total significant genes = 47), Bonferroni corrected $\alpha = 2.74 \times 10^{-06}$. n SNPs = number of SNPs found within NCBI 37 defined gene boundaries in sample. GeneID refers to Entrez Gene GeneID.

Sample	t	df	CI	r	p
$\text{Education}_{\text{Discovery}}$	274.9	17,241	0.890 - 0.905	0.902	$< 2.2 \times 10^{-16}$
$\text{Intelligence}_{\text{Discovery}}$	286.01	17,164	0.907 - 0.912	0.909	$< 2.2 \times 10^{-16}$
$\text{Education}_{\text{Replication}}$	375.95	17,079	0.943 - 0.946	0.945	$< 2.2 \times 10^{-16}$
$\text{Intelligence}_{\text{Replication}}$	281.83	17,200	0.904 - 0.909	0.907	$< 2.2 \times 10^{-16}$

Table 2.14: Correlation between gene level p-values calculated in MAGMA and Pascal (-$\log_{10}$ transform of p-values). Correlation and p for Pearson's product-moment correlation.

2.4 Gene Ontology Analysis of significant GWGAS genes

Pathway analysis of the GWA studies making up the discovery and replication samples have been published[383],[309],[180], but use more than one method and can be difficult to compare. For example Okbay et. al[309] used DEPICT for pathway analysis[329], and reported many more involved pathways than the studies using MAGMA GSA[383],[180]. Although the samples making up these studies have been combined to produce powerful meta-analyses, analysis as pairs of non overlapping samples allows comparisons between samples and phenotypes has not previously been reported and may have implications for subsequent analysis.

The discovery cohorts are likely to have more power being more homogeneous (all UK Biobank), having larger samples and using a single test of intelligence in the $\text{Intelligence}_{\text{Discovery}}$ sample. Gene ontology analysis of the genome wide significant genes (both for the whole genome and those found in the PSP) may show if there systematic differences across the samples, for example features common to a particular phenotype or sample (more significant sets in $\text{Education}_{\text{Discovery}}$).

The results reported using MAGMA GSA provide an estimate of the significance levels to be found using this method and can inform study design[383][180]. The non-adjusted p values represent a likely lower bound for the significance levels in these samples given the number of gene sets examined (6,166 GO terms in Sniekers et al[383], 10,891 in Hill et al.[180]). The fact that so few significant gene sets are reported (using MAGMA GSA) in some of the most well powered population studies in genetics limitation by type II error, or a very heterogeneous set of genetic variants [383],[180].

In the $\text{Intelligence}_{\text{Replication}}$ sample, the authors reported only one significant set using MAGMA GSA after multiple testing[383]. Hill et al. found 8 significant gene sets using MAGMA GSA (table 2.21), but the sample size (n=248,482) combines both $\text{Intelligence}_{\text{Discovery}}$ and IntelligenceReplication samples, and is further increased using MTAG to incorporate data from a genetically related trait (educational attainment)[180],[407].

Finally, it will provide a practical use of various gene ontologies enrichment platforms. The methods vary to the extent that they take into account the structure of the DAG, and hence the correlation between related terms in their implementation, how they deal with multiple comparisons, their frequency of update and variation in the results obtained[215],[357]. I have therefore used more than one method in this sec-

tion to discover to what extent these results are comparable, and any advantages and disadvantages the methods may have, in addition to those discussed in section 2.1.3.

2.4.1 Intelligence Replication GO analysis

Sniekers et al. tested 6,166 Gene Ontology terms and 674 Reactome terms using MAGMA GSA [37].

One significant gene set was identified after correction for multiple comparisons using MAGMA GSA: regulation of cell development ($p = 3.5 \times 10^{-6}$; corrected p=0.03)[383]. This is a large and general Biological Process (GO:0060284 - lookup at Quick GO) with 962 Homo sapiens genes annotated (using PANTHER at Gene Ontology Consortium site for lookup).

I carried out pathway analysis of the 47 genome-wide significant genes identified by GWGAS (section 2.3.2) using PANTHER[271], toppGene[73], FUMA[428], g:Profiler[352] and SynGO[226] (section 2.1.4). Little consistent enrichment was found.

2.4.1.1 Results

I found no significant gene ontology terms using PANTHER over representation analysis for the genome wide significant genes from MAGMA GWGAS for Biological Process or Molecular Function. Two cellular component terms were enriched: cytoplasm, GO:0005737 (41 of 11,717 genes in annotation; FDR adjusted q=0.0268) and synapse, (GO:0045202, 12 of 1383 in annotation, fold change 3.77, FDR adjusted q=0.0366).

Using the PANTHER GO SLIM ontology, no enrichment was found for biological process or cellular component. ORA analysis of molecular function revealed three nested terms, the most specific being nucleic acid binding, GO:0003676 (11 of 1325 genes in annotation, Fold enrichment 3.68, p=1.44×10^{-4}, FDR adjusted $q = 0.026$)[38]).

PANTHER failed to identify correctly formatted valid Ensembl ID and Entrez IDs as previously noted and I used a combination of HGNC symbols and synonyms to maximise the number of genes identified. This was a consistent issue in using PANTHER led me to use ToppGene once it enabled custom background sets. The exact formatting

[37] https://static-content.springer.com/esm/art%3A10.1038%2Fng.3869/MediaObjects/41588_2017_BFng3869_MOESM2_ESM.xlsx

[38] RProjects/paper_xls_output/R/chapter_2/FUMA/edit_gene_table/CTG_add_uniprot.R

ID	Cellular component	p	FDR BH	n tested	n Annotation
GO:0032584	growth cone membrane	1.416×10^{-4}	0.038	2	8

Table 2.15: $\text{Intelligence}_{\text{Replication}}$ pathway enrichment in ToppGene for gene ontology terms. Default background set used. FDR BH false discovery rate corrected using Benjamini and Hochberg method

of the genes is found in supplemental material (see section 2.1.4.1)[39]. There was no enrichment using PANTHER pathways or protein class.

There was no evidence of increased enrichment of expression in CNS using FUMA (fig 2.21 and 2.22). No pathway enrichment was found for biological function, molecular function or cellular component using g:Profiler. Only one term was found enriched in ToppFun a cellular component (see table 2.15). No enrichment was found of gene ontology terms in ToppFun using the PSP as a background set.

2.4.2 Education Replication Gene Ontology analysis

Okbay et al.[309] used DEPICT[329] to investigate pathway enrichment in genes near significant SNPs. 283 gene sets (far more with DEPICT than other methods) were determined to be significant and clustered into five groups: "neural progenitor cells, migration of new neurons to cortex, projection of axons to target, sprouting of dendrites and spines, and neuronal signalling and synaptic plasticity throughout the lifespan"[309]. An association with genetic variants identified and genes involved in intellectual disability was also reported. The $\text{Education}_{\text{Replication}}$ sample differs from that reported by Okbay et al. [309] as results obtained from UK Biobank and the individuals from 23andme are excluded (see section 2.3.3).

2.4.3 PANTHER Gene Ontology analysis Education Replication

The significant genes were again formatted to maximise the number identified in PANTHER[40].

[39] `RProjects/paper_xls_output/R/chapter_2/FUMA/edit_gene_table/CTG_add_uniprot.RRProjects/paper_xls_output/R/chapter_2/translate_sig_genes/ctg_NCBI_translate2.R`

[40] HGNC id from Symbols from NCBI from MAGMA`RProjects/paper_xls_output/R/chapter_2/translate_sig_genes/ea2_NCBI_translate2.R)`

Using the default background and FDR to correct for multiple comparisons, nineteen biological process terms were found to be enriched. No molecular function or cellular component terms were enriched. No enrichment was seen in PANTHER protein class, pathways (including Reactome pathways) or for any of the SLIM ontologies.

The PANTHER enrichment displays the hierarchy of the ontology terms; although the FDR value is just less than nominal significance a very specific term related to neurogenesis, "positive regulation of collateral sprouting" (FDR q = 0.041) is prominent in comparison to the same data ordered by FDR shown in table 2.5, figure 2.6

No significant enrichment was seen using PANTHER for molecular function or cellular function. Biological function was enriched for 19 terms using FDR to correct for multiple comparisons. The most enriched term is "nervous system development" (GO:0007399), FDR q = 0.006, but the term is the largest and most general with 2,430 members in the annotation (table 2.16). "Positive regulation of collateral sprouting", GO:0048672, is just below nominal significance after correction for multiple comparisons using FDR but is a very specific term with over 52 times fold enrichment and 3 of the 12 genes annotated to this term are significant genes in GWGAS. This balance between significance and specificity is an issue to some extent for all GO pathway analyses (see fig 2.5 and table 2.16). No PANTHER SLIM ontology term showed significant enrichment.

Using FUMA GTex enrichment I found no statistically significant differential enrichment for the genes in the central nervous system (figure 2.22 and fig 2.21)

2.4.3.1 ToppGene: ORA analysis Education Replication significant genes

The 99 significant genes were enriched compared to the rest of the genome using ToppFun and the FDR method, Benjamini and Hochberg (FDR BH) to correct for multiple comparisons. Seventy seven significant terms were found for Biological Process using the FDR method Benjamini and Hochberg (10 with Bonferroni correction and 22 using FDR Q value Benjamini and Yekutieli, FDR BY). Twenty annotations for cellular component were significantly over represented (two with Bonferroni correction and FDR BY). No molecular function annotations were found to be enriched in the 99 genes found at GWGAS. Enrichment was performed using default background and settings.

The significant biological process terms are shown in table 2.17. The top two terms are neurogenesis (GO:0022008, q FDR BH= 0.002) and gliogenesis (GO:0042063, q FDR BH= 0.002) both reported by Okbay et al. using DEPICT in an expanded sam-

	Homo sapiens (REF)	upload_1 (▽ Hierarchy NEW!)					
GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
positive regulation of collateral sprouting	12	3	.06	52.13	+	4.60E-05	4.88E-02
↳regulation of extent of cell growth	115	6	.55	10.88	+	2.51E-05	4.43E-02
↳regulation of cell size	187	7	.90	7.81	+	4.05E-05	4.60E-02
↳regulation of axonogenesis	192	7	.92	7.60	+	4.77E-05	4.46E-02
↳regulation of neuron projection development	521	11	2.50	4.40	+	4.70E-05	4.67E-02
↳regulation of neurogenesis	847	14	4.06	3.45	+	5.84E-05	4.88E-02
↳generation of neurons	1588	21	7.62	2.76	+	1.93E-05	3.84E-02
↳neurogenesis	1692	23	8.11	2.83	+	4.50E-06	1.43E-02
↳nervous system development	2432	30	11.66	2.57	+	7.75E-07	6.16E-03
negative chemotaxis	46	5	.22	22.66	+	4.42E-06	1.76E-02
membrane protein proteolysis	39	4	.19	21.39	+	5.23E-05	4.62E-02
negative regulation of axonogenesis	72	5	.35	14.48	+	3.37E-05	4.46E-02
↳negative regulation of cell morphogenesis involved in differentiation	100	6	.48	12.51	+	1.18E-05	3.12E-02
regulation of synapse organization	229	8	1.10	7.28	+	1.81E-05	4.11E-02
↳regulation of synapse structure or activity	240	8	1.15	6.95	+	2.51E-05	3.99E-02
neuron projection development	697	13	3.34	3.89	+	3.29E-05	4.75E-02
↳neuron development	848	17	4.07	4.18	+	6.68E-07	1.06E-02
↳cell development	1656	21	7.94	2.64	+	3.57E-05	4.37E-02
↳neuron differentiation	1050	18	5.04	3.57	+	2.72E-06	1.44E-02

Figure 2.5: GO enrichment of genome wide significant genes at GWGAS in the $\text{Education}_{\text{Replication}}$ sample. Enrichment calculated using PANTHER with Fisher's exact test for the over representation test and using the False Discovery Rate correction for multiple comparisons. The term with the lowest FDR is GO:0007399 nervous system development (FDR=0.006). The figure shows the hierarchy of Gene Ontology terms, the most general terms are the least indented, the deepest terms in the directed acyclic graph are the most indented. Parent terms are below and more indented than child terms. Terms with identical levels of indentation are at identical layers in the hierarchy.

Text output in `/home/grant/RProjects/paper_xls_output/data/PANTHER_txt_output/checked_for_thesis/ea2/BP_all.txt`

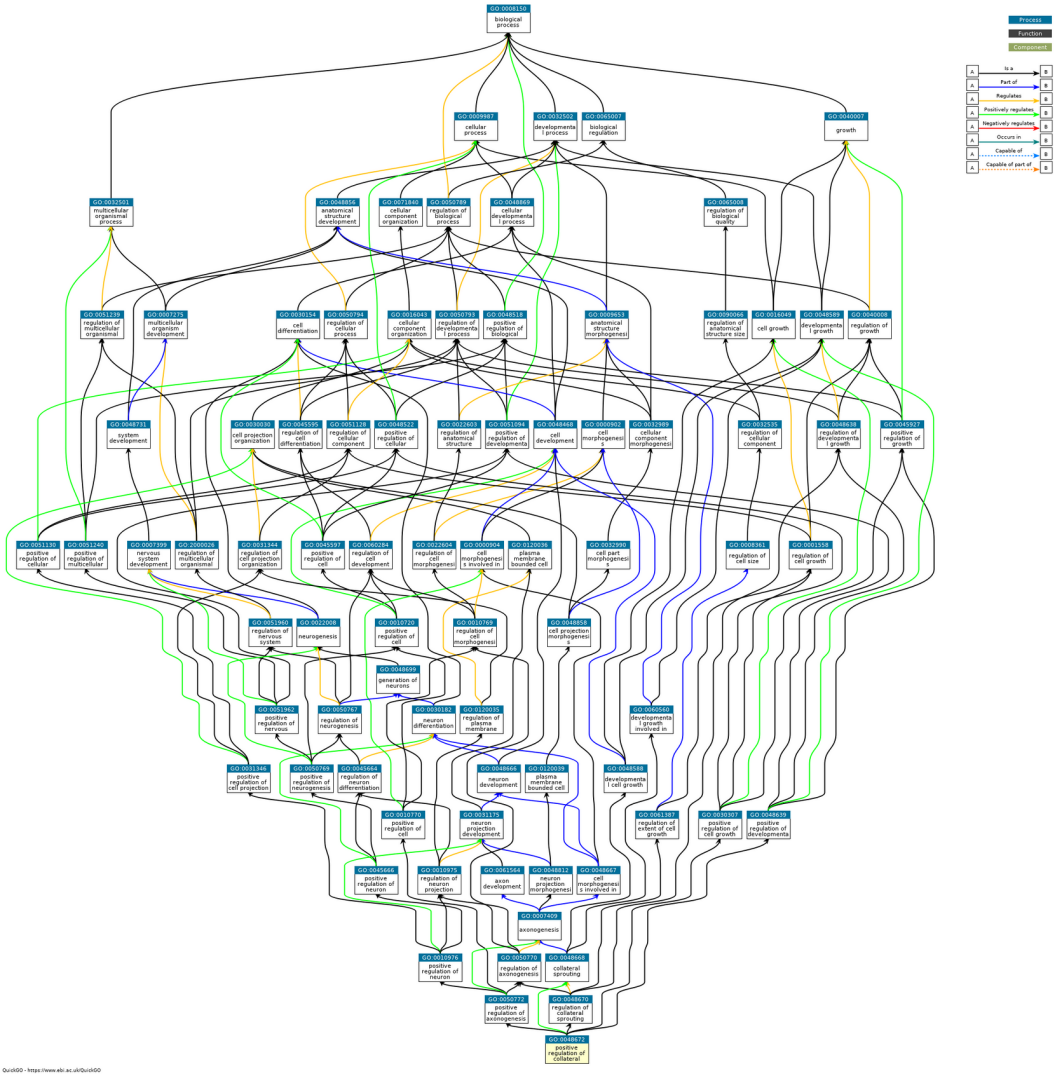


Figure 2.6: Gene Ontology Hierarchy of Biological Process term Collateral Sprouting from Quick GO. Although FDR q level is modest, in $\text{Education}_{\text{Replication}}$, the fold change for the term is marked. Image source Quick GO. The term related to other terms such as growth cone development and neuron projection development (GO:0031175) found in both Education samples and in the paper by [309].How to prioritise such specific terms over general terms is an issue faced by GO enrichment methods.

GO	description	Ref	Test	E	Fold	p	FDR q
GO:0048672	positive regulation of collateral sprouting	12	3	0.1	52.13	4.600×10^{-5}	0.049
GO:0050919	negative chemotaxis	46	5	0.2	22.66	4.420×10^{-6}	0.014
GO:0033619	membrane protein proteolysis	39	4	0.2	21.39	5.230×10^{-5}	0.046
GO:0050771	negative regulation of axonogenesis	72	5	0.3	14.48	3.370×10^{-5}	0.045
GO:0010771	negative regulation of cell morphogenesis involved in differentiation	100	6	0.5	12.51	1.180×10^{-5}	0.031
GO:0061387	regulation of extent of cell growth	115	6	0.6	10.88	2.510×10^{-5}	0.044
GO:0008361	regulation of cell size	187	7	0.9	7.81	4.050×10^{-5}	0.046
GO:0050770	regulation of axonogenesis	192	7	0.9	7.60	4.770×10^{-5}	0.045
GO:0050807	regulation of synapse organization	229	8	1.1	7.28	1.810×10^{-5}	0.041
GO:0050803	regulation of synapse structure or activity	240	8	1.1	6.95	2.510×10^{-5}	0.040
GO:0010975	regulation of neuron projection development	521	11	2.5	4.40	4.700×10^{-5}	0.047
GO:0048666	neuron development	848	17	4.1	4.18	6.680×10^{-7}	0.011
GO:0031175	neuron projection development	697	13	3.3	3.89	3.290×10^{-5}	0.048
GO:0030182	neuron differentiation	1046	18	5.0	3.59	2.580×10^{-6}	0.014
GO:0050767	regulation of neurogenesis	847	14	4.1	3.45	5.840×10^{-5}	0.049
GO:0022008	neurogenesis	1690	23	8.1	2.84	4.420×10^{-6}	0.018
GO:0048699	generation of neurons	1585	21	7.6	2.76	1.880×10^{-5}	0.037
GO:0048468	cell development	1654	21	7.9	2.65	3.510×10^{-5}	0.043
GO:0007399	nervous system development	2430	30	11.7	2.57	7.620×10^{-7}	0.006

Table 2.16: Gene ontology enrichment Biological Process using PANTHER $\text{Education}_{\text{Replication}}$ over representation analysis, with background all genome (PANTHER default). Ref: n in reference set, test:number of genes being tested present in ontology term, E expected number of genes being tested present in ontology term, Fold= Fold change, p = raw p value, FDR = false discovery rate

RProjects/paper_xls_output/R/chapter_2/make_PANTHER_tables/make_Panther_table_GO_col_nounderFDR.R

ple[309]. The significant genes are not enriched for molecular function but show enrichment for two cellular component terms, the top being "post synapse", GO:0098794.

There is also a considerable difference in actual enrichment values, including the raw p value calculated in ToppGene and PANTHER (compare tables 2.17 and table 2.16) . Widely variant scores are noted in Rhee et al.[357] and also see Khatri et al. for a detailed discussion [215].

Given that differences in significance values have been reported between enrichment packages[215](here differences can be seen in the number annotated for identical inputs), I calculated the correlation of the -log10 transform of the uncorrected p values for the seventeen annotations found in both PANTHER and ToppGene (17 of 19 for

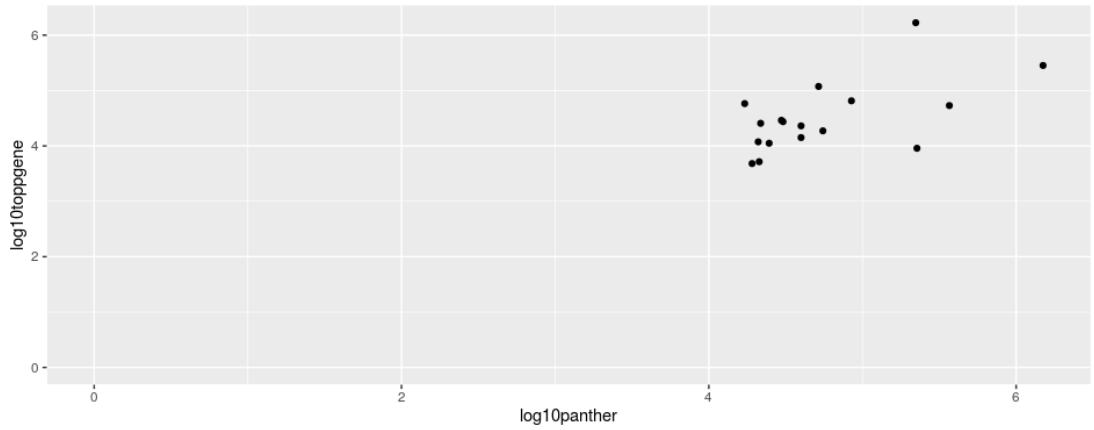
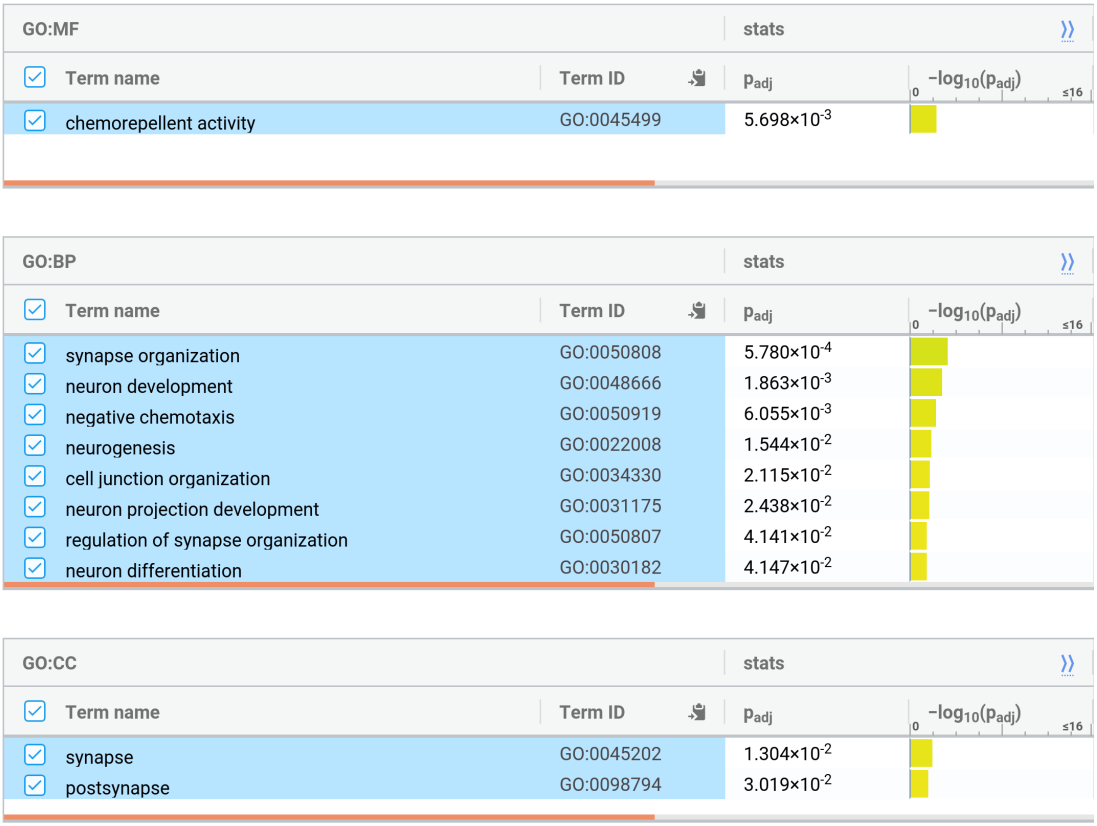


Figure 2.7: Correlation between gene ontology term enrichment values in ToppGene and PANTHER for $\text{Education}_{\text{Discovery}}$ sample. Scatterplot shows -log10 transform of uncorrected p value for PANTHER on x-axis and -log10 transform of ToppGene values on y axis. There appears to be a linear relationship between the two although the number of samples is small.

PANTHER, 17 of 77 for ToppGene). Pearson's product-moment correlation was 0.594 (95%CI 0.158-0.84); t=2.861, df=15, p-value= 0.012 (see figure 2.7).

2.4.3.1.1 SynGO analysis Sixteen genes were annotated ontology terms using SynGO and over representation was carried out using a background set, for which 1104 genes are SynGO annotated genes. The results are shown in table 2.19. The terms are more significant, with lower raw p values, due to a smaller background set (1104 SynGO annotated genes) and a decreased number of terms.

2.4.3.1.2 g:Profiler One molecular function term, two cellular component terms and eight biological process terms were found to be significant after corrections for multiple comparisons using the g:SCS method (see section 2.1.4.4). The two cellular component terms are the same as those discovered using ToppGene with the Bonferroni or Benjamini-Yekutieli methods for multiple comparisons (synapse and post synapse). Synapse organisation (GO:0050808) is the most enriched biological process term (p adjusted g:SCS 5.78×10^{-4}). Terms related to neurogenesis (GO:0022008) p adjusted = 0.0154) and neuron projection development (GO:0031175, p adjusted g:SCS=0.0024) are enriched as reported by Okbay et al[309] using DEPICT (see figure 2.8).



GO:MF

Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
chemorepellent activity	GO:0045499	5.698×10^{-3}	

GO:BP

Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
synapse organization	GO:0050808	5.780×10^{-4}	
neuron development	GO:0048666	1.863×10^{-3}	
negative chemotaxis	GO:0050919	6.055×10^{-3}	
neurogenesis	GO:0022008	1.544×10^{-2}	
cell junction organization	GO:0034330	2.115×10^{-2}	
neuron projection development	GO:0031175	2.438×10^{-2}	
regulation of synapse organization	GO:0050807	4.141×10^{-2}	
neuron differentiation	GO:0030182	4.147×10^{-2}	

GO:CC

Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
synapse	GO:0045202	1.304×10^{-2}	
postsynapse	GO:0098794	3.019×10^{-2}	

Figure 2.8: gProfiler $\text{Education}_{\text{Replication}}$. p adj is corrected p value using g:SGS method

ID	Name	p	q Bonf	q BH	q BY	n query	n genome
GO:0022008	neurogenesis	5.969×10^{-7}	0.002	0.002	0.014	26	1866
GO:0042063	gliogenesis	1.143×10^{-6}	0.003	0.002	0.014	11	352
GO:0007417	central nervous system development	1.812×10^{-6}	0.005	0.002	0.015	19	1129
GO:0048666	neuron development	3.528×10^{-6}	0.010	0.002	0.021	20	1297
GO:0010001	glial cell differentiation	5.159×10^{-6}	0.015	0.002	0.021	9	261
GO:0050808	synapse organization	5.386×10^{-6}	0.016	0.002	0.021	12	498
GO:0048667	cell morphogenesis involved in neuron dif.	5.919×10^{-6}	0.017	0.002	0.021	14	688
GO:0048699	generation of neurons	8.435×10^{-6}	0.025	0.003	0.027	23	1751
GO:0048812	neuron projection morphogenesis	1.389×10^{-5}	0.041	0.004	0.031	14	742
GO:0010771	negative regulation of cell morphogenesis dif.	1.534×10^{-5}	0.045	0.004	0.031	6	109

Table 2.17: Top 10 terms biological process Topp Gene $\text{Education}_{\text{Replication}}$. dif. = involved in differentiation. BH FDR q value Benjaminin and Hochberg, BY FDR Benjamini and Yekutieli, Bonf = Bonferroni, n query = number of genes in sample (GWGAS significant genes in $\text{Education}_{\text{Replication}}$) annotated to this term, n genome = number of genes in genome annotated to this term

RProjects/paper_xls_output/R/chapter_2/panther_and_toppgene/panther_and_toppgene_ea2.R

2.4.3.1.3 Disease Enrichment Disease enrichment in DisGeNet BeFree[335] searched by ToppFun (part of the ToppGene suite) is found for intellectual disability (ID:C3714756) which is the top term. Twenty genes are found from 1219 in the annotation ($p = 8.88 \times 10^{-7}$, FDR BH=0.00165). Intellectual disability was reported to be enriched by Okbay et al [309] for a larger related sample using DEPICT. Six of the genes are found in the PSP (See fig 2.20 for intellectual disability associated genes significant at GWGAS in $\text{Education}_{\text{Replication}}$ sample.) Testing enrichment with PSP as background found no significantly enriched disease terms other than inflammatory bowel disease .See fig 2.20 for intellectual disability associated genes significant at GWGAS in $\text{Education}_{\text{Replication}}$ sample.

2.4.4 Intelligence Discovery: Gene Ontology enrichment

Gene set analysis of this sample, combined with the $\text{Intelligence}_{\text{Discovery}}$ sample and augmented by educational attainment data using MTAG is reported by Hill et al. [180]. MAGMA GSA was carried out using 10,891 gene sets from Gene ontology, Reactome and MSigDB and using the Bonferroni correction for multiple comparisons (raw p

ID	Name	p	q Bonf	q BH	q BY	n query	n genome
GO:0098794	postsynapse	1.458×10^{-6}	0.001	0.001	0.003	16	825
GO:0045202	synapse	1.580×10^{-5}	0.005	0.003	0.018	20	1482
GO:0005794	Golgi apparatus	2.099×10^{-4}	0.073	0.015	0.097	19	1643
GO:0097418	neurofibrillary tangle	2.178×10^{-4}	0.075	0.015	0.097	2	5
GO:0035658	Mon1-Ccz1 complex	2.178×10^{-4}	0.075	0.015	0.097	2	5
GO:0031252	cell leading edge	2.705×10^{-4}	0.094	0.016	0.100	9	449
GO:0071458	integral component of cytoplasmic side of ERM	3.256×10^{-4}	0.113	0.016	0.103	2	6
GO:0044297	cell body	4.697×10^{-4}	0.163	0.020	0.131	12	820
GO:0036477	somatodendritic compartment	6.068×10^{-4}	0.210	0.023	0.146	14	1096
GO:0043025	neuronal cell body	6.590×10^{-4}	0.228	0.023	0.146	11	732

Table 2.18: Top 10 Terms Cellular component ToppGene Education$_{\text{Replication}}$. ERM=Endoplasmic reticulum membrane BH FDR q value Benjaminin and Hochberg, BY FDR Benjamini and Yekutieli, Bonf = Bonferroni, n query = number of genes in sample (GWGAS significant genes in Education extsubscriptReplication) annotated to this term, n genome = number of genes in genome annotated to this term

required = 0.05/10891 = 4.59×10^{-6})[41]. GSA revealed eight quite large gene sets[42] including neurogenesis (1,355 genes), genes associated with the synapse (717 genes), regulation of nervous system development (722 genes), neuron projection (989 genes), neuron differentiation (842 genes), CNS neuron differentiation (160 genes) and cell development (808 genes) (see table 2.21).

2.4.4.1 Intelligence Discovery Gene Ontology analysis PANTHER

A modified list of gene symbols was again used to ensure registration of genes using PANTHER. [43]

In contrast with the findings at MAGMA GSA analysis of the significant genes revealed terms associated with DNA packaging although terms associated with the synapse (GO:0045202) and vesicle membrane (GO:0012506) are enriched.

Only one biological process term was enriched: GO:0010468, regulation of gene

[41] Bonferroni α from supplementary table 5 2.75×10^{-6} which I calculate as 18,182 comparisons the same as the correction for multiple comparisons in the gene level statistics. The paper however says a correction of 10891 which I calculate means the corrections is 4.59×10^{-6}. This does not however result in any extra significant sets. 27 gene sets are obtained using the FDR correction

[42] See `RProjects/chapter2_checks/R/expected_gsa/ctg_size_plots.R` for correlation between raw p value and gene size.

[43] `RProjects/paper_xls_output/R/chapter_2/FUMA/edit_gene_table/UKBB_Int_FUMA2clipboard.R` gene list at supplemental `/home/grant/Dropbox/paper_xls_data/Panther/input.txt`

Biological Process	n	p	FDR q
synapse organization*	12	6.94×10^{-8}	3.47×10^{-7}
process in the synapse	13	4.14×10^{-4}	1.03×10^{-3}
regulation of synapse organization	3	6.98×10^{-4}	1.16×10^{-3}
synapse adhesion between pre- and post-synapse	3	1.35×10^{-3}	1.68×10^{-3}
trans-synaptic signaling	3	0.0644	

Cellular component	n	p	FDR q
synapse	14	8.68×10^{-4}	3.47×10^{-3}
postsynapse	9	5.50×10^{-3}	0.010

Table 2.19: Enrichment of Gene Ontology terms Education Replication using SynGO. SynGO only tests for biological process and cellular comparment. FDR q = adjusted significance level using FDR. Note the increased significance level despite correcting for multiple comparisons. For term Synapse Organisation (GO:0050808)* 12 genes (including child terms) in annotation, 286 in SynGO.

expression (p=4.78×10^{-6}, FDR BH=0.038; n in annotation 4907, n in set = 74). Seven molecular function terms were enriched the most specific and smallest being DNA binding GO:0003677 (n in annotation 2506, test 48, $p = 1.01 \times 10^{-6}$ FDR=0.0016 (table 2.22))

Twenty three cellular component terms were enriched. "DNA packaging complex" is the most enriched term using FDR BH 85 terms in annotation, 9 appear in the significant genes found at GWGAS, (GO:0044815), FDR=0.0002. The second most enriched term is the child term nucleosome, GO:000786, FDR 0.0007 (see figure 2.9 and table 2.23). Twenty seven genes (from 1383) are enriched for the cellular component term synapse (GO:0045202; FDR 0.0199).

Two SLIM molecular function terms were enriched: transcription factor binding (FDR=0.0145) and DNA binding (FDR=0.0189) table 2.24 continuing the finding of terms related to transcription.

Eighteen SLIM Biological process terms are over represented figure 2.10 for a screenshot that shows the hierarchy of terms. The terms are again related to transcription, including those related to nucleosome assembly, DNA packaging and those regulating RNA polymerase. No SLIM cellular component was enriched.

	Homo sapiens (REF)	upload_1 (▽ Hierarchy NEW!)					
GO cellular component complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
nucleosome	77	8	.71	11.28	+	1.10E-06	7.34E-04
↳protein-DNA complex	175	9	1.61	5.59	+	5.03E-05	5.62E-03
↳DNA packaging complex	85	9	.78	11.50	+	1.98E-07	1.99E-04
↳chromatin	1504	36	13.85	2.60	+	1.50E-07	3.01E-04
↳chromosome	2012	40	18.53	2.16	+	3.59E-06	1.80E-03
↳intracellular non-membrane-bounded organelle	5274	77	48.56	1.59	+	8.30E-06	2.08E-03
↳intracellular organelle	12856	148	118.38	1.25	+	6.65E-06	1.91E-03
↳intracellular	14704	163	135.40	1.20	+	4.77E-06	1.92E-03
↳organelle	13868	155	127.70	1.21	+	2.01E-05	2.89E-03
↳cellular anatomical entity	18767	188	172.81	1.09	+	2.91E-05	3.90E-03
↳non-membrane-bounded organelle	5283	77	48.65	1.58	+	8.50E-06	1.90E-03
cytoplasmic vesicle membrane	788	18	7.26	2.48	+	4.39E-04	3.67E-02
↳vesicle membrane	810	18	7.46	2.41	+	6.01E-04	4.83E-02
↳membrane-bounded organelle	12748	146	117.39	1.24	+	1.91E-05	2.95E-03
nuclear chromatin	1316	29	12.12	2.39	+	1.66E-05	3.04E-03
↳nuclear chromosome	1542	30	14.20	2.11	+	1.36E-04	1.24E-02
↳nuclear lumen	4993	73	45.98	1.59	+	1.84E-05	3.08E-03
↳intracellular organelle lumen	6114	82	56.30	1.46	+	8.88E-05	8.91E-03
↳organelle lumen	6114	82	56.30	1.46	+	8.88E-05	9.38E-03
↳membrane-enclosed lumen	6114	82	56.30	1.46	+	8.88E-05	8.49E-03
↳nucleus	7578	101	69.78	1.45	+	5.29E-06	1.77E-03
↳intracellular membrane-bounded organelle	11041	132	101.67	1.30	+	1.14E-05	2.28E-03
synapse	1383	27	12.73	2.12	+	2.28E-04	1.99E-02

Figure 2.9: Intelligence Discovery significant genes at GWGAS. PANTHER GO enrichment Cellular component. FDR correction. Screenshot for hierarchy. Matches table table 2.23

	Homo sapiens (REF)	upload_1 (▽ Hierarchy NEW!)					
PANTHER GO-Slim Biological Process	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
nucleosome assembly	30	5	.28	18.10	+	1.56E-05	4.60E-03
↳protein-DNA complex assembly	91	7	.84	8.35	+	3.24E-05	8.36E-03
↳protein-DNA complex subunit organization	116	9	1.07	8.43	+	2.23E-06	9.22E-04
↳nucleosome organization	54	7	.50	14.08	+	1.36E-06	9.34E-04
↳chromatin organization	253	11	2.33	4.72	+	3.26E-05	7.47E-03
↳chromosome organization	394	15	3.63	4.13	+	5.62E-06	1.93E-03
↳DNA packaging	56	7	.52	13.57	+	1.70E-06	8.75E-04
↳DNA conformation change	79	8	.73	11.00	+	1.31E-06	1.35E-03
↳chromatin assembly or disassembly	48	7	.44	15.84	+	6.61E-07	1.36E-03
positive regulation of transcription by RNA polymerase II	345	12	3.18	3.78	+	1.14E-04	1.95E-02
↳positive regulation of transcription, DNA-templated	421	13	3.88	3.35	+	1.85E-04	2.25E-02
↳positive regulation of nucleic acid-templated transcription	421	13	3.88	3.35	+	1.85E-04	2.39E-02
↳positive regulation of RNA biosynthetic process	421	13	3.88	3.35	+	1.85E-04	2.13E-02
↳positive regulation of RNA metabolic process	437	14	4.02	3.48	+	7.20E-05	1.49E-02
↳positive regulation of nucleobase-containing compound metabolic process	454	14	4.18	3.35	+	1.06E-04	2.00E-02
↳positive regulation of macromolecule biosynthetic process	458	14	4.22	3.32	+	1.16E-04	1.85E-02
↳positive regulation of biosynthetic process	468	14	4.31	3.25	+	1.45E-04	1.99E-02
↳positive regulation of cellular biosynthetic process	465	14	4.28	3.27	+	1.36E-04	2.00E-02

Figure 2.10: Intelligence Discovery large screenshot Biological Process SLIM.

`/home/grant/Dropbox/paper_xls_data/Panther/output_txt/ukbbint/bp_all_slim_checked_with_image_matches_current_image_overleaf.txt`

Entrez Gene ID	Gene Symbol	Gene Name
5144	PDE4D	phosphodiesterase 4D
5662	PSD	pleckstrin and Sec7 domain containing
4137	MAPT	microtubule associated protein tau
4208	MEF2C	myocyte enhancer factor 2C
4744	NEFH	neurofilament heavy
257194	NEGR1	neuronal growth regulator 1

Table 2.20: Genes in PSP and associated with intellectual disability found amongst genome-wide significant Education Replication genes.

Gene set name	N genes in gene set	Beta	SE(β)	p
Neurogenesis	1355	0.20	0.05	5.59×10^{-10}
Regulation of nervous system development	722	0.23	0.05	4.02×10^{-8}
Regulation of cell development	808	0.22	0.04	7.38×10^{-8}
Neuron projection	898	0.20	0.04	2.07×10^{-7}
Central nervous system neuron differentiation	160	0.47	0.04	5.33×10^{-7}
Synapse	717	0.21	0.04	1.43×10^{-6}
Neuron differentiation	842	0.19	0.04	1.62×10^{-6}
Oligodendrocyte differentiation	1037	0.17	0.04	1.75×10^{-6}

Table 2.21: MAGMA GSA significant sets from Hill et al. 2019 Bonferroni α=2.75×10^{-6} from supplementary table 5 [180].

PANTHER protein classes were enriched for histone and DNA binding protein (see table 2.25.

Chromatin, RNA polymerase and Pre-Notch expression and processing were among the most enriched of the 98 Reactome pathways tested using PANTHER with 98 terms in total with FDR < 0.05 (see figure 2.11). [44]

g:Profiler results are shown in figure 2.12 and are a good summary of the main sources of enrichment found in ToppGene and PANTHER (Transcription related and nucleosome assembly).

[44] table at `RProjects/paper_xls_output/R/chapter_2/make_PANTHER_tables/UKBB_int/make_Panther_table_GO_col_nounderFDRreactome.R` includes reactome iids

GO	description	Ref	Test	E	Fold	p	FDR
GO:0003677	DNA binding	2506	48	23.1	2.08	1.01×10^{-6}	0.0016
GO:0003676	nucleic acid binding	4034	64	37.1	1.72	5.51×10^{-6}	0.0052
GO:1901363	heterocyclic compound binding	5999	83	55.2	1.50	2.73×10^{-5}	0.0185
GO:0097159	organic cyclic compound binding	6086	83	56.0	1.48	4.27×10^{-5}	0.0254
GO:0005515	protein binding	14109	159	129.9	1.22	3.77×10^{-6}	0.0045
GO:0005488	binding	16469	174	151.7	1.15	2.39×10^{-5}	0.0190
GO:0003674	molecular_function	18321	188	168.7	1.11	9.41×10^{-7}	0.0022

Table 2.22: GO molecular function complete Intelligence Discovery FDR over representation only shown. Ref=reference set, test:number of genes being tested present in ontology term, E=expected number of genes, Fold= Fold change, P = raw p value, FDR = false discovery rate Benjamini and Hochberg. The most enriched term is DNA binding and all of the annotations are large (smallest is 2506 genes).

RProjects/paper_xls_output/R/chapter_2/make_PANTHER_tables/UKBB_int/make_Panther_table_GO_col_nounderFDR_int.R

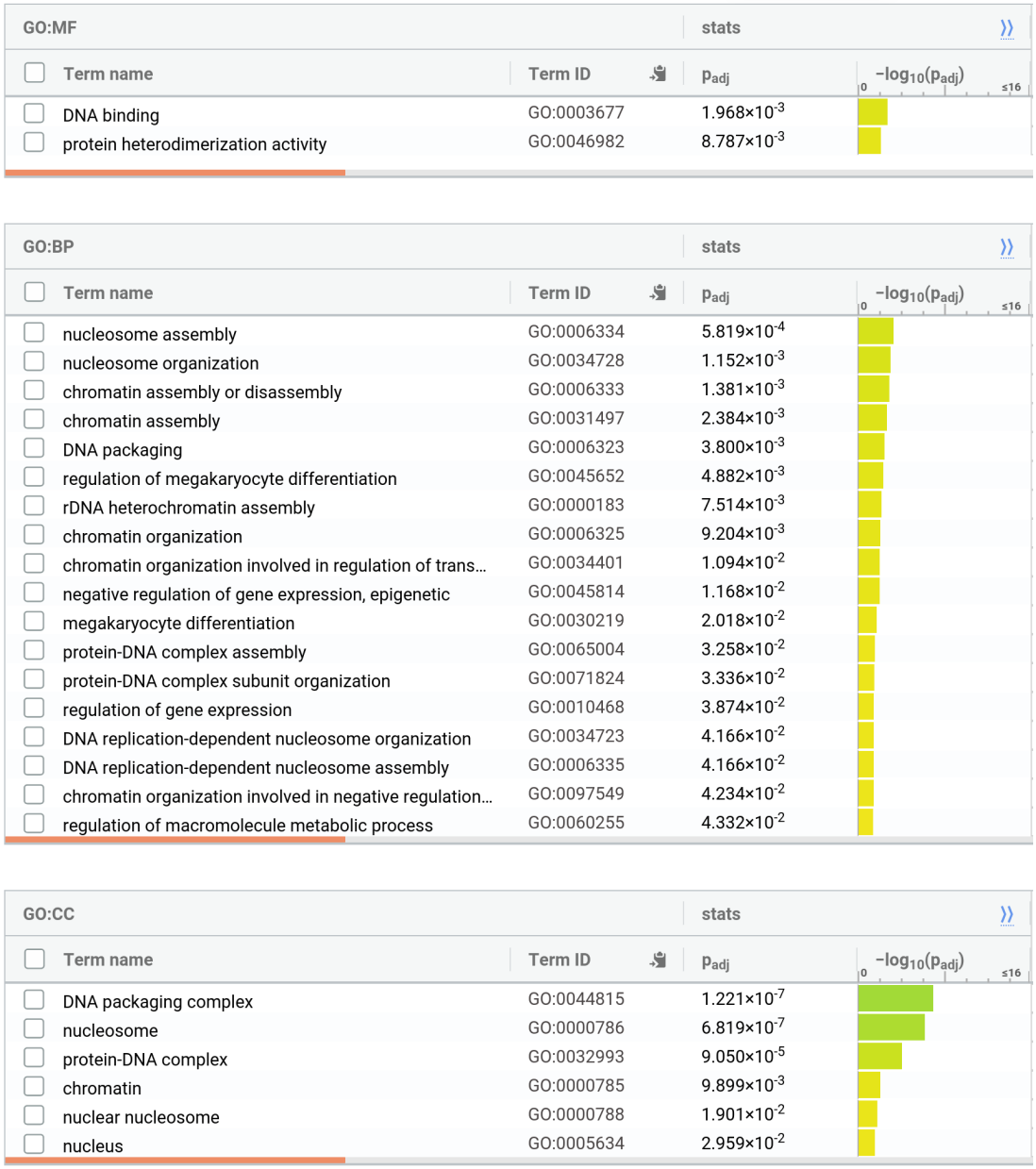
	Homo sapiens (REF)	upload_1 (Hierarchy) NEW!					
Reactome pathways	#	#	expected	Fold Enrichment	+/-	raw P value	▲ FDR
Pre-NOTCH Expression and Processing	78	10	.72	13.92	+	7.70E-09	1.76E-05
RHO GTPases activate PKNs	63	8	.58	13.79	+	2.70E-07	1.23E-04
Activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3	36	7	.33	21.12	+	1.15E-07	1.32E-04
Pre-NOTCH Transcription and Translation	62	8	.57	14.01	+	2.42E-07	1.38E-04
Chromatin organization	240	13	2.21	5.88	+	6.12E-07	1.40E-04
RUNX1 regulates genes involved in megakaryocyte differentiation and platelet function	66	8	.61	13.16	+	3.74E-07	1.42E-04
ERCC6 (CSB) and EHMT2 (G9a) positively regulate rRNA expression	45	7	.41	16.89	+	4.46E-07	1.46E-04
Formation of the beta-catenin:TCF transactivating complex	60	8	.55	14.48	+	1.92E-07	1.46E-04
Transcriptional regulation by RUNX1	204	12	1.88	6.39	+	7.36E-07	1.53E-04
Chromatin modifying enzymes	240	13	2.21	5.88	+	6.12E-07	1.55E-04
Signaling by NOTCH	200	12	1.84	6.52	+	6.04E-07	1.72E-04
RNA Polymerase I Promoter Opening	32	6	.29	20.36	+	1.16E-06	2.20E-04

Figure 2.11: Education Discovery most significantly enriched Reactome pathways PANTHER. NOTCH is a cell signalling pathway involved in neurogenesis

2.4.4.2 SynGO Intelligence Discovery

No significant terms were enriched for cellular component. The terms enriched for biological process are in figure 2.26. Although the FDR values are higher than found in the $\text{Education}_{\text{Discovery}}$, some synaptic terms are found to be enriched by using a limited ontology.

2.4.4.2.1 ToppGene Five biological processes terms, twenty nine molecular function terms and five cellular components had FDR < 0.05. See figure 2.13. The enrichment is that is present is also predominantly terms such as nucleosome function, DNA packaging and chromatin binding. Synaptic terms and neurogenesis terms are absent compared with the education samples.



GO:MF

Term name	Term ID	p_{adj}	stats: $-\log_{10}(p_{adj})$ (0 – ≤16)
DNA binding	GO:0003677	1.968×10^{-3}	
protein heterodimerization activity	GO:0046982	8.787×10^{-3}	

GO:BP

Term name	Term ID	p_{adj}	stats: $-\log_{10}(p_{adj})$ (0 – ≤16)
nucleosome assembly	GO:0006334	5.819×10^{-4}	
nucleosome organization	GO:0034728	1.152×10^{-3}	
chromatin assembly or disassembly	GO:0006333	1.381×10^{-3}	
chromatin assembly	GO:0031497	2.384×10^{-3}	
DNA packaging	GO:0006323	3.800×10^{-3}	
regulation of megakaryocyte differentiation	GO:0045652	4.882×10^{-3}	
rDNA heterochromatin assembly	GO:0000183	7.514×10^{-3}	
chromatin organization	GO:0006325	9.204×10^{-3}	
chromatin organization involved in regulation of trans...	GO:0034401	1.094×10^{-2}	
negative regulation of gene expression, epigenetic	GO:0045814	1.168×10^{-2}	
megakaryocyte differentiation	GO:0030219	2.018×10^{-2}	
protein-DNA complex assembly	GO:0065004	3.258×10^{-2}	
protein-DNA complex subunit organization	GO:0071824	3.336×10^{-2}	
regulation of gene expression	GO:0010468	3.874×10^{-2}	
DNA replication-dependent nucleosome organization	GO:0034723	4.166×10^{-2}	
DNA replication-dependent nucleosome assembly	GO:0006335	4.166×10^{-2}	
chromatin organization involved in negative regulation...	GO:0097549	4.234×10^{-2}	
regulation of macromolecule metabolic process	GO:0060255	4.332×10^{-2}	

GO:CC

Term name	Term ID	p_{adj}	stats: $-\log_{10}(p_{adj})$ (0 – ≤16)
DNA packaging complex	GO:0044815	1.221×10^{-7}	
nucleosome	GO:0000786	6.819×10^{-7}	
protein-DNA complex	GO:0032993	9.050×10^{-5}	
chromatin	GO:0000785	9.899×10^{-3}	
nuclear nucleosome	GO:0000788	1.901×10^{-2}	
nucleus	GO:0005634	2.959×10^{-2}	

Figure 2.12: g:profiler GO enrichment results for significant genes found at GWGAS for the $\text{Intelligence}_{\text{Discovery}}$ sample

1: GO: Molecular Function [Display Chart] 195 input genes in category / 638 annotations before applied cutoff / 19248 genes in category

	ID	Name	Source	pValue	FDR B&H	FDR B&Y	Bonferroni	Genes from Input	Genes in Annotation
1	GO:0008301	DNA binding, bending		5.412E-6	3.453E-3	2.429E-2	3.453E-3	12	236
2	GO:0031492	nucleosomal DNA binding		2.338E-4	3.883E-2	2.732E-1	1.492E-1	5	55
3	GO:0003682	chromatin binding		2.366E-4	3.883E-2	2.732E-1	1.509E-1	17	629
4	GO:0046982	protein heterodimerization activity		2.740E-4	3.883E-2	2.732E-1	1.748E-1	17	637
5	GO:0047894	flavonol 3-sulfotransferase activity		3.043E-4	3.883E-2	2.732E-1	1.941E-1	2	3

2: GO: Biological Process [Display Chart] 192 input genes in category / 3742 annotations before applied cutoff / 20148 genes in category

	ID	Name	Source	pValue	FDR B&H	FDR B&Y	Bonferroni	Genes from Input	Genes in Annotation
1	GO:0006334	nucleosome assembly		1.694E-7	3.490E-4	3.073E-3	6.339E-4	11	147
2	GO:0034728	nucleosome organization		2.476E-7	3.490E-4	3.073E-3	9.266E-4	12	187
3	GO:0006333	chromatin assembly or disassembly		4.340E-7	3.490E-4	3.073E-3	1.624E-3	12	197
4	GO:0006325	chromatin organization		4.655E-7	3.490E-4	3.073E-3	1.742E-3	25	841
5	GO:0071824	protein-DNA complex subunit organization		4.966E-7	3.490E-4	3.073E-3	1.858E-3	16	367

Show 24 more annotations

3: GO: Cellular Component [Display Chart] 195 input genes in category / 447 annotations before applied cutoff / 20584 genes in category

	ID	Name	Source	pValue	FDR B&H	FDR B&Y	Bonferroni	Genes from Input	Genes in Annotation
1	GO:0044815	DNA packaging complex		8.664E-11	3.873E-8	2.588E-7	3.873E-8	13	117
2	GO:0000786	nucleosome		5.227E-10	1.168E-7	7.805E-7	2.337E-7	12	109
3	GO:0032993	protein-DNA complex		2.102E-7	3.132E-5	2.093E-4	9.397E-5	13	222
4	GO:0000788	nuclear nucleosome		3.643E-5	4.071E-3	2.720E-2	1.629E-2	5	40
5	GO:0000790	nuclear chromatin		1.267E-4	1.133E-2	7.568E-2	5.664E-2	35	1924

Figure 2.13: Screen shot Gene ontology ToppGene Biological Process, Molecular Function and Cellular component Intelligence Discovery. Terms related to DNA binding, nucleosomes and chromatin are prominent

There is no enrichment for genes differentially enriched in the central nervous system areas using GTEx version 8 analysed using FUMA (figure 2.22 and fig 2.21).

2.4.5 Education Discovery Ontology Over representation analysis

The 235 genome-wide significant genes identified using MAGMA GWGAS for the $\text{Education}_{\text{Discovery}}$ sample (section 2.3.5)and were tested for over-representation of ontology terms. Using PANTHER, I was unable to have all of the significant genes output by MAGMA identified, and used the gene identifier with the most uniquely identified genes (233 terms, HGNC symbols with synonyms[45]). Thirty-three biological process ontology terms, were over-represented (FDR corrected q <0.05) amongst the 233 significant genes (see figure 2.14). There are several enriched terms related to neurogenesis and axonal development, similar to the $\text{Education}_{\text{Replication}}$ sample but distinct from the Intelligence samples. The most significant term (FDR=8.51×10^{-4}) was "nervous system development" (GO:0007399, 57 genes found amongst 2,432 in the annotation). Many of the sets are broad and non-specific; the lowest FDR for annotations with less than 500 genes was "negative regulation of cell projection" (GO:0031345;11 of 183 genes;5.4 fold change; FDR=0.0089). Considering only those terms with greater than

[45] `RProjects/paper_xls_output/R/chapter_2/FUMA/edit_gene_table/UKBB_Ed_FUMA2clipboard.R`

	Homo sapiens (REF)	upload_1 (▽ Hierarchy NEW!)					
GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
negative regulation of neuron projection development	153	9	1.75	5.15	+	9.66E-05	4.80E-02
↳regulation of neuron projection development	521	19	5.95	3.19	+	1.31E-05	1.10E-02
↳regulation of neurogenesis	847	26	9.67	2.69	+	6.72E-06	7.12E-03
↳regulation of nervous system development	962	29	10.98	2.64	+	2.68E-06	3.87E-03
↳regulation of multicellular organismal development	2106	50	24.04	2.08	+	7.50E-07	2.98E-03
↳regulation of multicellular organismal process	3243	67	37.02	1.81	+	9.27E-07	2.95E-03
↳regulation of cell development	977	28	11.15	2.51	+	1.01E-05	8.92E-03
↳regulation of cell differentiation	1883	41	21.49	1.91	+	6.28E-05	3.70E-02
↳generation of neurons	1588	42	18.13	2.32	+	4.68E-07	3.72E-03
↳neurogenesis	1692	43	19.31	2.23	+	1.06E-06	2.81E-03
↳cell differentiation	3781	70	43.16	1.62	+	2.99E-05	2.07E-02
↳cellular developmental process	3835	71	43.77	1.62	+	2.28E-05	1.72E-02
↳cellular process	15625	207	178.35	1.16	+	7.41E-06	6.93E-03
↳nervous system development	2432	57	27.76	2.05	+	1.43E-07	2.28E-03
↳anatomical structure development	5497	91	62.74	1.45	+	8.15E-05	4.62E-02
↳regulation of plasma membrane bounded cell projection organization	698	24	7.97	3.01	+	2.46E-06	3.91E-03
↳regulation of cell projection organization	707	24	8.07	2.97	+	3.04E-06	4.03E-03
↳regulation of cellular component organization	2439	54	27.84	1.94	+	2.27E-06	4.51E-03
↳negative regulation of cell projection organization	183	12	2.09	5.74	+	2.35E-06	4.14E-03
↳negative regulation of cellular component organization	724	24	8.26	2.90	+	4.50E-06	5.50E-03
regulation of cell size	187	10	2.13	4.68	+	8.47E-05	4.49E-02
↳cellular component organization	5695	100	65.00	1.54	+	1.20E-06	2.71E-03
↳cellular component organization or biogenesis	5915	101	67.52	1.50	+	4.54E-06	5.15E-03
regulation of synapse organization	229	11	2.61	4.21	+	9.43E-05	4.83E-02
negative regulation of cellular component movement	316	13	3.61	3.60	+	1.00E-04	4.82E-02
neuron projection development	697	22	7.96	2.77	+	2.37E-05	1.71E-02
↳neuron development	848	26	9.68	2.69	+	6.86E-06	6.81E-03
↳cell development	1656	43	18.90	2.27	+	4.74E-07	2.51E-03
↳plasma membrane bounded cell projection organization	1149	29	13.12	2.21	+	8.31E-05	4.56E-02
↳cell projection organization	1196	31	13.65	2.27	+	3.18E-05	2.10E-02
cellular protein modification process	3195	61	36.47	1.67	+	4.20E-05	2.67E-02
↳protein modification process	3195	61	36.47	1.67	+	4.20E-05	2.57E-02
↳macromolecule modification	3420	65	39.04	1.67	+	2.22E-05	1.76E-02

Figure 2.14: PANTHER Gene Ontology Enrichment Biological Process $\text{Education}_{\text{Discovery}}$ FDR. Terms related to neurogenesis are prominent.

mean fold change (2.61), the most enriched is "regulation of nervous system development" (GO:0051960,FDR 0.003). There was no significant over-representation of molecular function terms.

Forty cellular component terms were over-represented. Five of the ten most enriched terms had over 10,000 genes in their annotation. This includes the most significant term "Intracellular" (GO:0005622, 14,707 genes, FDR=4.9×10^{-5}) although only six of the forty significant terms contained over 10,000 genes (see figure 2.15 and table). Four terms have under 100 annotations amongst the ten most enriched, three are related to the endoplasmic reticulum including "lumenal side of membrane" (GO:0098576; FDR = 3.84×10^{-4}), one is annotated to GO:0042613 "MHC class II protein complex" ; (5 of 19 genes in annotation found; fold change 23.1, FDR=0.00126)."Synapse" (GO:0045202 ; 36 of 1,383 genes in annotation, ; fold change 15.8, FDR = 0.00116)

was also amongst the ten most enriched terms. No SLIM ontologies showed significant over-representation and there was no enrichment of PANTHER pathways. Only the protein class PC00149 “major histocompatibility complex protein” was over-represented 5 of 18 terms; 24.3 fold change; FDR=0.0010).

Seven reactome pathways were enriched (table 2.28).

2.4.5.1 ToppGENE

Seventy six biological process terms were over-represented using toppGene (FDR Benjamini and Hochberg - FDR-BH < 0.05). Using either the Bonferroni correction or FDR correction of Benjamini and Yekeutil (FDR-BY), seventeen terms are discovered. Terms related to neuron development, neurogenesis, synapse organisation, axon development and cell projection are prominent (Table 2.29). No terms related to molecular function were enriched. Forty five cellular component terms were enriched using FDR-BH table 2.30 with 10 significant using the Bonferroni correction or FDR-BY. The synapse and post synapse are over-represented as is MHC class II and terms related to the endoplasmic reticulum.

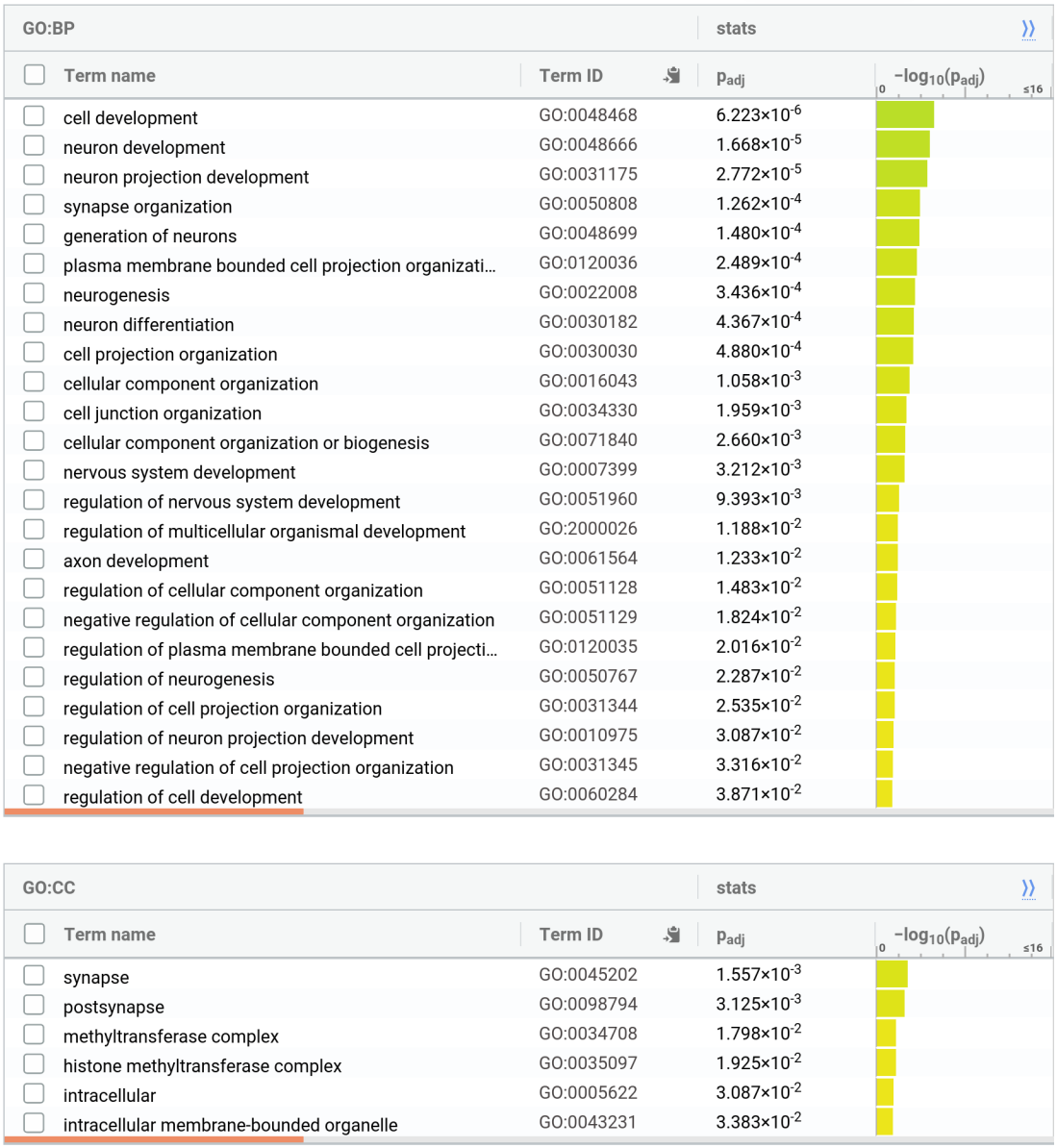
2.4.5.1.1 Phenotype 40 genes were associated with intellectual disability on searching DisGeNET using ToppGene. This was found in the $\text{Education}_{\text{Replication}}$ sample but in none of the Intelligence samples. Eleven of these genes are found in PSP and are shown in table 2.31. Of the 29 not in the PSP no molecular function or cellular component enrichment was found but the biological process GO:0007417 central nervous system development was enriched (0.02 FDR BH).

2.4.5.2 g:Profiler

Twenty four biological process terms were significantly over represented in GWGAS significant genes in the $\text{Education}_{\text{Discovery}}$ sample using g:profiler after correction for multiple comparisons. Six cellular component terms are enriched. The terms include neurogenesis and nervous system development and cellular components include synapse and post synapse and are clearly similar to those in PANTHER and ToppGene although exact adjusted significance levels differ (See figure 2.16). Figure 2.17 shows a summary of enrichment in g:profiler across the three samples in which enrichment for gene ontology terms was detected.

	Homo sapiens (REF)	upload_1 (▽ Hierarchy NEW!)					
GO cellular component complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
MHC class II protein complex	19	5	.22	23.06	+	6.26E-06	1.26E-03
↳MHC protein complex	28	5	.32	15.64	+	3.22E-05	4.31E-03
↳cellular anatomical entity	18767	232	214.21	1.08	+	1.56E-05	2.60E-03
↳protein-containing complex	5532	89	63.14	1.41	+	2.83E-04	2.03E-02
integral component of lumenal side of endoplasmic reticulum membrane	29	6	.33	18.13	+	2.39E-06	8.01E-04
↳integral component of endoplasmic reticulum membrane	164	8	1.87	4.27	+	7.71E-04	4.42E-02
↳intrinsic component of endoplasmic reticulum membrane	172	8	1.96	4.07	+	1.04E-03	4.95E-02
↳membrane-bounded organelle	12748	180	145.51	1.24	+	3.31E-06	8.30E-04
↳organelle	13868	193	158.29	1.22	+	1.04E-06	4.18E-04
↳lumenal side of endoplasmic reticulum membrane	29	6	.33	18.13	+	2.39E-06	6.87E-04
↳lumenal side of membrane	37	7	.42	16.57	+	5.73E-07	3.84E-04
histone methyltransferase complex	81	7	.92	7.57	+	6.23E-05	6.95E-03
↳nucleoplasm	3990	67	45.54	1.47	+	8.77E-04	4.76E-02
↳intracellular organelle lumen	6114	96	69.79	1.38	+	3.29E-04	2.20E-02
↳organelle lumen	6114	96	69.79	1.38	+	3.29E-04	2.28E-02
↳membrane-enclosed lumen	6114	96	69.79	1.38	+	3.29E-04	2.13E-02
↳intracellular organelle	12856	183	146.74	1.25	+	8.00E-07	4.01E-04
↳intracellular	14704	205	167.84	1.22	+	2.44E-08	4.90E-05
↳nucleus	7578	116	86.50	1.34	+	1.07E-04	9.73E-03
↳intracellular membrane-bounded organelle	11041	166	126.03	1.32	+	1.94E-07	1.95E-04
↳methyltransferase complex	110	8	1.26	6.37	+	5.89E-05	6.96E-03
ER to Golgi transport vesicle membrane	61	5	.70	7.18	+	8.96E-04	4.73E-02
↳cytoplasm	11717	165	133.74	1.23	+	4.24E-05	5.32E-03
↳Golgi-associated vesicle membrane	112	7	1.28	5.48	+	4.10E-04	2.49E-02
↳Golgi-associated vesicle	184	10	2.10	4.76	+	7.46E-05	7.88E-03
↳COPII-coated ER to Golgi transport vesicle	96	7	1.10	6.39	+	1.69E-04	1.41E-02
↳transport vesicle membrane	212	9	2.42	3.72	+	9.49E-04	4.76E-02
GABA-ergic synapse	79	6	.90	6.65	+	4.03E-04	2.53E-02
↳synapse	1383	36	15.79	2.28	+	5.18E-06	1.16E-03
↳cell junction	2122	44	24.22	1.82	+	1.39E-04	1.21E-02
trans-Golgi network membrane	95	6	1.08	5.53	+	1.01E-03	4.93E-02
↳Golgi membrane	769	20	8.78	2.28	+	7.70E-04	4.55E-02
lysosomal membrane	383	14	4.37	3.20	+	1.80E-04	1.39E-02
↳lytic vacuole membrane	383	14	4.37	3.20	+	1.80E-04	1.44E-02
↳vacuolar membrane	436	15	4.98	3.01	+	2.00E-04	1.49E-02
postsynaptic density	345	12	3.94	3.05	+	7.87E-04	4.39E-02
↳postsynapse	651	22	7.43	2.96	+	8.58E-06	1.57E-03
↳asymmetric synapse	351	12	4.01	3.00	+	9.09E-04	4.68E-02
neuron projection	1395	33	15.92	2.07	+	7.85E-05	7.88E-03
cytosol	5314	88	60.66	1.45	+	9.63E-05	9.21E-03

Figure 2.15: PANTHER Gene Ontology Enrichment Cellular Component $\text{Education}_{\text{Discovery}}$. The lowest FDR again is found in the most general and largest gene ontology terms with correspondingly small fold enrichment.



GO:BP

Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
cell development	GO:0048468	6.223×10^{-6}	
neuron development	GO:0048666	1.668×10^{-5}	
neuron projection development	GO:0031175	2.772×10^{-5}	
synapse organization	GO:0050808	1.262×10^{-4}	
generation of neurons	GO:0048699	1.480×10^{-4}	
plasma membrane bounded cell projection organizati...	GO:0120036	2.489×10^{-4}	
neurogenesis	GO:0022008	3.436×10^{-4}	
neuron differentiation	GO:0030182	4.367×10^{-4}	
cell projection organization	GO:0030030	4.880×10^{-4}	
cellular component organization	GO:0016043	1.058×10^{-3}	
cell junction organization	GO:0034330	1.959×10^{-3}	
cellular component organization or biogenesis	GO:0071840	2.660×10^{-3}	
nervous system development	GO:0007399	3.212×10^{-3}	
regulation of nervous system development	GO:0051960	9.393×10^{-3}	
regulation of multicellular organismal development	GO:2000026	1.188×10^{-2}	
axon development	GO:0061564	1.233×10^{-2}	
regulation of cellular component organization	GO:0051128	1.483×10^{-2}	
negative regulation of cellular component organization	GO:0051129	1.824×10^{-2}	
regulation of plasma membrane bounded cell projecti...	GO:0120035	2.016×10^{-2}	
regulation of neurogenesis	GO:0050767	2.287×10^{-2}	
regulation of cell projection organization	GO:0031344	2.535×10^{-2}	
regulation of neuron projection development	GO:0010975	3.087×10^{-2}	
negative regulation of cell projection organization	GO:0031345	3.316×10^{-2}	
regulation of cell development	GO:0060284	3.871×10^{-2}	

GO:CC

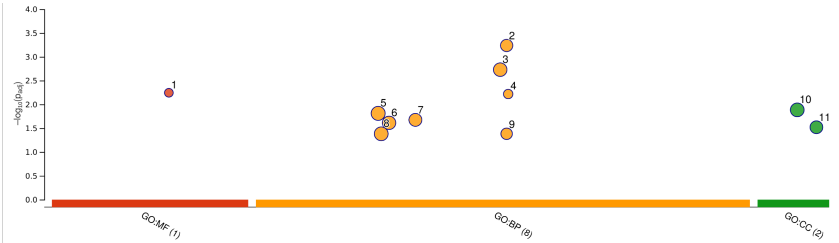
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
synapse	GO:0045202	1.557×10^{-3}	
postsynapse	GO:0098794	3.125×10^{-3}	
methyltransferase complex	GO:0034708	1.798×10^{-2}	
histone methyltransferase complex	GO:0035097	1.925×10^{-2}	
intracellular	GO:0005622	3.087×10^{-2}	
intracellular membrane-bounded organelle	GO:0043231	3.383×10^{-2}	

Figure 2.16: Table of all terms Education Discovery gProfiler

2.4.5.3 SynGO

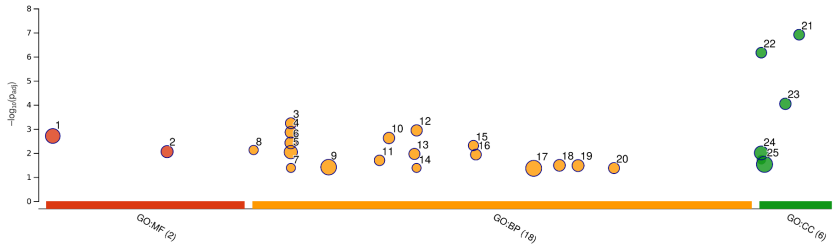
Synapse, GO:0045202, was the only enriched cellular component term using SynGO (table 2.33). SynGO annotates only 23 of the significant genes to synapse, compared to 40 genes in ToppGene table 2.30. Six biological process terms are enriched, including synapse organisation GO:0050807(table 2.32). Fifteen terms are annotated synapse organisation by SynGO compared with 19 in ToppGene (table 2.29 but the limited ontology results in a more significant FDR (6.69×10^{-5} compared to 0.018). These results show both the benefits and limitations of specialised ontologies.

Using FUMA to analyse gene expression the GWGAS significant $\text{Education}_{\text{Replication}}$ genes are over-represented amongst differentially expressed genes up-regulated in the brain using GTEx (figure 2.18). This is the only sample in which this is the case (see figure 2.22 and figure 2.21 which shows only over-representation).



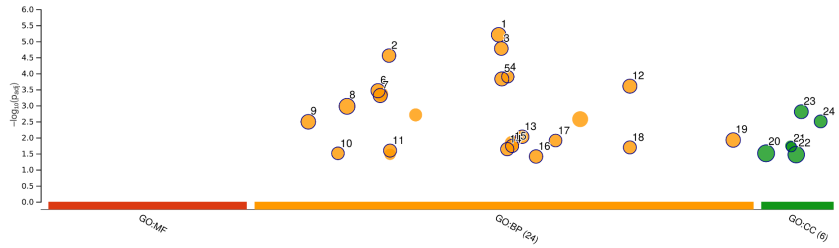
ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0045499	chemorepellent activity	5.698×10^{-3}
2	GO:BP	GO:0050808	synapse organization	5.780×10^{-4}
3	GO:BP	GO:0048666	neuron development	1.863×10^{-3}
4	GO:BP	GO:0050919	negative chemotaxis	6.055×10^{-3}
5	GO:BP	GO:0022008	neurogenesis	1.544×10^{-2}
6	GO:BP	GO:0031175	neuron projection development	2.438×10^{-2}
7	GO:BP	GO:0034330	cell junction organization	2.115×10^{-2}
8	GO:BP	GO:0030182	neuron differentiation	4.147×10^{-2}
9	GO:BP	GO:0050807	regulation of synapse organization	4.141×10^{-2}
10	GO:CC	GO:0045202	synapse	1.304×10^{-2}
11	GO:CC	GO:0098794	postsynapse	3.019×10^{-2}

(a) gProfiler. Education Replication



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0003677	DNA binding	1.968×10^{-3}
2	GO:MF	GO:0046982	protein heterodimerization activity	8.787×10^{-3}
3	GO:BP	GO:0006334	nucleosome assembly	5.819×10^{-4}
4	GO:BP	GO:0006333	chromatin assembly or disassembly	1.381×10^{-3}
5	GO:BP	GO:0006325	chromatin organization	9.204×10^{-3}
6	GO:BP	GO:0006323	DNA packaging	3.800×10^{-3}
7	GO:BP	GO:0006335	DNA replication-dependent nucleosome assembly	4.166×10^{-2}
8	GO:BP	GO:0000183	rDNA heterochromatin assembly	7.514×10^{-3}
9	GO:BP	GO:0010468	regulation of gene expression	3.874×10^{-2}
10	GO:BP	GO:0031497	chromatin assembly	2.384×10^{-3}
11	GO:BP	GO:0030219	megakaryocyte differentiation	2.018×10^{-2}
12	GO:BP	GO:0034728	nucleosome organization	1.152×10^{-3}
13	GO:BP	GO:0034401	chromatin organization involved in regulation of tr...	1.094×10^{-2}
14	GO:BP	GO:0034723	DNA replication-dependent nucleosome organizati...	4.166×10^{-2}
15	GO:BP	GO:0045652	regulation of megakaryocyte differentiation	4.882×10^{-3}
16	GO:BP	GO:0045814	negative regulation of gene expression, epigenetic	1.168×10^{-2}
17	GO:BP	GO:0060255	regulation of macromolecule metabolic process	4.332×10^{-2}
18	GO:BP	GO:0065004	protein-DNA complex assembly	3.258×10^{-2}
19	GO:BP	GO:0071824	protein-DNA complex subunit organization	3.336×10^{-2}
20	GO:BP	GO:0097549	chromatin organization involved in negative regula...	4.234×10^{-2}
21	GO:CC	GO:0044815	DNA packaging complex	1.221×10^{-7}
22	GO:CC	GO:0000786	nucleosome	6.819×10^{-7}
23	GO:CC	GO:0032993	protein-DNA complex	9.050×10^{-5}
24	GO:CC	GO:0000785	chromatin	9.899×10^{-3}
25	GO:CC	GO:0005634	nucleus	2.959×10^{-2}

(b) gProfiler Intelligence Discovery



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:BP	GO:0048468	cell development	6.223×10^{-6}
2	GO:BP	GO:0031175	neuron projection development	2.772×10^{-5}
3	GO:BP	GO:0048666	neuron development	1.668×10^{-5}
4	GO:BP	GO:0050808	synapse organization	1.262×10^{-4}
5	GO:BP	GO:0048699	generation of neurons	1.480×10^{-4}
6	GO:BP	GO:0022008	neurogenesis	3.436×10^{-4}
7	GO:BP	GO:0030030	cell projection organization	4.880×10^{-4}
8	GO:BP	GO:0016043	cellular component organization	1.058×10^{-3}
9	GO:BP	GO:0007399	nervous system development	3.212×10^{-3}
10	GO:BP	GO:0010975	regulation of neuron projection development	3.087×10^{-2}
11	GO:BP	GO:0031344	regulation of cell projection organization	2.535×10^{-2}
12	GO:BP	GO:0120036	plasma membrane bounded cell projection organi...	2.489×10^{-4}
13	GO:BP	GO:0051960	regulation of nervous system development	9.393×10^{-3}
14	GO:BP	GO:0050767	regulation of neurogenesis	2.287×10^{-2}
15	GO:BP	GO:0051129	negative regulation of cellular component organiz...	1.824×10^{-2}
16	GO:BP	GO:0060284	regulation of cell development	3.871×10^{-2}
17	GO:BP	GO:0061564	axon development	1.233×10^{-2}
18	GO:BP	GO:0120035	regulation of plasma membrane bounded cell proj...	2.016×10^{-2}
19	GO:BP	GO:2000026	regulation of multicellular organismal development	1.188×10^{-2}
20	GO:CC	GO:0005622	intracellular	3.087×10^{-2}
21	GO:CC	GO:0035097	histone methyltransferase complex	1.925×10^{-2}
22	GO:CC	GO:0043231	intracellular membrane-bounded organelle	3.383×10^{-2}
23	GO:CC	GO:0045202	synapse	1.557×10^{-3}
24	GO:CC	GO:0098794	postsynapse	3.125×10^{-3}

(c) gProfiler Education Discovery

Figure 2.17: Gene Ontology enrichment of significant genes using gProfiler

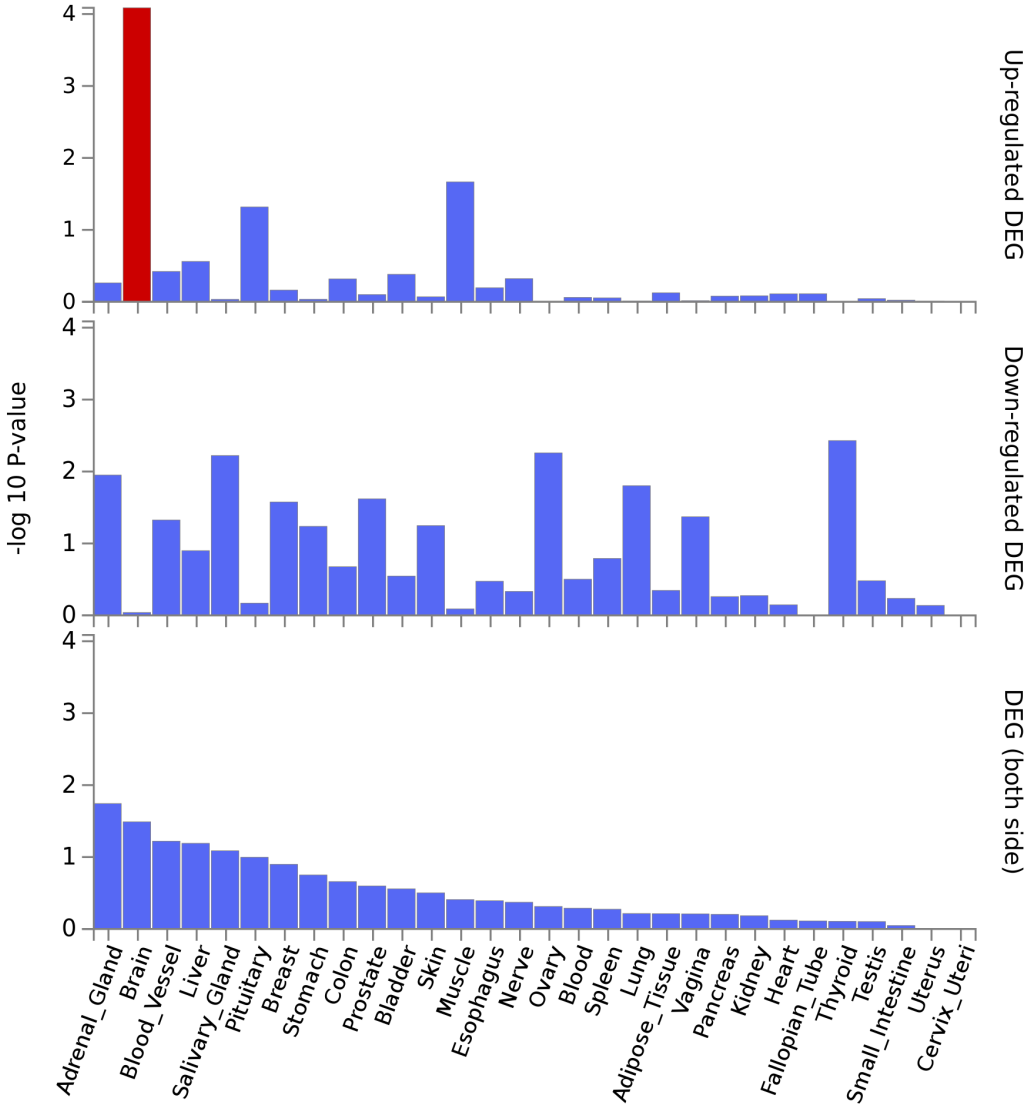


Figure 2.18: Gtex Analysis of Gene expression of genome wide significant Education Discovery genes shows enrichment for up-regulated differentially expressed genes in brain. Red columns are significantly enriched

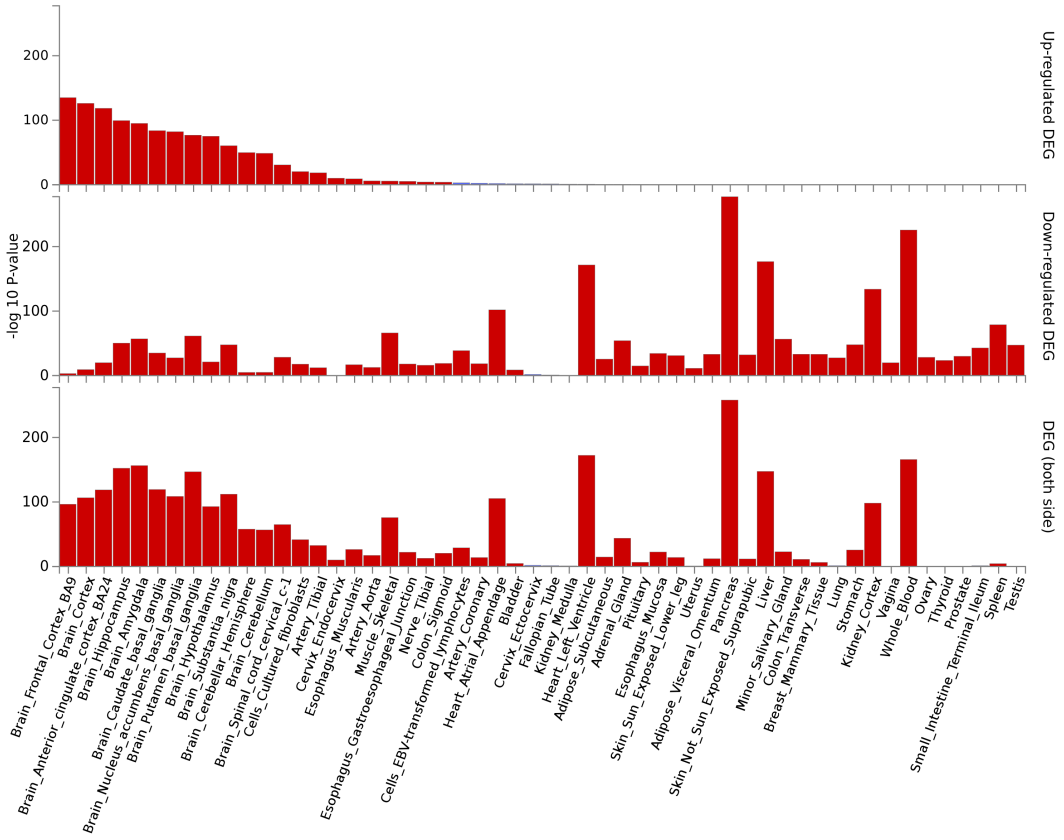


Figure 2.19: GTex all PSP. Note top row of upregulated genes. Shows that signal from expanded PSP is very specific to CNS expression. Red significant. Y axis -log10 of p value. Hypergeometric test of genes differentially expressed from pre-created list see FUMA [428] for details (two sided Bonferroni corrected p versus all labels with cut off fold change)

GO	Description	Ref	Test	E	Fold	p value	FDR
GO:0044815	DNA packaging complex	85	9	0.8	11.50	1.98×10^{-7}	0.0002
GO:0000786	nucleosome	77	8	0.7	11.28	1.10×10^{-6}	0.0007
GO:0032993	protein-DNA complex	175	9	1.6	5.59	5.03×10^{-5}	0.0056
GO:0000785	chromatin	1504	36	13.8	2.60	1.50×10^{-7}	0.0003
GO:0030659	cytoplasmic vesicle membrane	788	18	7.3	2.48	4.39×10^{-4}	0.0367
GO:0012506	vesicle membrane	810	18	7.5	2.41	6.01×10^{-4}	0.0483
GO:0000790	nuclear chromatin	1316	29	12.1	2.39	1.66×10^{-5}	0.0030
GO:0005694	chromosome	2012	40	18.5	2.16	3.59×10^{-6}	0.0018
GO:0045202	synapse	1383	27	12.7	2.12	2.28×10^{-4}	0.0199
GO:0000228	nuclear chromosome	1542	30	14.2	2.11	1.36×10^{-4}	0.0124
GO:0031981	nuclear lumen	4993	73	46.0	1.59	1.84×10^{-5}	0.0031
GO:0043232	intracellular non-membrane-bounded organelle	5274	77	48.6	1.59	8.30×10^{-6}	0.0021
GO:0043228	non-membrane-bounded organelle	5283	77	48.6	1.58	8.50×10^{-6}	0.0019
GO:0043233	organelle lumen	6114	82	56.3	1.46	8.88×10^{-5}	0.0094
GO:0070013	intracellular organelle lumen	6114	82	56.3	1.46	8.88×10^{-5}	0.0089
GO:0031974	membrane-enclosed lumen	6114	82	56.3	1.46	8.88×10^{-5}	0.0085
GO:0005634	nucleus	7578	101	69.8	1.45	5.29×10^{-6}	0.0018
GO:0043231	intracellular membrane-bounded organelle	11041	132	101.7	1.30	1.14×10^{-5}	0.0023
GO:0043229	intracellular organelle	12856	148	118.4	1.25	6.65×10^{-6}	0.0019
GO:0043227	membrane-bounded organelle	12748	146	117.4	1.24	1.91×10^{-5}	0.0029
GO:0043226	organelle	13868	155	127.7	1.21	2.01×10^{-5}	0.0029
GO:0005622	intracellular	14704	163	135.4	1.20	4.77×10^{-6}	0.0019
GO:0110165	cellular anatomical entity	18767	188	172.8	1.09	2.91×10^{-5}	0.0039
GO:0005575	cellular_component	18951	189	174.5	1.08	3.56×10^{-5}	0.0045
UNCLASSIFIED	Unclassified	1900	3	17.5	0.17	3.56×10^{-5}	0.0042

Table 2.23: GO cellular component complete Intelligence Discovery FDR over represenation only. Ref reference set, test:number of genes being tested present in ontology term, E expected number of genes being tested present in ontology term, Fold= Fold change, P = raw p value, FDR = false discovery rate

RProjects/paper_xls_output/R/chapter_2/make_PANTHER_tables/UKBB_int/make_Panther_table_GO_col_nounderFDR_int.R

GO-Slim Molecular Function	n annotation	n	expected	fold	p value	FDR
transcription factor binding	203	10	1.87	5.35	2.73×10^{-5}	0.0145
DNA binding	806	20	7.42	2.69	7.11×10^{-5}	0.0189

Table 2.24: Intelligence Discovery Molecular Function SLIM PANTHER. n annotation= number of genes in annotation; n= number of significant genes; fold = fold change; FDR = false discovery rate Benjamini and Hochberg

PANTHER Protein Class	n annot	n sample	expected	Fold	P	FDR
histone	55	7	.51	13.82	1.52×10^{-6}	2.96×10^{-4}
DNA binding protein	227	9	2.09	4.31	3.30×10^{-4}	0.032

Table 2.25: PANTHER protein class enrichment $\text{Intelligence}_{\text{Discovery}}$. n annot = number in annotation, Fold = fold change

Term	n	p	FDR q
synapse organization	11	1.81×10^{-4}	1.81×10^{-3}
process in the synapse	17	3.01×10^{-3}	0.0107
regulation of synapse organization	3	3.93×10^{-3}	0.0107
modification of postsynaptic structure	3	4.29×10^{-3}	0.0107
regulation of presynaptic membrane potential	3	9.62×10^{-3}	0.0192

Table 2.26: SynGO Enrichment $\text{Intelligence}_{\text{Discovery}}$ sample of genome wide significant genes at GWGAS. Correction for multiple comparisons using FDR. Background gene set is all brain expressed genes.

GO	Description	Ref	Test	E	Fold	p	FDR
GO:0042613	MHC class II protein complex	19	5	0.2	23.06	6.26×10^{-6}	0.0013
GO:0071556	integral component of lumenal side of ER membrane	29	6	0.3	18.13	2.39×10^{-6}	0.0008
GO:0098553	lumenal side of ER membrane	29	6	0.3	18.13	2.39×10^{-6}	0.0007
GO:0098576	lumenal side of membrane	37	7	0.4	16.57	5.73×10^{-7}	0.0004
GO:0042611	MHC protein complex	28	5	0.3	15.64	3.22×10^{-5}	0.0043
GO:0035097	histone methyltransferase complex	81	7	0.9	7.57	6.23×10^{-5}	0.0069
GO:0012507	ER to Golgi transport vesicle membrane	61	5	0.7	7.18	8.96×10^{-4}	0.0473
GO:0098982	GABA-ergic synapse	79	6	0.9	6.65	4.03×10^{-4}	0.0253
GO:0030134	COPII-coated ER to Golgi transport vesicle	96	7	1.1	6.39	1.69×10^{-4}	0.0141
GO:0034708	methyltransferase complex	110	8	1.3	6.37	5.89×10^{-5}	0.0070
GO:0032588	trans-Golgi network membrane	95	6	1.1	5.53	1.01×10^{-3}	0.0493
GO:0030660	Golgi-associated vesicle membrane	112	7	1.3	5.48	4.10×10^{-4}	0.0249
GO:0005798	Golgi-associated vesicle	184	10	2.1	4.76	7.46×10^{-5}	0.0079
GO:0030176	integral component of ER membrane	164	8	1.9	4.27	7.71×10^{-4}	0.0442
GO:0031227	intrinsic component of ER membrane	172	8	2.0	4.07	1.04×10^{-3}	0.0495
GO:0030658	transport vesicle membrane	212	9	2.4	3.72	9.49×10^{-4}	0.0476
GO:0098852	lytic vacuole membrane	383	14	4.4	3.20	1.80×10^{-4}	0.0144
GO:0005765	lysosomal membrane	383	14	4.4	3.20	1.80×10^{-4}	0.0139
GO:0014069	postsynaptic density	345	12	3.9	3.05	7.87×10^{-4}	0.0439
GO:0005774	vacuolar membrane	436	15	5.0	3.01	2.00×10^{-4}	0.0149
GO:0032279	asymmetric synapse	351	12	4.0	3.00	9.09×10^{-4}	0.0468
GO:0098794	postsynapse	651	22	7.4	2.96	8.58×10^{-6}	0.0016
GO:0045202	synapse	1383	36	15.8	2.28	5.18×10^{-6}	0.0012
GO:0000139	Golgi membrane	769	20	8.8	2.28	7.70×10^{-4}	0.0455
GO:0043005	neuron projection	1395	33	15.9	2.07	7.85×10^{-5}	0.0079
GO:0030054	cell junction	2122	44	24.2	1.82	1.39×10^{-4}	0.0121
GO:0005654	nucleoplasm	3990	67	45.5	1.47	8.77×10^{-4}	0.0476
GO:0005829	cytosol	5314	88	60.7	1.45	9.63×10^{-5}	0.0092
GO:0032991	protein-containing complex	5532	89	63.1	1.41	2.83×10^{-4}	0.0203
GO:0043233	organelle lumen	6114	96	69.8	1.38	3.29×10^{-4}	0.0228
GO:0070013	intracellular organelle lumen	6114	96	69.8	1.38	3.29×10^{-4}	0.0220
GO:0031974	membrane-enclosed lumen	6114	96	69.8	1.38	3.29×10^{-4}	0.0213
GO:0005634	nucleus	7578	116	86.5	1.34	1.07×10^{-4}	0.0097
GO:0043231	intracellular membrane-bounded organelle	11041	166	126.0	1.32	1.94×10^{-7}	0.0002
GO:0043229	intracellular organelle	12856	183	146.7	1.25	8.00×10^{-7}	0.0004
GO:0043227	membrane-bounded organelle	12748	180	145.5	1.24	3.31×10^{-6}	0.0008
GO:0005737	cytoplasm	11717	165	133.7	1.23	4.24×10^{-5}	0.0053
GO:0005622	intracellular	14704	205	167.8	1.22	2.44×10^{-8}	0.0000
GO:0043226	organelle	13 868	193	158.3	1.22	1.04×10^{-6}	0.0004

Table 2.27: GO cellular component complete Education Discovery FDRover represenation only Ref reference set, test:number of genes being tested present in ontology termE expected number of genes being tested ,present in ontology term, Fold= Fold changeP = raw p value, FDR = false discovery rate. ER = endoplasmic reticulum Ordered by fold change.GO:0110165 Cellular anatomical entity and GO:0005575 cellular component omitted along with one unclassified term.

GO	description	Ref	Test	E	Fold	P	FDR
R-HSA-202430	Translocation of ZAP-70 to Immunological synapse	25	5	0.3	17.52	1.99×10^{-5}	0.0454
R-HSA-202427	Phosphorylation of CD3 and TCR zeta chains	28	5	0.3	15.64	3.22×10^{-5}	0.0368
R-HSA-389948	PD-1 signaling	29	5	0.3	15.11	3.74×10^{-5}	0.0285
R-HSA-202433	Generation of second messenger molecules	40	5	0.5	10.95	1.48×10^{-4}	0.0484
R-HSA-3247509	Chromatin modifying enzymes	240	11	2.7	4.02	1.40×10^{-4}	0.0639
R-HSA-4839726	Chromatin organization	240	11	2.7	4.02	1.40×10^{-4}	0.0533
R-HSA-194315	Signaling by Rho GTPases	417	15	4.8	3.15	1.26×10^{-4}	0.0717

Table 2.28: Reactome pathways Education Discovery FDRover represenation only Ref reference set, test:number of genes being tested present in ontology termE expected number of genes being tested ,present in ontology term, Fold= Fold changeP = raw p value, FDR = false discovery rate

ID	Name	n	n annot	p	FDR BH	Bonferroni
GO:0048666	neuron development	38	1297	6.045×10^{-8}	2.629×10^{-4}	0.0003
GO:0030182	neuron differentiation	42	1578	1.580×10^{-7}	3.299×10^{-4}	0.0007
GO:0031175	neuron projection development	34	1150	2.686×10^{-7}	3.299×10^{-4}	0.0012
GO:0022008	neurogenesis	46	1866	3.362×10^{-7}	3.299×10^{-4}	0.0015
GO:0048699	generation of neurons	44	1751	3.792×10^{-7}	3.299×10^{-4}	0.0016
GO:0007417	central nervous system development	33	1129	5.320×10^{-7}	3.760×10^{-4}	0.0023
GO:0051129	negative regulation of cellular component organization	27	819	6.635×10^{-7}	3.760×10^{-4}	0.0029
GO:0051960	regulation of nervous system development	32	1087	6.916×10^{-7}	3.760×10^{-4}	0.0030
GO:0060284	regulation of cell development	32	1132	1.655×10^{-6}	7.231×10^{-4}	0.0072
GO:0120036	plasma membrane bounded cell projection organization	43	1789	1.721×10^{-6}	7.231×10^{-4}	0.0075
GO:0050767	regulation of neurogenesis	29	971	1.829×10^{-6}	7.231×10^{-4}	0.0080
GO:0030030	cell projection organization	43	1827	2.968×10^{-6}	1.075×10^{-3}	0.0129
GO:0050808	**synapse organization**	19	498	4.261×10^{-6}	1.425×10^{-3}	0.0185
GO:0120035	regulation of plasma membrane bounded cell projection organization	25	821	7.140×10^{-6}	2.218×10^{-3}	0.0311
GO:0031344	regulation of cell projection organization	25	831	8.787×10^{-6}	2.548×10^{-3}	0.0382
GO:0031345	negative regulation of cell projection organization	12	226	1.084×10^{-5}	2.921×10^{-3}	0.0471
GO:0061564	axon development	20	583	1.142×10^{-5}	2.921×10^{-3}	0.0496

Table 2.29: Biological Process Education Discovery ToppGene Bonferroni and BY significant results n=17

RProjects/paper_xls_output/R/chapter_2/toppgene/do_tables/recent/topp_gene_bp_education_discovery_and_bonfe.R)

ID	Name	n	n annot	p	FDR BH	Bonferroni
GO:0045202	**synapse**	40	1482	1.881×10^{-7}	9.876×10^{-5}	0.0001
GO:0098794	postsynapse	26	825	2.158×10^{-6}	5.666×10^{-4}	0.0011
GO:0098553	lumenal side of endoplasmic reticulum membrane	5	29	1.625×10^{-5}	2.132×10^{-3}	0.0085
GO:0071556	integral component of lumenal side of ER membrane	5	29	1.625×10^{-5}	2.132×10^{-3}	0.0085
GO:0042613	MHC class II protein complex	4	16	2.530×10^{-5}	2.462×10^{-3}	0.0133
GO:0043005	neuron projection	37	1624	2.814×10^{-5}	2.462×10^{-3}	0.0148
GO:0098576	lumenal side of membrane	5	34	3.637×10^{-5}	2.728×10^{-3}	0.0191
GO:0005798	Golgi-associated vesicle	10	186	4.982×10^{-5}	3.269×10^{-3}	0.0261
GO:0034708	methyltransferase complex	8	121	6.751×10^{-5}	3.938×10^{-3}	0.0354
GO:0035097	histone methyltransferase complex	7	94	9.152×10^{-5}	4.400×10^{-3}	0.0481

Table 2.30: Cellular component Education Discovery ToppGene top10 these are those of bonferroni significance. Identical to graph above other than including bonferroni results. ER = endoplasmic reticulum

RProjects/paper_xls_output/R/chapter_2/toppgene/do_tables/toppgene_ukbbed_cc.R

Gene ID	Gene Symbol	Gene Name
5144	PDE4D	phosphodiesterase 4D
5662	PSD	pleckstrin and Sec7 domain containing
4137	MAPT	microtubule associated protein tau
7248	TSC1	TSC complex subunit 1
4208	MEF2C	myocyte enhancer factor 2C
120	ADD3	adducin 3
9378	NRXN1	neurexin 1
257194	NEGR1	neuronal growth regulator 1
23334	SZT2	SZT2 subunit of KICSTOR complex
22907	DHX30	DExH-box helicase 30
8085	KMT2D	lysine methyltransferase 2D

Table 2.31: Genes associated with intellectual disability found in PSP C3714756 DisGeNET. ORA in ToppGene. Education Discovery. Gene ID = Entrez Gene Gene ID

Cellular component	n	p	FDRq
synapse organization	15	7.44×10^{-6}	6.69×10^{-5}
process in the synapse	20	4.06×10^{-3}	0.0183
trans-synaptic signaling	6	0.0234	0.0428
synapse adhesion between pre- and post-synapse	3	0.0146	0.0428
regulation of synapse assembly	3	0.0238	0.0428
synapse assembly	4	0.0289	0.0433

Table 2.32: SynGO Biological Process Education$_{\text{Discovery}}$

Cellular component	n	p	FDRq
synapse	23	0.00451	0.0316
postsynapse	14	0.0245	0.0858

Table 2.33: syngo CC Education Discovery

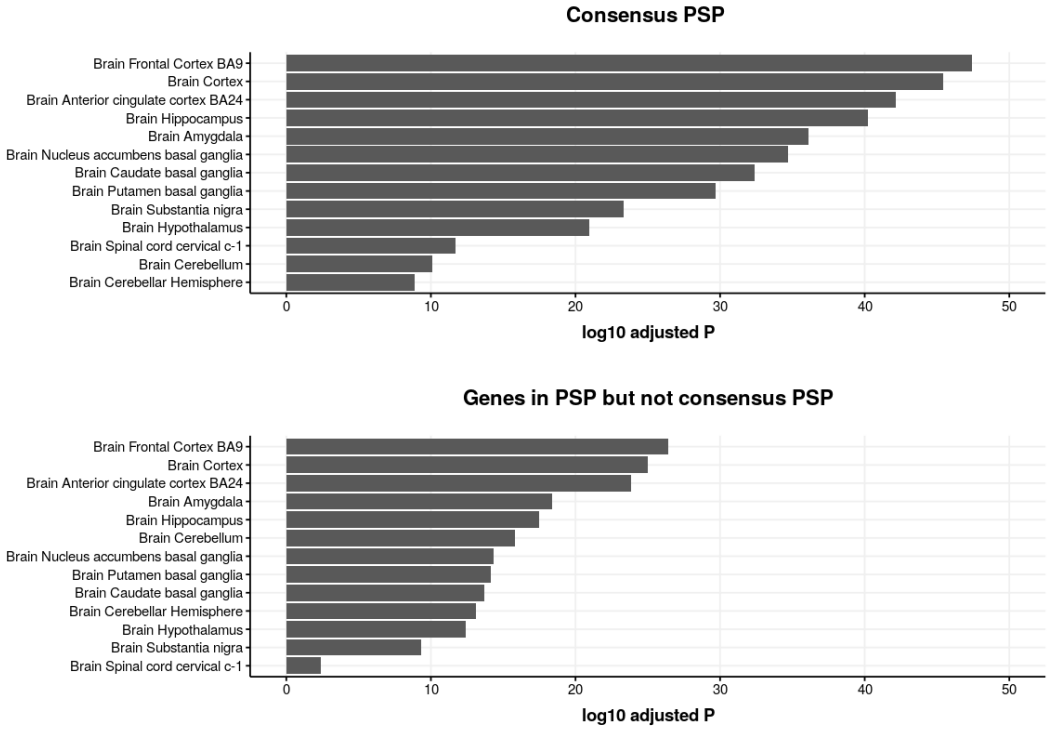
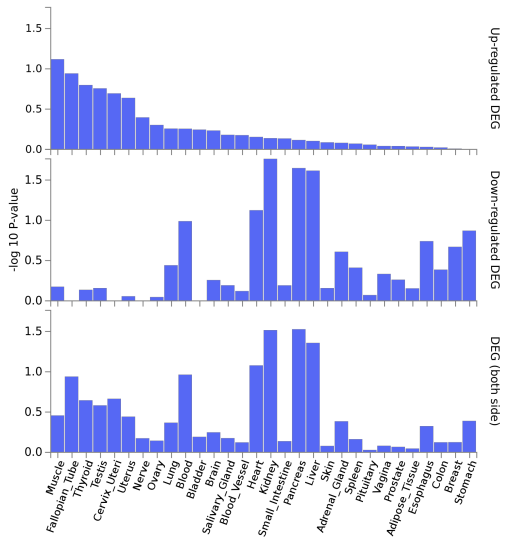


Figure 2.20: Differentially expressed genes over representation analysis in Gtex. Top panel genes in consensus PSD $n = 1312$. Lower panel genes not in consensus PSD but in PSP.

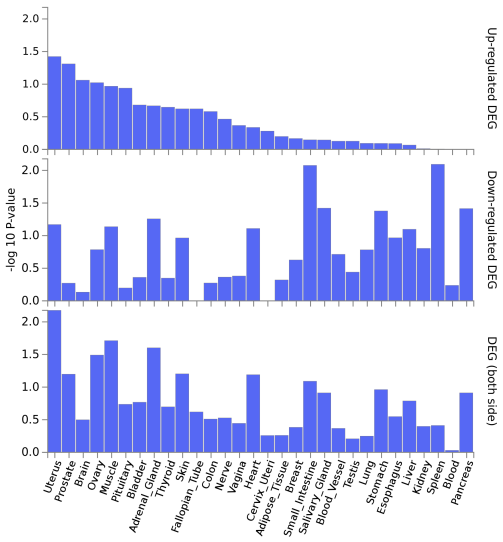
RProjects/paper_xls_output/R/chapter_2/FUMA/plot_gtex/plot_consensus_multiplot_theme.R



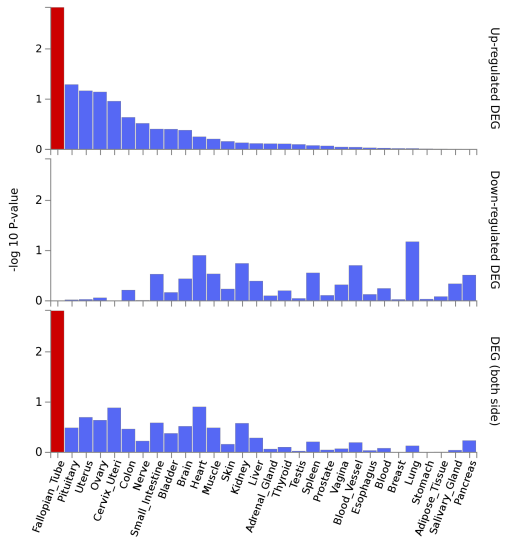
Figure 2.21: Over representation of differentially expressed genes (up regulated) amongst significant genes at GWGAS for each of the samples clockwise from top left Intelligence Replication, Intelligence Discovery, Education Discovery and Education replication using FUMA to perform over representation. Gene expression data GTEx v8. x axis is -log10 transform of p value for cumulative hypergeometric test. Significant tissues are shown in cyan. The only sample with over representation of genes with up-regulated differential expression in the brain is $\text{Education}_{\text{Replication}}$ `RProjects/paper_xls_output/R/chapter_2/FUMA/plot_gtex/plot_study/multiplot_samples.R`



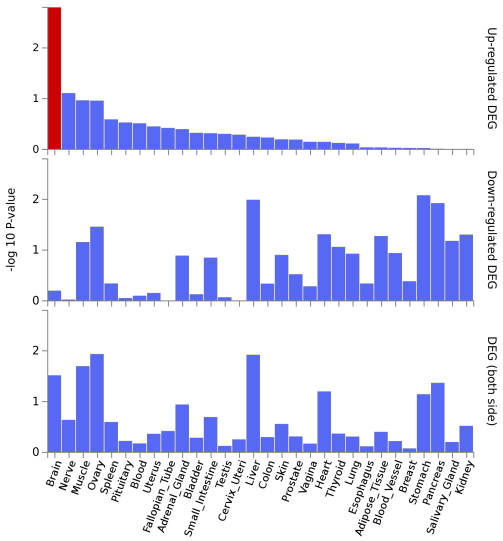
(a) Gene expression Intelligence Replication



(b) Education Replication

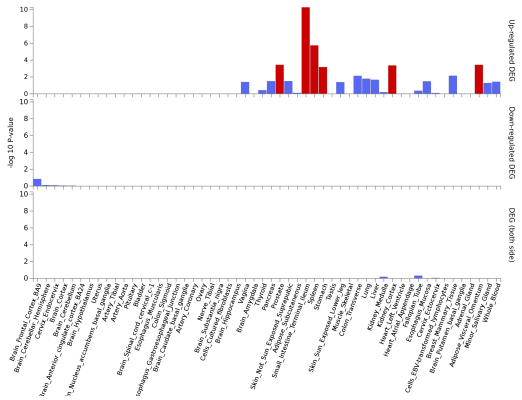


(c) DEG Up Regulated Intelligence Discovery

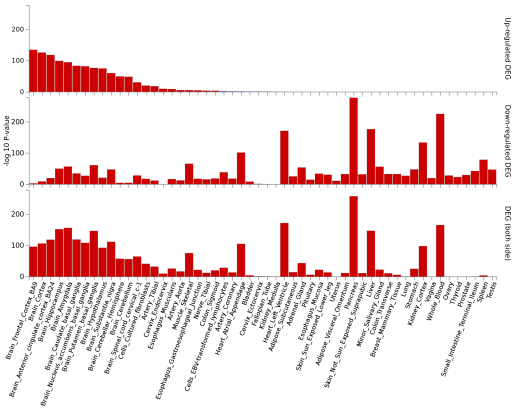


(d) Education Discovery

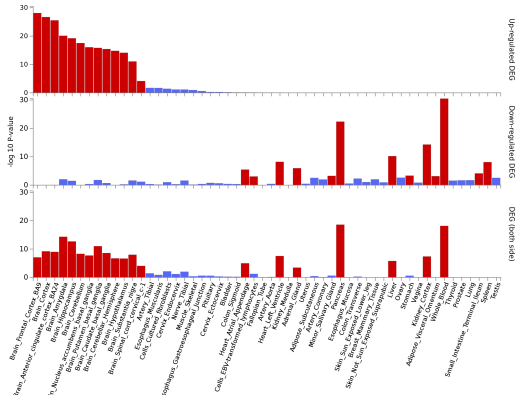
Figure 2.22: FUMA output. Enrichment for differentially expressed genes ordered by up regulated genes. Only education discovery shows evidence of up regulated differentially expressed genes in CNS



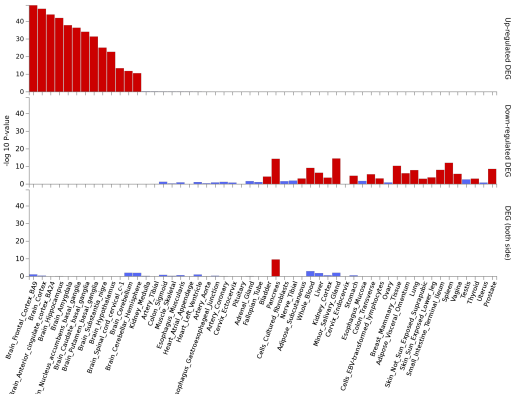
(a) Gene expression genes not in PSP



(b) Genes in PSP



(c) Genes in PSP but not in consensus PSD. ie new genes



(d) Genes in consensus PSD

Figure 2.23: Comparison of gene expression in genes not in the PSP, genes in the PSP, genes in PSP but not in the consensus PSD (new genes), genes in the consensus PSD - that which is conserved in different vertebrate species. Images from FUMA. GTEx v8 54 tissue types. Ordered from left to right by up-regulated differentially expressed genes p values (hypergeometric test). Tissues with significant differential enrichment after bonferroni correction for multiple comparisons are shown in blue. Y axis -log10(p) of enrichment for gene set (eg PSP)

2.5 Synaptic component enrichment MAGMA

Functional synaptic group gene sets are curated at the Complex Trait Genetics lab [46] based on stuies by Ruano et al. (2012)[363] and Lips et al. (2012)[250]. Using these sets with MAGMA will provide an estimate of how effective a purely functional annotation approach is and an estimate of the value of significance tests in these samples.

MAGMA GSA was performed using the raw output of the GWGAS analyses in MAGMA as part of a standard two stage MAGMA GSA. Default options were used.

The annotated set was composed of 1,028 genes. 736 of these were in the PSP, 292 were not. A large number of the missing ones were non glutamatergic receptors and the proteins associated with these. There are 18 gene sets each a functional group such as G-protein-coupled receptor signalling.

MAGMA accepts gene sets in gene matrix transposed format (gmt) where each gene set is a row in a text file that is tab separated. gmt files used for MAGMA enrichment are often those curated by MSigDB. I have used an Rscript to create a gmt file from the supplied data[47].

To reduce type II error, interesting sets are identified as those of nominal significance in the discovery samples and these selected sets are then tested using the replication samples.

2.5.0.0.1 Results For the $\text{Intelligence}_{\text{Discovery}}$ sets Cell adhesion and trans-synaptic signaling p=0.009 , Excitability 0.005 and Exocytosis p=0.048 are of potential interest having $p < 0.05$ (supplementary table B.16). Only exocytosis is of nominal significance in the $\text{Intelligence}_{\text{Replication}}$ cohort (p=0.016; p= 0.048 with Bonferroni correction see table B.18).

For $\text{Education}_{\text{Discovery}}$ two sets achieve nominal levels of significance: Cell adhesion and trans-synaptic signaling p=0.001 and Excitability n=57 genes 0.003 (table B.17). None of these achieve nominal significance in the $\text{Education}_{\text{Replication}}$ cohort (table B.19).

The only significant group we have identified using the curated synaptic sets provided by the CTG lab and our samples are Exocytosis for the Intelligence sample. The test statistic in the discovery cohort (0.48) is close to the nominal significance ($p<0.05$) threshold. Enrichment by functional class has been of limited benefit, but has proven a

[46] https://ctg.cncr.nl/software/GENESETS/synapse.zip

[47] RProjects/paper_xls_output/R/magma_synapse/make_synapse_gmt.R

Term	Phenotype	n	Discovery	Replication
Cell adhesion and trans-synaptic signalling	Intelligence	76	0.009	0.155
Excitability	Intelligence	57	0.005	0.159
Exocytosis	Intelligence	83	**0.048**	**0.016**
Excitability	Education	57	0.003	0.122
Cell adhesion and trans-synaptic signalling	Education	76	0.001	0.067

Table 2.34: Discovery and Replication cohorts of curated synaptic genes from CTG site.

Group	n	Min.	1Q	Median	Mean	3Q	Max.
all sets	10,673	3.0	17.0	34.0	100.4	88.0	1898.0
$p < 0.05$	850	4.0	20.0	52.0	180.2	155.8	1898.0
$p < 0.001$	50	6.0	20.5	61.0	320.7	439.2	1674.0

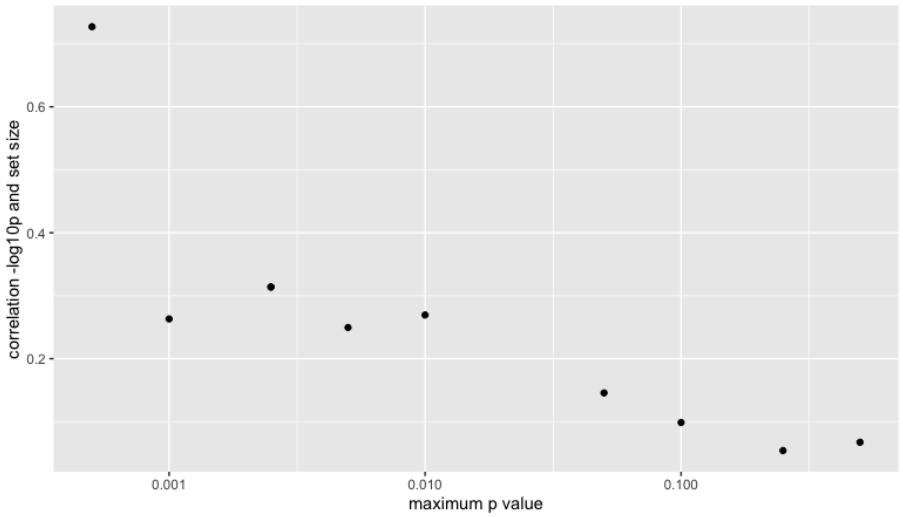
Table 2.35: Summary statistics for gene set size for different levels of significance from Hill et al. (2019)[180]

useful guide to the likely significance levels found in MAGMA GSA of synaptic sets. See table 2.34

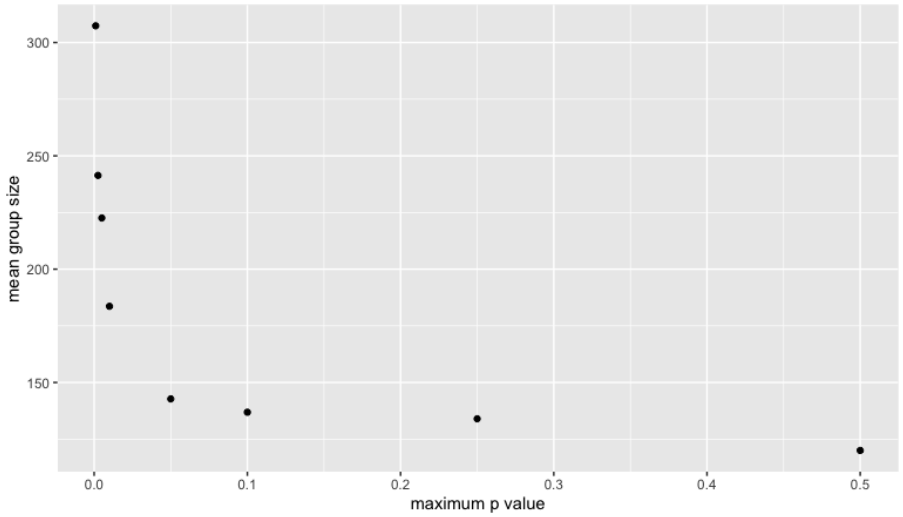
2.5.0.1 Gene size and enrichment intelligence studies MAGMA

Nominally significant gene sets are larger using 0.05 and size 50 as cutoff cor test 0.22 $p = 2.13 \times 10^{-6}$ [48].

[48] `RProjects/paper_xls_output/R/exploratory_data_analysis/gene_size_and_enrichment_magma_sets.R`



(a) Correlation of gene significance and set size with decreasing p. X axis maximum p value, y Pearsons correlation between -log10(p) of sets and set size. Data from supplementary material Sniekers et al[383]



(b) Changes in mean gene set size with decreasing maximum p values. Mean gene size increases.

Figure 2.24: Correlation between gene size and level of significance in MAGMA GSA and changes in mean gene size with increasingly significant gene sets (max p).Data from supplementary material Sniekers et al[383]

2.6 Summary

The results obtained using MAGMA GSA correspond closely to those reported in the literature despite possible variations in versions, platforms and reference panels. In addition, results from the PASCAL scoring algorithm correlate closely with those from MAGMA.

Most of the potential PSP genes identified by proteomics are in the large connected component. Isolated genes, excluded from analysis, are few and show no evidence of possessing shared features that would bias results. Similarly, only a few PSP genes do not have boundaries defined in the reference gene boundaries used for MAGMA, and again, they show no evidence of common features.

There are enriched disease terms related to intellectual disability in both education samples and neither intelligence sample. An association is reported in Sniekers et al. [383] using DEPICT but I am not aware that the difference with intelligence samples has been reported. It suggests a qualitative difference between enriched genes in the education and intelligence samples and that we may expect some subsequent findings to hold for only one of the phenotypes.

PSP genes are expressed robustly in all brain regions using the GTEx data, and this suggests that the increase in gene number in the network has still resulted in a set of genes strongly expressed in the CNS. The pattern of correlation of gene expression between brain regions in the PSP is similar to that recorded of all genes literature [269] suggesting that we do not have a very unusual sample.

The difference between the findings for MAGMA GSA and ORA of significant genes in the intelligence discovery cohort show the limitations that network approaches limited to over-representation analysis face[144]. I found ToppGene[72] had the best overall performance of a Gene Ontology enrichment package now that it accepts custom background sets.

The relative lack of statistically significant gene associations in published studies using MAGMA GSA suggests it would be reasonable to reduce unnecessary comparisons. A reasonable way to achieve this is to use a discovery and replication cohort design and only test synaptic communities sets we have a reason to believe would be of interest.

Sniekers et al.[383] tested 6,155 gene sets using MAGMA GSA, an extensive search over possible gene sets. The uncorrected p values should provide a rough guide to likely significance levels in using this sample ($\text{Intelligence}_{\text{Replication}}$ - this is the only

sample with published MAGMA GSA and the summary data for the identical population available).

SynGO shows the value of using fewer tests and more specific terms (particularly for $\text{Education}_{\text{Discovery}}$ Biological Process).

Three hundred sixty-five gene sets are of greater than nominal significance (5.91% of terms). There is no reason to expect a *particular* network community defined by topology to be associated with a particular phenotype (intelligence or educational attainment). If there are between 15 and 120 communities (gene sets)[49] in the PSP network (see chapter 4), and if Sniekers et al. is an *approximate* guide to the distribution of MAGMA significance levels, then using nominal significance ($p < 0.05$) to select interesting sets would result in around six genes being tested in the replication samples.

The lowest p value found in any of the samples using the 18 member functionally defined synaptic set was 0.001 (Cell adhesion and trans-synaptic signalling $\text{Education}_{\text{Discovery}}$) and occurred in the best-powered cohort. This group also did not have the constraint of forming a topological unit.

Few significant gene sets are reported in MAGMA GSA studies of intelligence or educational attainment. Those tended to be large (see table 2.35) and I find an association between group size and significance (figure and figure 2.24). The gene sets with very low unadjusted p values in these studies would be larger than we would expect to find in the PSP.

Re-Analysis of the gene-level scores of significant genes in GWGAS studies of intelligence and education suggest that around 25% of these genes are in the postsynaptic proteome. This is not unexpected, many of the genes found in GWAS studies are related to neural development[309] and neurogenesis [180]. Many key genes associated with neurogenesis are not the the synaptic proteome. The gene ontology biological process GO:0022008 Neurogenesis is annotated with 1,940 genes (using ToppGene [72]). Of these only 677 genes are in the PSP (approximately 34.9%).

Neurogenesis occurs prior to synaptogenesis and can have significant effects upon phenotype and function. Nervous tissue derives from ectoderm that undergoes differentiation into neuroectoderm. The neuroectoderm initially divides, migrates to the appropriate tissues and then extends axons and dendrites towards the tissue (under the control of neurotrophic factors)[44].

The checks carried out in this chapter such as the correspondence of MAGMA

[49] a rough estimate extrapolating from McLean et al. (2016)

scores in the analysis pipeline to those in the literature do not show any evidence of systematic bias and appear sound. Therefore in the next chapters I will move on to test the two hypotheses described in chapter 1.

Chapter 3

Do centrality measures in the post synaptic proteome correlate with function?

> I asked a person of intelligence how many steps he thought it would take, and he said that it would require 100 intermediate persons, or more, to move from Nebraska to Sharon. [276]
>
> Stanley Milgram 1969

This chapter addresses measures of network centrality. No association is found between measures of network centrality and the probability of synaptic genes having variants associated with educational ability or intelligence. Central genes are found to be more loss of function intolerant than the rest of the PSP and significantly more so than the rest of the genome. Disease terms associated with neurodegenerative disorders are more commonly associated with central synaptic genes.

Networks can be studied at three levels: the properties of the network as a whole, such as the small-world property (section 3.1.1), the properties of individual nodes, including their importance to the network, and the properties of mesoscale structures within the network such as communities (chapter 4).

The principle node properties are measures of *centrality*. The concept of centrality arose in the social sciences literature in the late 1940s, and its early evolution is reviewed in Freeman [137]. Centrality measures use network topology to determine which nodes are important [289] and where they are (central or peripheral) in the network [137].

Real-world networks often have a small number of nodes that are critical to net-

work function, for example, a hub airport in an air transportation network[52]. Centrality statistics measure different ways in which nodes can be important in a network allowing key nodes to be identified, ranked, and any association between centrality and network function assessed.

There is evidence that genes encoding proteins that are central in the proteome network are more likely to be essential. The most established finding being that null mutations in hub (high degree centrality) genes in the yeast proteome are more likely to be lethal (the *centrality lethality* effect)[197]. Central genes have also been reported to be more commonly associated with diseases, most notably cancer[417], [4],[441] although the relationship is complex and recent results question even the previously accepted degree centrality effects in yeast (section 3.5).

Before addressing the association between synaptic proteome centrality and cognition, it is necessary to introduce the centrality measures, briefly review their previous use in the biomedical literature and their effects on the function of the network.

As the synapse is important in cognition, critical nodes in the synaptic proteome network may be over-represented in animal models of cognition or human differences in intelligence. There is evidence that central nodes affect numerous genes, and, in a protein network, impact *pleiotropic* traits and diseases [71]. The genetic architecture of intelligence is subject to pleiotropic effects[336],[415]. Conversely there may be no effect; genetic variation in critical nodes may be so deleterious as to be lethal, or cause levels of impairment that would absent sufferers from large scale genetic studies of cognitive ability. Central genes have been reported to be essential rather than associated with disease[26]. However there is no necessary dichotomy between a gene being associated with disease and being essential. Recent databases of population level genetic variation have shown that there exists a spectrum of susceptibility to loss of function mutations including mutations resulting in gene products with reduced function (which might give rise to disease causing mutation in a gene regarded as essential)[241],[206]. In addition the definition of variations associated with disease and disease genes varies (see section 4.2.1) in a way in which has not always been recognised in the literature (for example, see [151]).

Therefore, what is needed is an empirical approach to study the effects of network topology on complex traits and diseases. This should include an appreciation that effects may be specific to only some traits or diseases.

3.1 Properties of 'real world' networks

Real world networks often have properties that would not be expected to arise by chance. These include: a power law degree distribution [24],[25] (section 3.2.3) and 'small world' structure [429]. Network level properties in turn arise from the collective properties of individual nodes; for example a degree sequence (node level) gives rise to a power law distribution (network level) or individual local transitivity to global clustering coefficient[429]. In addition to centrality measures (section 3.2), nodes may have properties including non-topological node data[427] or metadata[325] of which the gene level significance scores from GWAS studies are an example. Chapter 4 will address meso-scale structures and community detection.

Network-level statistics reflect how networks form and evolve, their patterns of information flow and how resistant they are to disruption. Not all naturally occurring networks share the same network-level properties; for example, road networks do not show a scale-free property[439][1], although air travel networks do[49]. Comparing network centrality statistics to a random model (section 3.2.2), helps show if they are of significance or interest.

3.1.1 Small world property

The "small world" property means that a short number of hops from any node along the network edges can reach any other node. The maximum distance, the *diameter D*, between any two nodes in the network is therefore small[457]. Random networks are also small, but their nodes do not cluster together, unlike most real-world networks.

Milgram[276][2] famously demonstrated this phenomenon in a 1967 experiment; he tried to find the number of steps (the average path length) it would take for a letter to reach someone unknown to the first letter holder if they and everyone else in the chain could only mail it to people they knew personally. The average number of links was six, which was less than expected[3].

Watts and Strogatz [429] reported small-world properties in naturally occurring networks: the neural network of the nematode *Canenorhabditis elegans*; the power network of the western United States, and the collaboration graph of film actors gen-

[1] their degree distribution is similar to that found in a random graph

[2] although it was previously described by Pool and Kochen, although published later [384]see Newman [289] p 62

[3] Milgram reports it as the median number of intermediaries as five, which is six edges between sender and recipient using the definition of path length

erated from the Internet Movie Database (IMDb). These properties are a short average path length, with a high clustering coefficient in contrast to a random graph which has a short path length but low clustering coefficient. The average path length of small-world networks and random networks are proportional to the logarithm of the number of nodes in the network. Scale-free networks scale as loglogN[78].[4].

Watts and Strogatz[429] proposed a generative model for this phenomenon in which a circular lattice network has edges rewired randomly with probability p in which interconnections are rewired to span the circle[289][5]. A small amount of rewiring allows a short path length and preserves the high clustering coefficient. Figure 3.1 shows the generation of a small world graph from a circular lattice with increasing rewiring probability.

Only a small number of edges need to be rewired to produce the small-world property, and it has been argued that this is why it is so common in nature [289]. Small world networks in the Watts-Strogatz model (WS) are an interpolation between random networks and lattices, and some have suggested that it is the usefulness of the combination of clustering and connectivity[235] that explains why they are ubiquitous.

The small-world property allows rapid information transfer within networks. It has recently been posited as a model for widespread genetic effects on traits (the omnigenic model)[54], and as a developmental stage in the evolution of brain networks[413].

3.1.2 The configuration model

To assess the significance of a network or a vertex statistic, we can measure how much they differ from their expected value given a random network. Simple random networks such as Erdős-Rényi (ER) random graphs have already been introduced (see section 2.1.2).

Erdős-Rényi models generate networks with a vertex degree distribution that is approximated by a Poisson distribution (see section 3.2.2), and hence the mode of the probability mass is centred around the mean degree $< k >$ of the network. Many real-world networks have more complicated degree structures, often a heavy-tailed distribution such as a power-law distribution (section 3.2.3).

The configuration model generates a new random network that retains the degree sequence of the original network. Network properties can be compared between the existing network and a configuration model with an identical degree sequence.

[4] scales as loglogN in scale-free with core-periphery [49]

[5] p312

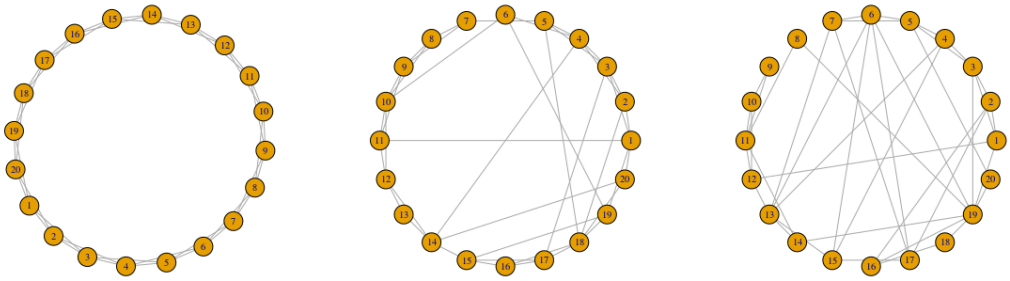


Figure 3.1: Development of small world property. A small world lattice with n=20 and probability of rewiring (left to right)= 0, 0.1, 0.3.

https://github.com/14760/thesis/blob/main/R/chapter2/smallworld/plot_small_world.R

The process for generating the configuration model is as follows: break each edge into two stumps and then rewire each of the stumps at random to another stump with uniform probability. Self-loops and multi edges can result, which do not occur in most real-world networks, but the number of these are low as the network size increases[289].

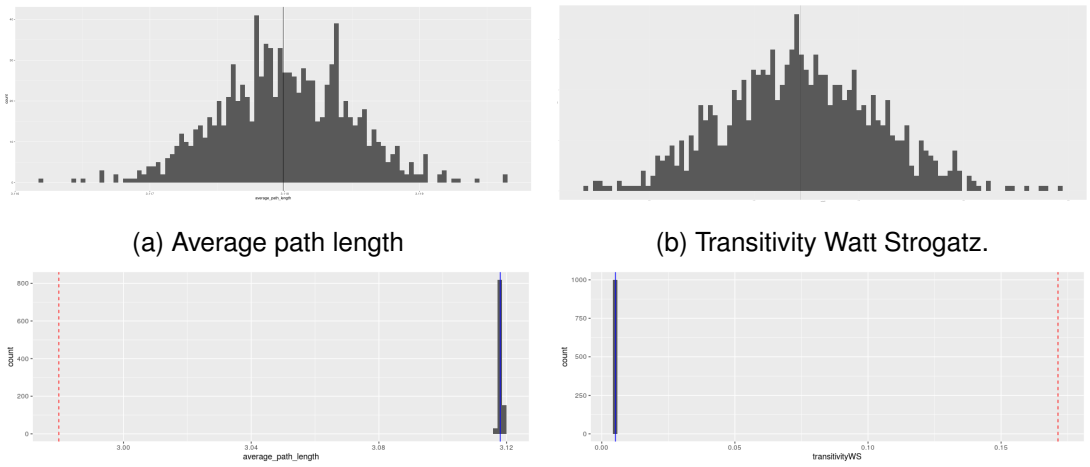
There are also specific models to generate random power-law distribution networks such as the Barabasi-Albert model[24] and earlier Price model[342], but if one is studying a *particular* network, and one has the degree sequence, as is the case here, the configuration model is a good choice[289].

3.2 Centrality measures

Centrality measures identify important, central nodes using network topology[289][6]. Node centrality also determines how information flows through a network [53]. Several measures of importance have been described and can be used to rank, group or identify key nodes. All measures are derived from the adjacency matrix.

Degree centrality is the simplest and is the number of edges that connect to a node. Other vertex measures include eigenvector centrality, a measure of the importance of

[6] p159



(a) Average path length

(b) Transitivity Watt Strogatz.

(c) Average path length showing mean

(d) Transitivity Watt Strogatz.

Figure 3.2: Figure shows average path length and transitivity (Watts Strogatz mean of local transitivity) for an Erdős-Rényi graph of 3,457 nodes and 30,498 edges (i.e. the same as the PSP graph). The difference in values is small but the sampling distributions do not overlap. Upper panels show the distribution of these values over 1000 iterations

`RProjects/chapter3/R/small_world/eda_small_world`

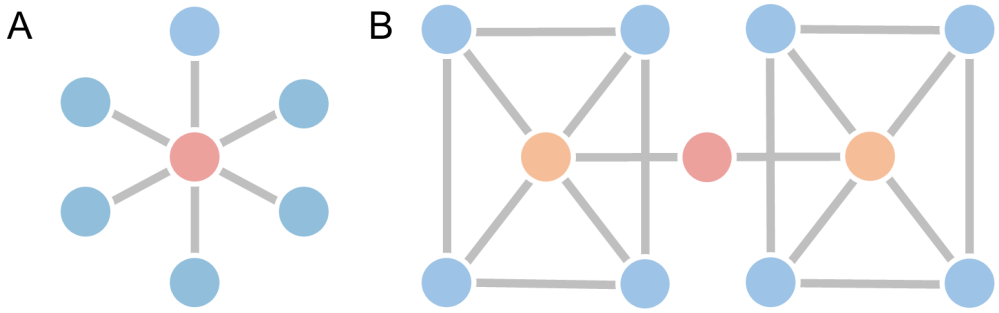


Figure 3.3: Centrality measures from Oldham et al. (2019)[310]. In panel A the red node has both high degree and closeness. In contrast in panel B the red node has high centrality but low degree. This image illustrates that although these measures can be correlated they can also represent different properties.

the node that includes the importance of its neighbours[51]. Some nodes are bottlenecks in the network. If a large number of the shortest (geodesic) paths pass through a node, it is a potential bottleneck, and its removal can disrupt the integrity of the network. The measure of a nodes propensity to be a bottleneck is the betweenness centrality[136]. Closeness centrality measures how close a vertex is on average to all other nodes in the network. Though not a classic centrality measure, transitivity can be seen as a local form of betweenness centrality [289] and is therefore included here along with kcoreness.

Centrality measures are frequently correlated with each other [409], although their correlation varies[310].

Figure 3.3 shows the fact that centrality measures can be concordant (for example, high degree and closeness in panel a) or discordant (red node has high centrality and betweenness in panel b but low degree).

Over 200 centrality measures have been described[196][7]. This chapter will consider degree centrality, eigenvector centrality, betweenness, closeness and transitivity in the PSP and their association with animal models of learning and human cognitive ability. The six measures primarily used in this thesis have been those that are implemented in igraph and are sufficiently popular to appear in a general textbook[289]. Degree centrality and eigenvector centrality are neighbourhood based centrality measures, and closeness and betweenness shortest path-related[207]. The simplest centrality measure, degree centrality, [289] is considered below.

3.2.1 Degree centrality

The degree, k, of a node is the number of edges connected to it. A node connected to a large number of other nodes can influence more of the network. In a protein-protein interaction network, changes in a high degree protein may give rise to a variety of different phenotypes or, if the protein is sufficiently essential, give rise to severe disease. In the PSP proteome network, the degree is the number of proteins a protein (represented by its encoding gene) directly interacts with.

All centrality measures can be derived from the adjacency matrix. The adjacency matrix $A_{i,j}$, represents an undirected, unweighted network such as the PSP, where $A_{i,j} = 1$ if there is an edge between nodes i and j and 0 otherwise, the degree of

[7] definitions of "classic" centrality measures vary [66] has closeness, betweenness, information and degree but not eigenvector centrality which is otherwise common. [55] has closeness, betweenness and degree joined later by eigenvector centrality

node i, k_i is[49]:

$$k_i = \sum_{j=1}^{n} A_{i,j}. \tag{3.1}$$

Nodes with a high degree, often called 'hubs'[457][8] may have more influence in a network. In a social network, individuals with a high degree centrality have many contacts and may be influential in the network. In a protein-protein interaction network, high degree proteins will have numerous interactions, take part in many biological processes and are likely to be necessary.

Disruption of high degree nodes also leads to disproportionate effects upon a network structure[197], [6]. Scale-free networks are relatively immune to disruption by removing random nodes but break apart with disruption of hubs. Hub effects on PPI networks are reviewed by Zotenko[458].

3.2.2 Degree distribution

The degree distribution describes the probability, $p(k)$, that a node has degree k. It is a probability distribution, and so

$$\sum_{k} p(k) = 1. \tag{3.2}$$

For an empirical network $P(k)$ is the proportion of nodes with degree k.

If a network were to have edges distributed between nodes with a uniform probability, p, we would expect the degree distribution ($P(k) = k$)) to be distributed according to the binomial distribution.

The probability of a node having k links is p^k, the probability of there being no links to the remaining $N-1-k$ nodes (-1 as we exclude self loops) is $(1-p)^{N-1-k}$, there are $\binom{N-1}{k}$ ways in which this arrangement of present and missing edges can be arranged, so the degree distribution follows the binomial distribution[23]:

$$\binom{N-1}{kp^k(1-p)^{N-1-k}} \tag{3.3}$$

In the limit, where the mean degree is much smaller than the size of the network, the degree distribution is closely approximated by a Poisson distribution[23]. The Poisson distribution

[8] In a directed network, hubs may refer to nodes that point to authorities nodes with important information. As the postsynaptic proteome network is undirected, hubs will be used in the first sense in this thesis[224]

$$p(k) = \frac{\lambda^k e^{-}\lambda}{!k}, \tag{3.4}$$

has only one parameter for an empirical network, λ is $E[k]$. The probability mass should be centred around the empirical mean (17.6 in the case of the PSP). The variance of a Poisson distribution is $\lambda^{1/2}$, which is approximately 4.12. The twenty postsynaptic proteome genes with highest degree (table 3.1) have a degree of 221 or greater. Under a Poisson model, these are around 100 standard deviations from the mean.

Entrez ID	k	Symbol	Description
351	535	APP	amyloid beta precursor protein
2335	474	FN1	fibronectin 1
1994	473	ELAVL1	ELAV like RNA binding protein 1
1956	450	EGFR	epidermal growth factor receptor
7514	396	XPO1	exportin 1
10482	375	NXF1	nuclear RNA export factor 1
7316	363	UBC	ubiquitin C
26270	342	FBXO6	F-box protein 6
7412	327	VCAM1	vascular cell adhesion molecule 1
55832	316	CAND1	cullin associated and neddylation dissociated 1
8454	316	CUL1	cullin 1
2885	313	GRB2	growth factor receptor bound protein 2
7534	298	YWHAZ	tyrosine 3-monooxygenase tryptophan 5-monooxygenase
3320	266	HSP90AA1	heat shock protein 90 alpha family class A member 1
10075	263	HUWE1	HECT, UBA and WWE domain containing 1, E3 UPL
4869	263	NPM1	nucleophosmin 1
7415	245	VCP	valosin containing protein
3178	227	HNRNPA1	heterogeneous nuclear ribonucleoprotein A1
22938	221	SNW1	SNW domain containing 1
8453	221	CUL2	cullin 2

Table 3.1: The twenty post synaptic proteome genes with highest degree in their cognate proteins. Degree = k. UPL = Ubiquitin protein ligase.

Code to generate table at `RProjects/gridsearch_gamma/R/degree_distribution`

3.2.3 Scale free

The work of Barabasi, Albert and Jeong (1999) led to the recognition that real-world networks often have a scale-free degree distribution [24],[25][9]. In a scale-free network, the probability of a vertex having degree k, $p(k)$, is

$$p(k) \sim k^{-\gamma}, \tag{3.5}$$

where γ is the degree exponent[23].

Power law distributions are 'scale free' because increasing the degree, k, changes the probability of the degree by a common multiplicative factor. This pattern of change leads to a straight line graph on a log-log plot of degree k and probability of degree $p(k)$, the signature of the scale-free distribution[37].

Many other things in nature follow a scale-free distribution (e.g. the size of craters on the moon, frequency of words and size of cities), and subsequent research showed that many other networks (the first being the wiring diagram of an IBM chip, the map of the power grid and the Hollywood actor database) are all scale free[23]. [10]

Albert, Jeong and Barabasi [5] built a web robot to map out the links in the `nd.edu` domain of the world wide web. They found 325,729 documents and 1,469,680 links. As documents point to other documents on the world wide web, they were represented by a directed graph. The network they constructed displayed a power-law distribution with $\gamma_{out} = 2.45$ and γ_{in}=2.1. At that time (1998), there was '*no trace in the literature of a network with a power-law distribution*'[23][11] although Price's network model, a generative model for power-law networks, was described in the 1960s[342].

The Barabási-Albert model for generating a scale-free network distribution requires network growth and preferential attachment. Preferential attachment means that as a network grows, new nodes are usually attached to high degree nodes[24].

The characterisation of power-law distributions remains an active area of research in statistics. Many networks display the power-law distribution only in the tails, and some seem to conform to log-normal distributions. Recent work has suggested that

[9] more recent work[58] has argued that scale-free networks are relatively uncommon and that a log-normal or other distribution may be a better statistical fit. When discussing the literature on scale-free networks, I have retained the use of scale-free or power-law from the original literature; however, this should be understood that this includes the possibility of another significantly right-skewed distribution

[10] p11 in [23]. Chapter 0 of this book gives a detailed personal history of the discovery of the scale-free property in networks and its historical antecedents

[11] p10

power-law distributions are comparatively rare compared to other right-skewed distributions[58]. In any case, many real-world networks are right-skewed with heavy tails and deviate markedly from a Poisson distribution.

3.2.4 Betweenness centrality

Betweenness centrality and closeness centrality (section 3.2.6) are both vertex centrality measures based on shortest paths between nodes[289]. Betweenness centrality, popularised by Freeman (1977) [136][12]is a measure of the number of geodesic (shortest) paths in the network passing through a particular node.

Following Newman's notation[289], if $n_{st}^i = 1$ if a shortest path between nodes s and t pass through node i, and g_{st} is the number of shortest paths between s and t (as there may be more than one shortest path), then the betweenness centrality of node i is

$$x_i = \sum_{st} \frac{n_{st}^i}{g_{st}}. \tag{3.6}$$

There is a shortest path in both directions between nodes in an undirected network, e.g. from s to t and t to s. This double-counting does not alter the rank of the nodes ordered by betweenness centrality and allows for easier application to directed networks[289]. The betweenness centrality can be normalised by dividing by the number of possible node pairs in the network n^2[289] [13].

Nodes with high betweenness centrality connect different areas on the network (i.e. paths between nodes in the network disproportionately flow through them) and are sites of increased flow (of information, for example). The description above assumes shortest paths are known and taken through the network[53] (random walk betweenness centrality measures betweenness when paths are random). Other nodes depend on nodes of high betweenness centrality to connect, and in the sociological literature, they are sometimes called 'brokers'[289].

3.2.4.1 Edge betweenness centrality

Edge betweenness centrality is the number of shortest paths that pass through a particular edge [146]. The removal of high edge betweenness centrality edges leads a network to break into communities (see section 4.3.1)

[12]Newman [289] points out that Freeman acknowledges an earlier unpublished account by Anthosine
[13]p176

3.2.5 Random walk betweenness

The betweenness centrality described above assumes that information takes the shortest path through the network although this may not always be true. Random walk betweenness is a betweenness metric dependent on the property of random walks in the network. The results are often similar and where they differ it is not clear if this is significant[289]. For this reason I have not included random walk betweenness.

3.2.6 Closeness centrality

Closeness centrality was described by Alex Bevelas in studies of information flows through social networks [33].

Closeness centrality measures how centrally located a node is in a network (i.e. closeness to all other nodes in the network). Closeness centrality is the inverse of the average length of the shortest path from a node to other vertices in the graph [137]. l_i is the mean distance between node i and every other node in the network:

$$l_i = \frac{1}{n}\sum_j d_{ij}. \tag{3.7}$$

As a *distance* measure it is small when nodes are close to all other nodes. The closeness *centrality* is its inverse:

$$C_i = \frac{1}{l_i}. \tag{3.8}$$

[289]

While betweenness centrality measures the dependence of other nodes in the network on a particular node (as a bottleneck for information), the closeness centrality has been interpreted both as a measure of proximity and efficient access to other nodes, and alternatively as a measure of independence from other nodes [55].

3.2.7 Eigenvector centrality

Degree centrality assumes that all the connections between a node and others in the network are equally important. However, adding a connection to a critical node may make a node more important than the increase in degree centrality by one would suggest.

Eigenvector centrality, proposed by Bonacich [50], takes into account the importance of a node's neighbours in calculating centrality and does not consider all of the

edges of a vertex to be equal. Instead, connecting to a critical node, results in a greater increase in importance.

Using Newman's notation [289], the importance of a node is defined as the sum of the importance of its neighbours if x_i is the importance of node i

$$x_i = \kappa^{-1} \sum_{j \in nei.i} x_j, \tag{3.9}$$

where κ is a constant of proportionality and j denotes all neighbours of node i. This can be written using the adjacency matrix as

$$x_i = \kappa^{-1} \sum_j A_{ij} x_j, \tag{3.10}$$

so avoiding the introduction of j as the neighbours of x_i, since the value of the adjacency matrix for nodes, i, j, that are not neighbours is 0.

In matrix form:

$$\mathbf{x} = \kappa^{-1} \mathbf{A}\mathbf{x}, \tag{3.11}$$

so,

$$\kappa \mathbf{x} = \mathbf{A}\mathbf{x}. \tag{3.12}$$

This is an eigenvector problem and the vector of centralities is one of the eigenvectors of $\mathbf{A}$. The eigenvector will have as its ith entry the importance of the ith node. As all centralities should be positive, the appropriate eigenvector is associated with the largest eigenvalue (by the Perron-Frobenius theorem).

3.2.7.1 Related measures

Katz centrality is a solution to the problem of calculating eigenvector centrality in directed networks. If a vertex is not in a strongly connected component, it will have an eigenvector centrality of 0. The Katz centrality includes a constant term β which is added to the eigenvector centrality, and a term α that determines the balance between the eigenvector centrality and the constant term. It is not widely used in undirected networks[289].

Page rank is a variant of Katz centrality widely used in web-search. Here the benefit of an important neighbour is divided by their out-degree (or degree in an undirected

network). If a node is linked to a very important neighbour, this is much less remarkable if the neighbour has a million other connections. A review is provided in Gleich [150]. There are also constant terms α and β analogous with Katz centrality[289].

3.2.8 Transitivity

Transitivity is a measure of how connected the neighbours of nodes are either globally or at the level of individual nodes.

Transitivity is the phenomena where two nodes are more likely to be connected if they share common neighbours. Transitivity is measured by clustering indices, forms a basis of the small world effect and can be thought of as a centrality measure. There is some inconsistency in the literature over the use of the terms clustering, clustering coefficient and transitivity, and it is worth clarifying this before going on.

A network's global transitivity measures the probability that two nodes, each connected to another node, will be connected (if someone has two friends, how likely are those people to be friends). The transitivity is the ratio of the number of these connections, *closed triads*, occurring in a network to the maximum possible (range 0-1).

Transitivity of relations, analogous to mathematical transitivity, has been used in the sociological literature since the 1940s (see Louch[257] for details). Heider writing in 1946 states: 'A relation R is transitive if aRb and bRc imply aRc'[172][14]. It is natural, therefore, to extend the description to networks describing social relations.

Watts and Strogatz described a *global clustering coefficient* in 1989[429], the arithmetic mean of local clustering coefficients. The local clustering coefficient measures transitivity (as described by Heider[172]), and the global measure is a global measure of transitivity. Watts and Strogatz do not use the term transitivity, but the global clustering coefficient is a mean of a measure of transitivity. Newman described a global measure of clustering (again a measure of transitivity) as an alternative formulation of the Watts and Strogatz global clustering Equation 3.13[291]. This was in the context of random graphs, and except in some circumstances (e.g. degree the same), these transitivities are not identical (see Schank and Wagner[369]).

There are, therefore *two* global measures of global transitivity, one described by Newman and one due to Watts-Strogatz (most often referred to as the clustering coefficient). These can and do diverge. There is a local clustering coefficient (described in Watts and Strogatz[429]) which is sometimes referred to as local transitivity. I will

[14] p109

Name	Global/Local	Measures	Symbol	Estrada[125]
Global clustering coefficient	Global	Global transitivity	C_{WS}	the Watts–Strogatz clustering coefficient
Global transitivity	Global	Global transitivity	C	global transitivity index
Local clustering coefficient	Local	Local transitivity	C_i	Watts–Strogatz clustering coefficientof a node i,

Table 3.2: Disambiguation of clustering and transitivity measures

refer to the transitivity due to Newman as C and the Watts-Strogatz as C_{WS}. Local transitivity is C_i

Attempts have been made to clarify this nomenclature[125], but these introduce more terms and definitions. A full exploration of this is beyond the scope of this thesis, but the interested reader will find all the relevant details Louch[257], Schank and Wagner[369] and Newman [289].

3.2.8.1 Global transitivity C (Newman)

A triangle of vertices, all connected, forming a loop is a *closed triad* and is a *path* of length two (as paths have distinct edges)[289]. The possible triads that could be formed by adding an edge are all of the paths of length two in the network, so the transitivity C is:

$$C = \frac{\text{number of closed paths of length two}}{\text{number of paths of length 2}}. \tag{3.13}$$

Since six paths are present in a closed triad (counting once for each direction the path can take in an undirected network, e.g. *ab* and *ba* are distinct), this is:

$$C = \frac{6 \times \text{number of triangles}}{\text{number of paths of length 2}} \tag{3.14}$$

.

Using *connected triple* to refer to any connected three vertices (whether closed or not), the transitivity can be written,

$$C = \frac{6 \times \text{number of triangles}}{2 \times \text{number of } \textit{connected triples}}. \tag{3.15}$$

.

The average connectivity of two nodes joined by a neighbour for a network or graph is thus its transitivity[290],[289]. This is the form found in [299] where it is said to be another form of the clustering coefficient of Watts and Strogatz (incorrectly, Schank and Wagner [369])

3.2.8.2 Global clustering coefficient C_{WS}

The local transitivity defined below (section 3.2.8.3), is also called the *local clustering coefficient*. The global clustering coefficient is defined by Watts and Strogatz[429] as the arithmetic mean of the vertex local clustering coefficients[125],

$$\bar{C} = \frac{1}{n}\sum_{i}^{n} C_i. \tag{3.16}$$

The global clustering coefficient can become dominated by nodes of low degree[289][15], and nodes with low degree tend to have high values of local clustering coefficient (section B.1.2). The transitivity above described by Newman[299] (see equation 3.15) is referred to as the *transitivity index*[125]. To avoid confusion I will follow Newman's [289] terminology and refer to $\hat{C}$ as C_{WS}.

3.2.8.3 Local clustering coefficient C_i

Local transitivity (also known as the *local clustering coefficient*, or specifically by Estrada[125] as the *Watts–Strogatz clustering coefficient* of a node *i*) is defined for a node with k_v neighbours as:

$$C = \frac{c_n}{c_{max}}, \tag{3.17}$$

where c_n is the number of connections between a nodes neighbours, and c_{max}, the maximum number of connections is,

$$c_{max} = \frac{k_v(k_v) - 1}{2}. \tag{3.18}$$

This [289] is precisely equivalent to the measure in Watts and Strogatz, which is the 'fraction of allowable edges that exist' amongst the neighbour of a node where node v has k_v neighbours is identical to 3.18.

The local clustering coefficient can be regarded as a local type of betweenness centrality and hence a type of centrality[289][16]. Local clustering measures the connectivity of nodes in the neighbourhood of a particular node. If the local clustering is low, then there exist disconnections between a nodes neighbours, which, in social networks, have been called structural holes [64]. The node connecting these neighbours has influence and importance or centrality similar to that measured by betweenness

[15] p188
[16] p187

centrality but confined to the immediate neighbourhood of the node[17]. I have therefore included analysis of local transitivity with that of other centrality measures.

It is clear from equation 3.17 and 3.18 that there is a problem for the local transitivity for nodes with degree 0 or 1. These *isolates* have been treated either as having an undefined value or as zero by convention (although see [369] where they are treated as 0 or 1). This has an effect both on local transitivity (Schank and Wagner [369]) and on the correlation of tranisitivity with other measures (see section B.1.2). See table 3.2

Schank and Wagner[369] discuss isolates in-depth; they hold that the transitivity is only defined for nodes with a degree greater than or equal to two. For those with degree at most one, it is either zero or one [18] and show that they get significantly different results using one rather than zero.

3.2.9 kcores

kcores can be viewed as a centrality measure and a solution to finding a cohesive set of actors in a graph efficiently.

The kcore of a graph is the maximum induced subgraph of a graph, such that each vertex has a degree of at least k. Kcores can be nested, and cores are not necessarily connected subgraphs[31].

H is a subset of g such that the degree of h is $>= k$

$$h \subseteq g \mid \delta'(h) >= k, \tag{3.19}$$

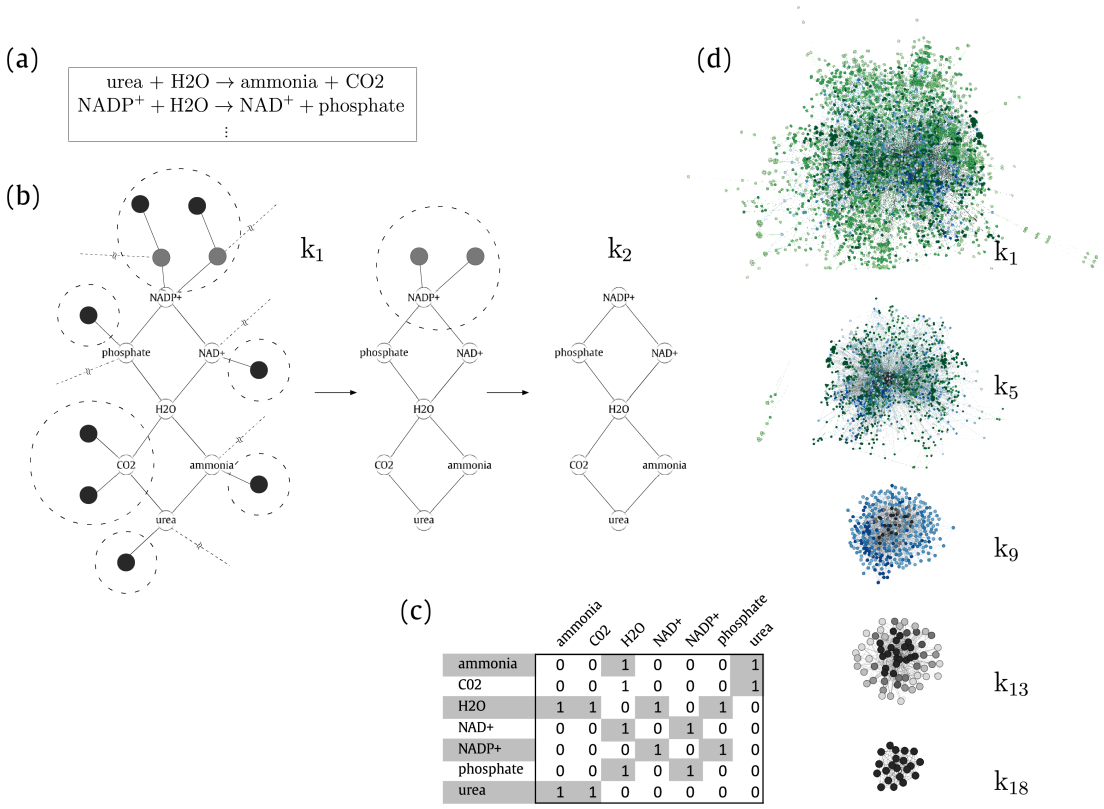
where g is the network, h the induced subgraph and $\delta'(g)$ the minimum of the degree of the vertices of h[374].

For a vertex its *kcoreness* is the maximal k core subgraph it is a member of (alternatively its shell index is c if it belongs to the c core but not the $c+1$ core)[374][19], [12]. The K-core is the connected graph left after components with degree less than k are removed[374][20].

The kcore algorithm has proved helpful in graph analysis as it runs in linear time and has been used as a measure of core-periphery structure in graphs[289].

K coreness was used to study the hierarchy of biological networks in plants. High core metabolic nodes represented core pathways conserved across different plant species[129]

[17] In social network analysis, they have been proposed to be a source of social capital and innovation
[18] see (text under equation 3 in [369]
[19] original description
[20] equation from Seidman p 272



	ammonia	CO2	H2O	NAD+	NADP+	phosphate	urea
ammonia	0	0	1	0	0	0	1
CO2	0	0	1	0	0	0	1
H2O	1	1	0	1	0	1	0
NAD+	0	0	1	0	1	0	0
NADP+	0	0	0	1	0	1	0
phosphate	0	0	1	0	1	0	0
urea	1	1	0	0	0	0	0

Figure 3.4: Image from Filho et al [129] showing k cores arising in the metabolic network of plants. Panel d shows the k core dissection of the network for k 1,6,9,13,18

(figure 3.4).

Wuchty and Almass carried out kcore dissection of the protein interactome of *Saccharomyces cerevisiae* to better understand the association between lethality and degree. The maximal kcore in their network was nine, and kcore was higher for essential proteins and in those with a eukaryote ortholog[436].

Newman reports that the sociological literature finds high k core actors to be important but that the evidence for this is limited[289]. It has been viewed as a kind of centrality measure core-periphery [289][21]. Given its reported utility in identifying vital elements in biological networks and as a potential supplement to degree centrality lethality[436] I have investigated kcoreness as a measure of centrality and report them along with other centrality measures.

[21] p180

3.3 Assortativity and assortativity of measures

Assortativity is a measure of the connectivity pattern of the network. If like node links to like the network is assortative, if dissimilar nodes link then the network is disassortative.

An example from a standard text is celebrities marrying one another more than would be expected by chance[23].

Assortativity was described and formalised by Newman as a network level measure of how likely nodes with similar properties are to be connected [291]. It can be calculated for scalar or categorical vertex properties. For example if supporters of the same sports team are more likely to be friends their social network would be assortative for this categorical vertex property (team supported). If in a social network people are more likely to be friends if their total wealth is different then the network would be disassortative for this scalar property. The degree assortativity is an important form of scalar assortativity with a significant impact on network structure and behaviour. We can use assortativity to see if significant nodes in GWA studies tend to be interconnected and if the same is true of animal models of cognition and LTP.

For a categorical vertex property, the measure of assortativity is the fraction of edges between vertices that share the property compared with the number of edges expected by chance.

This measure is the modularity which we will see again in chapter 4 where the groups are communities.

The modularity (Q) is

$$Q = \frac{1}{2m}\sum_{ij}(A_{ij} - \frac{k_i k_j}{2m})\delta_{g_i g_j}, \tag{3.20}$$

where m is the number of edges in the network, A is the adjacency matrix, k_i is the degree of node i, g_i is the group or category to which node i belongs and δ is the Kroneker delta.

The modularity lies in the range (-1,1) and takes on a negative value for a disassortative network structure.

The measure of assortativity for scalar quantities is the covariance of the scalar quantity between connected nodes. The mean used in calculating the covariance is the mean value of the scalar quantity at the end of each edge. This is:

$$\mathrm{cov}(x_i, x_j) = \frac{1}{2m}\sum_{i,j}(A_{i,j} - \frac{k_i, k_j}{2m})x_i x_j, \tag{3.21}$$

where x_i is the value of the scalar property at node i. The *assortativity coefficient*, r is the covariance normalised by the maximum 'value of the covariance in a perfectly mixed network'[289]

$$r = \frac{\sum_{ij}(A_{ij} - k_i k_j/2m)x_i x_j}{\sum_{ij}(A_{ij} - k_i \delta_{ij}/2m)x_i x_j}, \tag{3.22}$$

which is Pearson's correlation coefficient[289].

We can use assortativity to determine if vertices associated with a high z score in an intelligence GWAS (more associated with the trait) are more likely to be connected across the network. If there is widespread 'guilt by association'[311] then the network would be assortative for that particular property. Similarly the categorical measure of assortativity (equation 3.20) can be used to determine if genes that are involved in animal models of cognition are connected. This provides an objective measure of guilt by association and the 'disease module'[28] hypothesis (or trait module).

3.3.1 Degree assortativity

The most commonly studied scalar property is degree assortativity [289][22]. The degree assortativity is the Pearson correlation coefficient between node degree and the mean degree of its neighbours (for connected vertices)[307]. Assortative networks tend to form communities and are more robust to disruption by node removal [291] and increased assortativity leads to increased speed of transmission of information through a graph [307].

Social networks tend to be assortative with a core structure of highly connected hubs linking to other hubs[289] however biological networks tend to have a hub and spoke structure.

3.4 Centrality and essentialness

Hub proteins/genes tend to be essential to life. This was the finding of an analysis of the in the proteome of *Saccharomyces ceresvisiae* proteome) [197] shortly after the discovery that scale free networks are vulnerable to disruption by the removal of hubs[6].

[22] p210

The *Saccharomyces ceresvisiae* PPI network's (1,870 nodes, 2,240 edges) degree distribution followed a power-law distribution in its tail with 78% of its proteins in the largest connected component[197]. Elimination of the most connected nodes led to an increase in network diameter while random elimination did not, and mutations in high degree proteins were more likely to be lethal. Only 21% of the genes encoding proteins with degree less than five (93% of the proteome) had lethal effects from mutation, whereas although only 0.7% of yeasts had more than fifteen links, 62% of these had lethal phenotypes. The authors [197] concluded that the combination of centrality and lethality was due to the "*topological position of its protein product in the complex hierarchical web of molecular interactions*". The *centrality-lethality* hypothesis holds that high degree hubs are more likely to be essential for life[458].

He and Zhang (2006)[170] proposed an alternative mechanism: high degree node were more likely to be essential as they participated in more protein interactions, a proportion of which were essential and so had a higher probability of containing one of these essential links (rather than being integral to the network structure). Zotenko et al. [458] rejected these hypotheses noting that Gene Ontology terms suggested that hub proteins were more likely to take part in essential gene processes and that these clustered in modules they called ECOBIMs (Essential Complex Biological Modules - induced subgraphs around seed nodes representing gene ontology terms). Later proteomic work further complicated this picture: more complete networks constructed using high quality yeast interactomes did not show evidence of the centrality lethality effect[275],[450], [351].

Estrada[123] used additional centrality measures to assess essentialness in the yeast protein interaction network (degree centrality, eigenvector centrality, information centrality, closeness centrality, betweenness centrality and subgraph centrality). High centrality measures predicted essentialness better than random selection, and eigenvector centrality gave the best results and closeness and betweenness centrality the poorest.

Raman [348] examined multiple centrality metrics (degree, betweenness centrality, closeness centrality and pairwise dis-connectivity index) and their association with gene essentiality for 20 organisms. Using the STRING database, filtered for high-quality interactions, they found a disassortative pattern of connection of essential genes (section 3.3.1) and that essential genes had a higher degree and betweenness centrality. The twenty organisms they studied ranged from bacteria and yeasts to the more complex *Caenorhabditis elegans*.

3.5 Biomedical significance of centrality measures

Hub genes are reported to be more likely to be expressed ubiquitously[151] and critical genes in cancer proteomic networks are often of high degree[419].

High degree nodes were involved in more diseases in the Online Mendelian Inheritance in Man (OMIM) compendium of genetic phenotypes[441]. Goh et al., however, found that disease genes were likely to be tissue-specific and were not more likely to be hubs [151]. High degree hub genes were found to be more commonly essential than genes associated with disease[26] although all of these findings depend upon how disease-gene associations are defined.

Lee et al.[238] used the National Human Genome Research Institute (NHGRI) catalogue of complex trait GWAS[23] and found that complex trait loci were more likely to be found in genes that encoded proteins that were hubs or bottlenecks. Here, hub and bottleneck are defined as the top 20% of the degree centrality and betweenness centrality distributions in a network derived from the STRING protein-protein interaction database (14025 nodes and 492087 edges). The authors used their network findings to re-prioritise GWAS datasets composed of type 2 diabetes studies and Inflammatory Bowel Disease Genetics Consortium (IBDGC) and increased the number of significant genes found in independent sources such as OMIM[168]. The study found that SNP loci are commoner in genes that encode hubs or bottlenecks in the protein interaction network for at least these two polygenic disorders.

Hahn and Kern[166] examined protein-protein interactions networks in fly, yeast and worm, measuring betweenness centrality, degree centrality and closeness centrality. They found that more centrally located proteins in all networks evolved more slowly. Bayes (2011)[35] found more sequence conservation in PSP hub proteins, and Hill found that SNPs associated with differences in intelligence are found in genomic regions that are evolutionarily conserved[181]. These provide further reasons to investigate the association between centrality measures and genes associated with differences in intelligence.

Ghasemi et al. (2014) review centrality measures in biological networks and illustrate each centrality measure using Zachary's Karate network[143]. Some of the papers in this review are discussed below (see Ghasemi et al.[143] for full review).

Joy et al.[202] found that in the yeast protein-protein interaction network, essential nodes had 77% higher degree centrality and 82% higher betweenness centrality

[23] now the NHGRI-EBI GWAS catalogue

compared with non-essential nodes. They found several proteins with high betweenness and low degree centrality (High Betweenness Low Connectivity - HBLC), which standard models of scale-free network generation did not predict (e.g. Barabsi Albert). Degree and betweenness correlated with age and essentialness, but there were too few high betweenness low degree nodes to state if they differed from nodes with a similar degree. They make the interesting point that betweenness is less robust to mutation as it is not protected from mutation by the constraint of maintining multiple interactions partners.

Yu et al.[449] found that 'bottlenecks', nodes with high betweenness centrality, were more likely to be essential, particularly in directed and regulatory networks. They found that betweenness correlated with degree (Pearson correlation coefficient = 0.49). Non-hub bottlenecks, were also associated with essentiality, but degree was a better predictor of essentiality. Bottleneck nodes had lower than average co-expression supporting the hypothesis that they connect different modules.

Ozgur et al.[317] developed a curated gene interaction network and used centrality measures to attempt to predict the genes associated with prostatic cancer using the prostate gene database (PGDB). They measured degree, eigenvector, betweenness and closeness centrality and found that high scores on the first three measures predicted increased rank in the PDGB.

Milenkovic et al.[275] investigated whether biologically central genes are topologically central using dominating networks, an extension of the idea of a dominating set, to derive a centrality measure. The four categories of biologically essential were host-pathogen interactions, cancer, HIV and ageing. They also assessed betweenness centrality, subgraph centrality and graphlet centrality. They conclude that biologically essential genes are more enriched with all measures but most for the graphlet centrality, although the differences are modest[24].

Pandey et al.[321] analysed GWAS results for bipolar affective disorder using eigenvector centrality. They used the WTCCC-BD GWAS (1,868 cases and 2,938 controls) as a discovery sample and an NIMH GWA study (1,001 cases; n=1,033 controls) of individuals of European ancestry as replication. The network was a genetic association network and limited by the lack of power provided by GWA study sample size.

Barrenas et al. [29] found that fifty four complex diseases found at GWAS[183] were less central in the human interactome (using HPRD v7[213]) than monogenic

[24] other than ageing genes (see figure p8 C)

diseases or essential genes.

Authoritative reviews of the literature have concluded that disease genes are less likely to be essential[26] and more peripheral than essential genes, while others have held that the literature supports their being central in the protein interaction network[192].

3.6 Methods

The evidence of the effect of protein and gene centrality and biological function is heterogeneous but supports highly important nodes being essential. Their roles in specific human disorders are undetermined and may vary between disorders.

I will calculate the centrality for all proteins in the network and see if they correlate with the level of significance in GWAS of intelligence and educational attainment in the gene that encodes the cognate protein. To test if the centrality lethality effect holds in human populations, I will test the correlation with measures of genetic selection and determine the nature of these central nodes using Gene Ontology enrichment. I will determine if the PSP network has properties such as being scale free and will investigate how significant genes are distributed through the PSP, and whether this distribution is similar to or different to those of animal models of cognition and LTP.

3.6.1 Calculation of centrality measures

To test the hypothesis that genes encoding central proteins in the PSP network were more likely to be associated with genetic variants associated with differences in intelligence or educational attainment I have calculated the following centrality measures for each gene in the PSP: degree, eigenvector, closeness, transitivity, kcoreness and betweenness, and tested their correlation with each other and with each gene's significance scores (GWGAS p values in MAGMA) in the discovery and replication educational attainment and intelligence samples (section 2.2.1).

3.6.2 Calculating centralities

Centrality measures for the PSP network (section 2.1) were calculated using the igraph package, version 1.2.6 for R. The functions used are shown in table 3.3[25]. All nodes in the graph are included unless specifically reported otherwise. The centrality calculations are available as a csv file in the supplementary material.

[25] RProjects/chapter3_mac/R/utils/get_centrality_df.R)

Centrality metric	method	Options
Degree	`degree`	default
Eigenvector centrality	`eigen_centrality`	default (scale=TRUE)
Betweenness centrality	`betweenness`	directed=FALSE
Closeness centrality	`closeness`	default
TransitivityNA	`transitivity`	type='localundirected', isolates='NaN'
Transitivity0	`transitivity`	type='localundirected', isolates='zero'
K-core	`coreness`	default

Table 3.3: Calculating centrality metrics showing igraph methods for R. The local node level transitivity is synonymous with the Watts-Strogatz local clustering coefficient. The Watts-Strogatz clustering coefficient is transitivity0. The igraph default is to treat isolates as NA (transitivityNA)

The local transitivity was calculated using the transitivity method in `igraph` for R with the argument type being set to 'localundirected'. Where isolates - nodes with degree zero or one - exist there arises the question if we are to treat them as undefined (NaN) and exclude them from calculations or as zero (section 3.2.8.3; corresponding to arguments isolates='NA' and isolates='zero'). I will refer to transitivity with isolates='NA' as transitivityNA (C^{NaN}) and isolates = 'zero' as transitivity0 (C^{0}) and include both as outlined in section 3.2.8.3[369].

3.6.3 Correlation between centralities and measures of genetic significance

Centrality measures are often correlated[409]. Few have distributions that appear normally or even linearly distributed(figure 3.7). The correlation between centrality measures and the significance of genetic variants for intelligence and education were calculated using Spearman's rank correlation method (non-parametric) in R. This was also used to calculate the correlation between centrality measures and with other measures as is reported in other analyses of centrality in the literature [310]. The correlation was also measured with other numeric node metadata (such as *loss of function intolerance*). The -log10 transform of the cohort p values are reported in calculating the significance and degree of association between centrality and genetic significance in studies of intelligence and educational attainment. This means highly central nodes with highly

significant results in GWA studies will have a positive correlation.

3.6.4 Characterising the degree distribution

If a network follows a power law degree distribution this has implications for its growth, function and resilience (section 3.2.2). The characteristic feature of a power law degree distribution is a straight line in a log-log plot of the probability distribution $P(k)$ against degree k.

Calculating the gradient of this line results in a biased estimate of the exponent [289][26]. The method of Clauset, Shalizi and Newman is widely used instead[77].

The degree distribution often conforms to a power law distribution only in its tail (high degree)[289]. There is therefore a minimum degree (k_{min}) above which the power law distribution holds. This value (k_{min}) is found such that the distance between the network degree distribution and the best fit power law distribution is minimised. The distance between distributions is measured by the Kolmogorov-Smirnov statistic following Clauset, Shalizi and Newman [77][27]. The value of the power-law exponent γ varies as a function of k_{min} as shown in figure B.5a.

The R package `poweRlaw` (version 0.70.6)[145] was used to calculate x_{min} and the exponent of the distribution[28] and has been widely used in other network analyses (for example[433] and [273]). After k_{min} is calculated by minimising the Kolmogorov-Smirnov statistic, the exponent is obtained by numerically optimising the log-likelihood of the fit initialised using the analytic maximum likelihood estimate (MLE)[145].

Recent research has argued that the terms 'scale free distribution' and 'power law distribution' are used inconstantly in the network literature and that power law distributions may be rarer in empirical networks than previously thought. For many networks a log-normal or other heavy tailed distribution may be a better statistical fit[58]. This remains controversial[186], but what is clear is that even the log-normal is a significant deviation from the Poisson distribution found in an Erdös-Rényi random graph generated with uniform probability of an edge between nodes.

The power law, log-normal, exponential and Poisson distribution were fitted to the empirical distribution following the instructions in the poweRlaw[145] vignette.

A plot comparing potential models to the complementary cumulative distribution

[26] p324

[27] section 3.3 p 11-12

[28] p.7 `https://cran.r-project.org/web/packages/poweRlaw/vignettes/a_introduction.pdf`

function (cCDF) of the empirical degree distribution of the PSP network was made to identify clearly implausible models. A bootstrap estimate p of the plausibility of a power law distribution was made using the poweRlaw package[29]. The parameters and log-likelihood of each model were calculated using the x_{min} of the power law distribution[30]. A likelihood ratio test was carried out (Vuong's closeness test) to compare models[418]. The methods in the poweRlaw package[145] implementing the methods of Clauset, Shalizi and Newman [77] are also those that were used to test for the presence of power law distributions in the recent paper questioning their frequency by Broido and Clauset [58]. The level of confidence in the existence of the power law distribution is reported using the scale of Broido and Clauset (not scale free, super weak, weakest, weak, strong and strongest)[58].

3.6.5 Calculating small world property

The small world property is present if a network has a short average path length and high clustering coefficient compared to a random model. The average path length and global clustering coefficient of the PSP network were calculated using igraph. Both global clustering coefficients were calculated (section 3.2.8); the proportion of connected triples described by Newman[299] and the mean of the local transitivity due to Watts and Strogatz (C and C_{WS}) (section 3.2.8)[429]. Given the ubiquity of the small world phenomena in complex networks[289], a measure of small worldness was calculated, the small world metric of Humphreys and Gurney[190].

The global and local transitivity are calculated using the `transitivity` function in igraph for R. The global transitivity measure described by Newman (C)[299] is calculated by specifying type as 'globalundirected'. The Watts-Strogatz clustering coefficient (global transitivity - C_{WS})[429] using the argument type="globalaverageundirected". In calculating the local transitivity or Watts-Strogatz global clustering one has to decide how to deal with isolates, nodes with degree zero or one as this can have a significant effect on the calculation[371]. The type="localaverageundirected" global transitivity in igraph has as its default the mean of local transitivity with isolates removed (isolates='NaN'). The implementation of local clustering in graph tool [31], by contrast, sets isolates to zero[326], and global clustering [32] returns the transitivity defined by New-

[29] p 17 rule out powerlaw on bootstrap p if $p <= 0.1$, p large when good fit

[30] therefore if $x_{min} = y$ for a power law distribution other distributions are fitted with $x_{min} = y$ and the log-likelihood reported for the range x_{min} to maximum degree

[31] `graph_tool.clustering.local_clustering`

[32] `graph_tool.clustering.global_clustering`

man[299]. The Watts-Strogatz clustering coefficient (C_{WS}) is, by convention, the mean of local transitivity with isolates set to zero[289]. Treating isolates as zero for this and local transitivity seems to be the majority practice [426],[289] and is better suited for analytic solutions. I will follow it here, although there are reasons to prefer treating these nodes as undefined ('NaN'), one is that it complicates the relation of the transitivity to other measures(section 3.2.8.2). How isolates are treated only affects the Watts-Strogatz *global* transitivity which I will refer to as C_{WS_0} if isolates are zero and $C_{WS_{NaN}}$ if isolates are excluded from the calculation of the mean. For local transitivity I will append 0 or NaN to the transitivity to show how isolates have been handled (e.g. local transitivity0) and as symbol C_i^0 or C_i^{NaN}.

The PSP network was compared to two different Monte Carlo estimates of expected average path length and global clustering coefficient. These random graph models had the same number of nodes and edges as the PSP. The first, the Erdös-Rényi random graph, is a graph g with parameters n and p; n the number of nodes and p the probability of two nodes being connected by an edge. A network with the small world property has a high clustering coefficient compared to an Erdös-Rényi model and an average path length closer to a random graph than a lattice (section 3.1.1).

The second is the degree corrected random model (configuration model), in which the degree sequence of the network is maintained by breaking all edges into two stubs which are then randomly rewired. The rewiring model has the potential for self loops and multiedges but the probability is small as the number of nodes increase. Eliminating the possibility of self edges is undesirable for technical reasons[289]. The configuration model is implemented in igraph for R in the function `degree.sequence.game` using default arguments[33]. The Erdös-Rényi random graph using `erdos.renyi.game` . One thousand random graphs were generated, and the mean and standard deviation of the average path length and clustering coefficient calculated[34]. As the two forms of global transitivity (C_{WS} and C) often diverge I have reported both values[125],[190].

The small world metric S, was calculated where S, is

$$S = \frac{C/C_r}{L/L_r}, \tag{3.23}$$

if C is the transitivity and L is the average path length (see [288] for other metrics).

[33] Three possible methods are available: simple, simple avoiding multiedges and vi (Viger and Latapy), simple is the default

[34] `RProjects/chapter3/R/small_world/eda_small_world.R` using `erdos.renyi.game`, type gnm, 1000 iterations.

3.6.6 Calculating assortativity

To determine whether genes in the PSP network associated with intelligence and educational attainment tend to disproportionately connect with one another on a network level[35], I have calculated the assortativity between nodes for the z score levels for gene significance in the GWA studies (section 3.3). Calculations of assortativity were carried out using igraph. The degree assortativity was calculated and compared to other complex networks. I have also calculated the assortativity by the other centrality measures. I have calculated whether the network is assortative by evolutionary pressure using essential gene status (nominal assortativity) and probability of loss of function intolerance (scalar)[36](See section3.3).

The degree assortativity of the PSP was compared with the mean and standard deviation of the degree assortativity of 1000 Erdos-Renyi random graphs with parameters matching the PSP[37] and the mean and standard deviation of degree assortativity of 1000 degree sequence (configuration) models[38].

Nominal assortativity for categorical variables was calculated by encoding the property (e.g. membership of gene set) as a numeric vector and using the `assortativity_nominal` function in igraph. The assortativity for murine models of LTP and related gene sets was calculated using the gene sets in table 3.13[39]. The nominal assortativity (Q) is the modularity of the division of the PSP into two groups using this property (section 3.3).

To determine if there was 'guilt by association' amongst genes associated with intelligence and educational attainment I have calculated the network assortativity for gene z-scores in the population intelligence and educational attainment samples. This allows us to see if significant genes are more likely to be connected to one another without specifying a cut-off level of significance. In calculating the assortativity with z score, PSP genes that do not have associated MAGMA genes scores had missing values replaced by the median of the Z score of the particular sample[40].

3.6.7 Gene Ontology Analysis of extreme centrality genes

Section 3.4reviews evidence that that central nodes may be more essential because they are qualitatively different and more involved in core processes.

[35] as opposed to areas where they interact as modules

[36] `RProjects/chapter3/R/assortativity_distribution/z_score/assortativity_z.R` degree and kcore

[37] `RProjects/chapter3/R/assortativity_distribution/assortativity_er.R`

[38] `RProjects/chapter3/R/assortativity_distribution/assortativity_config.R`

[39] `RProjects/chapter3/R/assortativity_distribution/nominal_murine/murine_LTP_assortativity.R`

[40] the assortativity function does not handle missing values

To determine whether central nodes possess distinctive non topological properties I performed gene ontology enrichment analysis using ToppGene and as a background set all PSP genes (see section 2.1.3). If there is no association between centrality and genes being associated with intelligence and educational attainment genes this would be unremarkable if genes with high or low centrality were not in some way distinct (one might expect all phenotypes to show no association). In addition this may support or refute the variant of the centrality lethality hypothesis of Zotenko and colleagues [458] as it applies to humans (section 3.4).

In order to define a set of extremely central genes (which we need for GO enrichment analysis) we need to decide on a level beyond which values are considered extreme. The use of the top 10 or 20% is common in the literature. Lee et al.[238] used the top 20% of degree and betweenness centrality (section 3.5). Hsing et al.[187] report the use of gene ontology terms to predict if proteins are hubs using boosting trees. To create their training set they used the 90th centile of the degree distribution as it represented an inflection point in the cumulative distribution.

Given contradictory findings in the literature on the nature of central nodes in model organisms such as budding yeast, the nature and function of high degree or high centrality nodes in the synaptome is, to my knowledge, an open question. In addition low degree or centrality nodes have been neglected.

To calculate gene ontology enrichment I have used the ToppGene platform[73] using Entrez ID to identify nodes. Enrichment was carried out using the PSP as the background set of genes, as the central genes can only come from the PSP. This is in contrast to testing GO enrichment for significant intelligence and educational attainment GWA genes (where they were not constrained to be in the PSP - see section 3.6.7. The ToppGene report contains the results for all Gene Ontology clades and gene-ontology mappings resulting in easier and more accurate data storage and parsing.

The methods for ToppGene have been described in section 2.1.4.2. The API does not support the use of a custom background set (the PSP) and so the web interface was used[41].

The top and bottom 10% of the centrality distributions were used for ontology enrichment. In some cases this results in more than 10% of the genes in the PSP (for example where over 10% have degree $k = 1$). The choice of 10 or 20% to represent extreme centralities is common in the literature[310][187]. Gene ontology enrichment was carried out using the genes with the highest and lowest 10% of values for degree,

[41] https://toppgene.cchmc.org/API/

betweenness centrality, eigenvector centrality, closeness centrality, transitivity and k-coreness. As several of the centralities have a right skewed distribution, low centrality measures may be less useful to discriminate between nodes and it may be most useful in kcore, transitivity and closeness.

Local transitivity was calculated using NaN for isolates[42]. Centrality measures were calculated using igraph for R (section 3.6.2). The text output from ToppGene is available in the supplementary material.

3.6.8 Disease term enrichment - DisGenNet

To test whether central nodes are associated with specific disorders or groups of disorders I used DisGeNET v7.0 accessed using the disgenet2R package for R[335]. DisGeNet is a comprehensive database of disease gene associations integrating data from a variety of sources[334]. The R package remains under active development and changed in the course of writing the thesis. I have therefore used ToppGene to query the DisGeNet database with the PSP as background set as this platform is relatively stable.

Enrichment was carried out in the same manner as for gene ontology enrichment, using the top and bottom 10% of genes in each centrality measure. The disease enrichment of all PSP genes is shown in table 3.4.

Specific diseases were identified using Universal Medical Language System (UMLS) Concept Unique Identifiers (CUI)[68]. Central nervous system and neuropsychiatric disorders were identified using the disease codes in the supplemental material to generate enrichment results for all diseases and neuropsychiatric/central nervous system only.

3.6.9 Methods Murine LTP and centrality

Long term potentiation (LTP) has been associated with learning in animals, and so following the genes to cognition approach determined whether murine genes associated with LTP had high or low centrality.

Murine genes associated with LTP were identified using Mammalian Phenotype (MP)[43] identifiers[62] and downloaded from ToppGene[73]. Mean centrality measures

[42] as otherwise there is no way to distinguish the very different topologies implied by the transitivity of a star graph or an isolated node

[43] `https://www.ebi.ac.uk/ols/ontologies/mp`

	ID	Name	p	q FDR BH	PSP	Genome
1	C0524851	Neurodegenerative Disorders	1.25×10^{-54}	2.47×10^{-50}	484	1480
2	C3714756	Intellectual Disability	2.98×10^{-50}	2.94×10^{-46}	412	1219
3	C0014544	Epilepsy	3.03×10^{-46}	1.99×10^{-42}	380	1124
4	C0002736	Amyotrophic Lateral Sclerosis	9.16×10^{-41}	4.52×10^{-37}	354	1069
5	C0030567	Parkinson Disease	2.80×10^{-39}	1.10×10^{-35}	548	1952
6	C0338656	Impaired cognition	1.39×10^{-37}	4.59×10^{-34}	415	1368
7	C0036572	Seizures	2.82×10^{-36}	7.96×10^{-33}	339	1052
8	C0557874	Global developmental delay	2.37×10^{-33}	5.84×10^{-30}	224	608
9	C0036341	Schizophrenia	5.08×10^{-33}	1.11×10^{-29}	291	883
10	C0004352	Autistic Disorder	2.03×10^{-32}	4.01×10^{-29}	285	864

Table 3.4: Ten most enriched disease terms in the PSP using ToppGene PSP

`RProjects/disgen2/R/toppgene_all_PSP/print_table_toppgene.R`

were calculated for murine LTP genes.

Potential Murine cognition gene sets were identified by manually searching the top 50 enriched mouse phenotypes for terms related to learning, potentiation or cognition. Ten particularly relevant gene sets were identified and are listed in table 3.5[44].

3.6.10 Centrality and Essential Genes

To test if central genes in the PSP support the centrality-lethality hypothesis (section 3.4) in human populations, and to investigate whether genes associated with differences in intelligence were under purifying selection (as described in [181] section 3.4), we used the database of essential genes (DEG)[261] and the Exac and the Gnomad databases to provide evidence of whether human genes were under selection presure.

3.6.10.1 Methods: Database of Essential Genes

The Database of Essential Genes (DEG) hosted at Tianjin University Bioinformatics (Tubic) centre, records genes found in experiments to be essential for life in a variety of organisms along with details of those experiments. The database was first reported in 2004 and the most recent documented version for which a version number is available is 15.2[45][260].

[44] `https://hpo.jax.org/app/download/annotation`

[45] `http://tubic.org/deg_bak/`

Rank	ID	Name	*n* PSP	*n* total
8	MP:0002063	abnormal learning/memory/conditioning	304	742
10	MP:0014114	abnormal cognition	304	744
11	MP:0002207	abnormal long term potentiation	142	258
20	MP:0001473	reduced long term potentiation	100	172
25	MP:0001898	abnormal long term depression	66	102
29	MP:0002062	abnormal associative learning	150	339
30	MP:0001469	abnormal contextual conditioning	95	181
31	MP:0001468	abnormal temporal memory	107	215
44	MP:0001463	abnormal spatial learning	121	272
46	MP:0004859	abnormal synaptic plasticity	50	77

Table 3.5: Mouse phenotypes associated with learning in the top 50 mouse phenotypes found on enrichment analysis of all PSP genes in order to identify potential animal cognitive gene sets for network analysis. n PSP is the number of genes in the PSP with an annotation and n total the number of annotated genes in the genome.

An update became available from 1st September 2020 with an increase in the number of organisms in the database (9 to 12), the number of genes reported (34,590 to 43,294) and the number of experiments and publications cited (16 to 20). This is the version used in all analyses presented here.

In the course of writing this thesis the publication reporting the update became available[261] but at the time of writing the links on the website no longer work. I have therefore stored the data in the supplementary electronic data store (SEDS). The database was downloaded on 18th October 2020.

Twelve organisms are found in the database, although the majority of genes are human (32,578 of total 43,294 - see table 3.6). A gene appears once in the record each time it appears in one of the twenty studies making up the database; the number of genes appearing in multiple studies is shown in figure 3.5. 9,085 unique human genes are reported using gene symbols using the gprofiler conversion tool 9,001 unambiguous unique human genes can be assigned Entrez Ids. The number of genes is larger than the number typically regarded as being essential.

Recently Zhang et al.[454] used a combination of sequence and network data, with deep learning and graph embedding features, to predict essential human genes using the DEG. They found 8,256 human genes (identical to DEG 15.2) across 16 studies and used those reported in five or more experiments to accord with the general approximate 10% estimate of the number of human genes that are essential[425].

I calculated the number of essential genes found in the PSP compared to the rest of the genome (Genes with loci in NCBI 37.3) using Fisher's exact test.

I compared the distribution of six centrality measures used in this section between essential genes and non essential genes using both parametric and non parametric tests of significance. I have carried out these calculations for all genes appearing in the essential list[46] and the 2,181 genes found in five or more[47] studies[48]. I have tested the correlation between centrality scores and the number of experiments a gene has been shown to be essential in.

[46] RProjects/db_essential_genes/R/get_essential_genes/db_2020/findforfive/get_db_essential_greater_five_results_centrality.R
[47] RProjects/db_essential_genes/R/get_essential_genes/db_2020/findforfive/get_db_essential_greater_five_results_centrality.R
[48] RProjects/db_essential_genes/R/get_essential_genes/db_2020/get_db_essential_greater_five.R

	Species	Number of Entries
1	Homo sapiens	32578
2	Plasmodium falciparum	2680
3	Mus musculus	2524
4	Schizosaccharomyces pombe 972h-	1260
5	Saccharomyces cerevisiae	1110
6	Bombyx mori	1006
7	Arabidopsis thaliana	356
8	Drosophila melanogaster	339
9	Caenorhabditis elegans	338
10	Danio rerio	315
11	Komagataella phaffii GS115 chromosome 1	214
12	Komagataella phaffii GS115 chromosome 2	201
13	Komagataella phaffii GS115 chromosome 3	192
14	Komagataella phaffii GS115 chromosome 4	146
15	Aspergillus fumigatus	35

Table 3.6: Organisms represented in Database of Essential genes 2020 updates. Number of entries is not equal to number of unique genes. All four of the chromosomes of the yeast K.phaffii (Pichia Pastoris)[41] have separate species entries perhaps because of its centrality in pharmaceutical processes. There are therefore a total of twelve organisms in the DEG.

RProjects/db_essential_genes/R/get_essential_genes/db_2020/organism_table.R

3.6.11 Loss of function and evolutionary pressure ExAC and gnomad

The Exome Aggregation Consortiuim (ExAC) database collected the data from the exomes of over 60,000 individuals to provide a measure of the tolerance of mutation in genes and genetic loci[241]. Comparing the number of loss of function mutations to the the number of harmless synonymous mutations provides a measure of how intolerant to mutation a gene or genetic region is in a population. The consortium provide a statistic, the probability of loss of function intolerance (pLI), that measures how much purifying selection a gene is under[49], taking into account confounding factors such as

[49]The data was downloaded from "https://storage.googleapis.com/gnomad-public/legacy/exac_browser/forweb_cleaned_exac_r03_march16_z_data_pLI_CNV-final.txt.gz" (valid on 16.06)

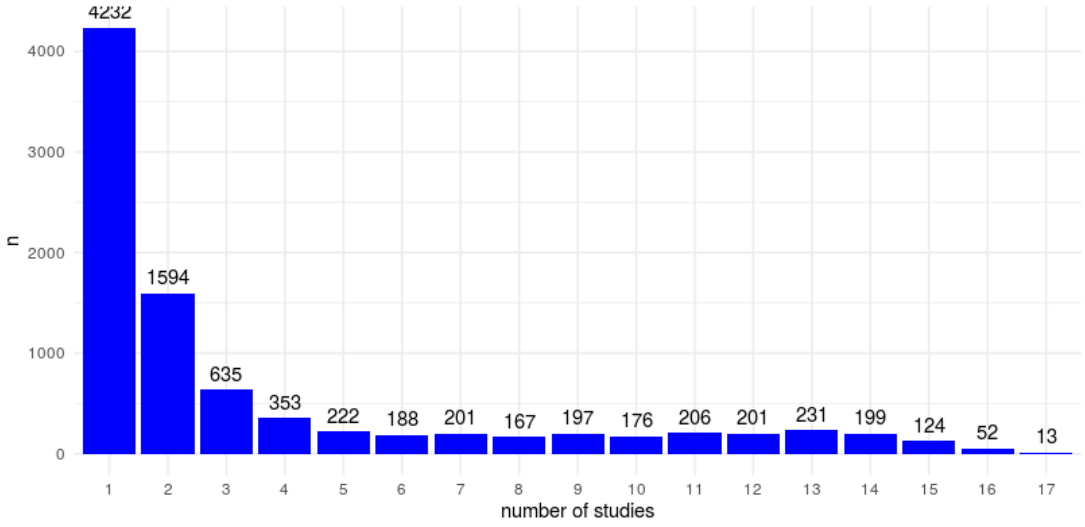


Figure 3.5: Number of datasets genes in Database of Essential Genes appear in. Twenty nine genes appearing in forty two studies for which consistent entrez ID could not be established were removed.

gene size.

To test the hypothesis that central genes are more likely to be essential and, by extension, intolerant of mutation, I measured the correlation coefficient between pLI and centrality measures such as degree. Likewise, to determine whether genes associated with intelligence and educational attainment were more likely to be loss of function intolerant, I calculated the correlation between gene z score and the -log10 transform of the gene p value.

Very recently a greatly enlarged successor the Genome Aggregation Database (gnomAD) has become available and I have repeated the analysis using this which I report here. It aggregates 15,708 whole genomes and 125,748 exomes across 141,456 subjects. This increases the number of small genetic variations reported to 241 million compared with 7.4 million by ExAC and thus a more complete picture of genetic constraint can be obtained[206]. Educational attainment (and schizophrenia) have been reported to be the most enriched traits with constraint[206][50] but the level of constraint in the PSP and the association between constraint and network features are unreported.

Gnomad v2.1.1 loss of function metrics by gene were downloaded from the consortium's cloud storage url[51]. This is the 6th of March 2019 release and contains loss

[50] specifically decile of LOEUF '*90% upper bound of this confidence interval, which we term the loss-of-function observed/expected upper bound fraction (LOEUF) (Extended Data Fig. 7b, c),*[206]'

[51] `https://storage.googleapis.com/gnomad-public/release/2.1.1/constraint/gnomad.v2.1.1.lof_metrics.by_gene.txt.bgz`

of function summary metrics by gene[52].

The file contained 19,704 rows and 77 columns. Three gene identifiers are provided: Ensembl Gene ID, Ensembl transcript ID and HGNC Symbol with 19,658 unique Gene Symbols and 19,704 unique Ensembl IDs.

Translating Ensembl ID to Entrez ID does not always result in a one to one mapping. I tried Ensembl ID and Gene Symbol to determine which performed best and used both gprofiler2[53] R interface (v0.2) to the g:Profiler toolset[352] and the ToppGene API. The best result was obtained using Ensembl ID and gprofiler2 resulting in 18,782 unique entries, 46 ambiguous/duplicated which were removed, and 983 which could not be mapped to an Entrez Gene ID[54].

The ToppGene translation API using a custom python script[55] resulted in 18,604 unambigous, unique Entrez ID terms. Both translations are available in csv format in the supplementary material.

1,185 Ensembl ID could not be translated to Entrez ID and were excluded. Forty eight duplicate Entrez ID were generated in the translation process and these too were excluded. The data cleaning is documented in an R markdown file[56].

3.6.12 Cohorts and samples

The population data used to assign significance values to the association of genes and intelligence and educational attainment and to test the association of theses genes with centrality measures are described in sections 2.2.1 and 2.2.2.

Gene level statistics were calculated using MAGMA as outlined in section 2.3.

[52] gnomad v3 is available but does not have these summary data

[53] `https://CRAN.R-project.org/package=gprofiler2`

[54] `RProjects/db_essential_genes/R/gnomad/download_gnomad.R`

[55] `RProjects/db_essential_genes/python/toppgene_api.ipynb`

[56] `RProjects/db_essential_genes/R/gnomad/mkd_download_gnomad.Rmd`. 18,717 genes have an Entrez ID associated with a Gnomad record. Of these 338 have NaN pLI values and were excluded resulting in 18,379 records. 3,407 PSP gnomad records are found (98.6% of PSP). Translation was stored as RDS and plain text files.

3.7 Results Summary of centrality measures

Degree, eigenvector centrality and betweenness centrality have heavy right tails with the mean greater than the median. The mean degree for the PSP network, was 17.6 and its median 8. A histogram showing the distribution of centrality measures is shown in figure 3.7. None of the distributions resemble a Gaussian distribution other than the closeness centrality. The K core index has a more uniform distribution (figure 3.6) with more nodes in the lowest kcores [114]. The U shaped kcore distribution of some complex networks is note seen[63] . The transitivity has a marked peak at zero as discussed previously.

Summary statistics for the centrality measures are shown in table 3.7. The local transitivity is reported first with isolates (nodes with degree $k = 1$[57]) set to zero (C_i^0 - transitivity0) and second (C_i^{NaN} - TransitivityNA) where nodes of degree one ($k = 1$) have an undefined transitivity and are removed from caclulations of summary statistics or correlation[58]. There were 304 nodes of degree one ($k = 1$).

The eigenvector centrality has been scaled to have a maximum value of one and local transitivity (local clustering coefficient) lies in the range 0-1. The betweenness is shown both as the number of shortest paths going through a node and scaled in the range 0-1. Closeness centrality does not follow a power law distribution [289][59] as it has a limited distribution as the reciprocal of the shortest path length from a node to all other nodes.

3.7.1 Testing the normality of centrality measures

Inspection of the histogram of the distribution of centrality measures in figure 3.7 suggest only the closeness centrality is approximately normal and degree, betweennness and eigenvector centrality are heavy tailed.

To determine the most appropriate method to test the correlation of centrality measures the Shapiro-Wilk test was performed using R and is shown along with the results of the Kolmogorov-Smirnov test, calculating the statistic as the distance between the empirical distribution and the normal distribution with identical mean and standard deviation is shown in table 3.8.

These confirm that the centrality measures are not normally distributed, and a non

[57] there are no nodes with $k = 0$

[58] because for a node with degree=1 the denominator $\frac{k_v(k_v)-1}{2}$ in equation 3.17 would be 0)

[59] p331-332

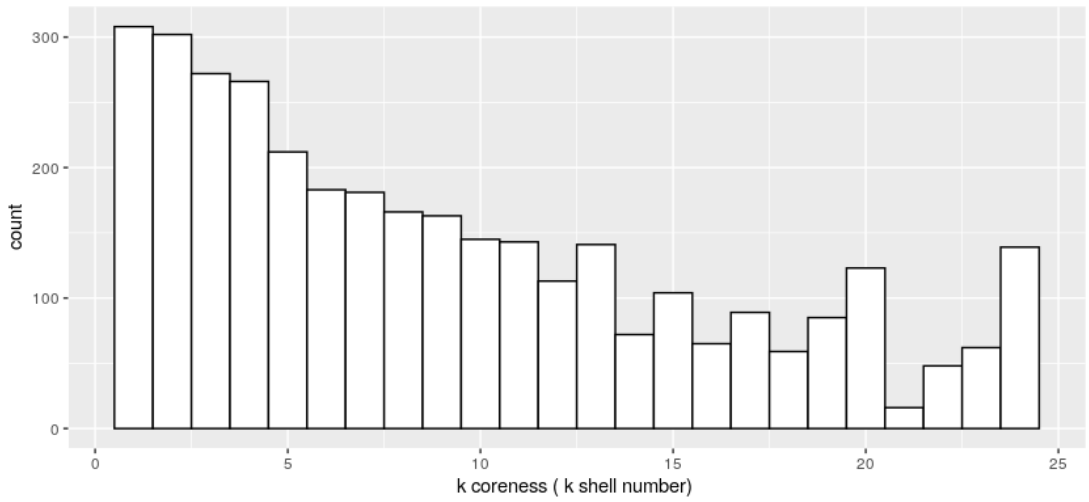


Figure 3.6: Histogram of shell index of vertices in calculation of kcore. Shell index is the maximum kcore a vertex is a member of and not a member of shell index n+1.

parametric test of correlation (Spearman's rho) was used [60] in accordance with some of the recent literature [310].

	k	Bet	$\text{Bet}_{0\text{-}1}$	Eigen	Closeness	C_i^0	C_i^{NaN}	kcore
Min.	1.0	0.0	0.0	1.82×10^{-6}	5.70×10^{-5}	0.00	0.00	1.0
1Q	4.0	43.3	7.25×10^{-6}	8.83×10^{-3}	9.22×10^{-5}	0.00	0.04	3.0
Median	8.0	317.0	5.31×10^{-5}	2.19×10^{-2}	9.88×10^{-5}	0.11	0.13	8.0
Mean	17.6	3421.1	5.73×10^{-4}	4.79×10^{-2}	9.85×10^{-5}	0.16	0.17	9.2
3Q.	19.0	1571.6	2.63×10^{-4}	5.32×10^{-2}	1.06×10^{-4}	0.22	0.24	13.0
Max.	535.0	644670.7	1.08×10^{-1}	1.0	1.40×10^{-4}	1.00	1.00	24.0

Table 3.7: Summary statistics for centrality measures in the PSP. 1Q 3Q = first and third quartiles.Bet=betweenness, $\text{Bet}_{0\text{-}1}$ is betweenness scaled by maximum possible betweenness[61]. C_i^{NaN}= TransitivityNA, isolates set to NA. Number of NA: 304. The value of transitivityNA (C_i^{NaN}) is slightly greater than transitivity with isolates set to zero (Transitivity0 C_i^0)

https://github.com/14760/thesis/blob/main/R/chapter3/centrality/summary_centrality_range_bet.R

[60] RProjects/paper_xls_latex/R/centrality_summary/centrality_summary_latex

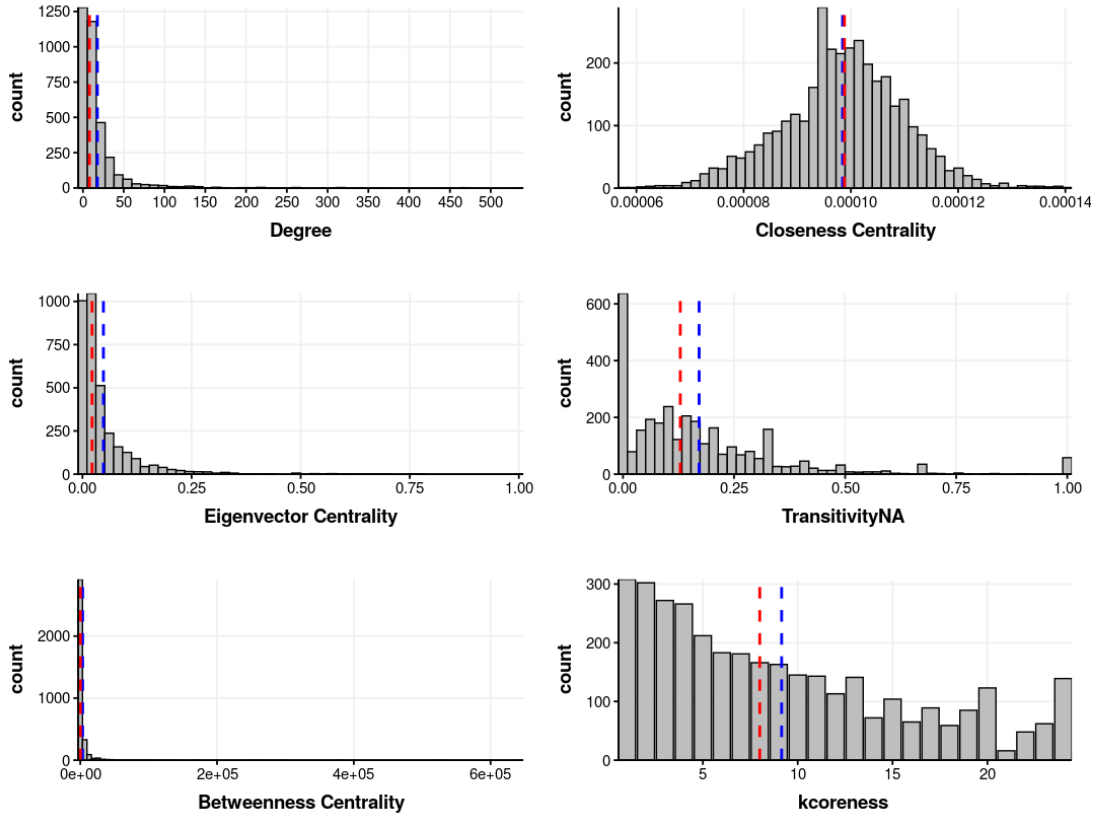


Figure 3.7: Multiplot histogram of centrality measures. Bin size kept constant at 50 other than barplot for kcoreness. Median is red dashed line, mean is blue dashed line. Plots are of (clockwise from top right: closeness centrality, local transitivity with isolates removed (transitivityNA), kcoreness, betweenness centrality, eigenvector centrality and degree centrality. Degree, eigenvector and betweenness centrality have particularly marked heavy tailed distributions suggestive of a power law. The closeness centrality looks similar to a normal distribution.

https://github.com/14760/thesis/blob/main/R/chapter3/histogram_centralities/multiplot_centralities.R

	W	p value (SW)	D	p value (KS)
Degree	0.425	2.52×10^{-74}	0.311	$< 2.2 \times 10^{-16}$
Betweenness	0.123	2.76×10^{-83}	0.434	$< 2.2 \times 10^{-16}$
Eigenvector	0.561	6.55×10^{-69}	0.269	$< 2.2 \times 10^{-16}$
Closeness	0.995	4.69×10^{-9}	0.046	7.174×10^{-7}
Transitivity0	0.775	2.65×10^{-56}	0.198	$< 2.2 \times 10^{-16}$
Kcoreness	0.911	4.32×10^{-41}	0.126	$< 2.2 \times 10^{-16}$

Table 3.8: Test for normality (Shapiro-Wilk and Kolmogorov-Smirnov) for centrality measures of PSP.

3.8 Correlation of centrality measures

The pairwise relation between centrality measures is shown in the scatterplot in figure 3.9 and figure 3.8.

3.8.1 Correlation of centrality measures

All of the centrality measures appear significantly correlated except between transitivity and betweenness centrality (table 3.9). The transitivity has been viewed as an inverse betweenness (section 3.2.8.3) and if so it is more marked with isolates set to NA when the distribution is not as dominated by zero values. Spearman's rank correlation with pairwise complete observations included was used to calculate the correlation of centrality measures. Transitivity is most weakly correlated with other measures[62].

See supplemental section B.1 for additional details.

3.9 Global network measures

3.9.1 Results Degree distribution

The degree distribution of the PSP is right-skewed. The mean degree is greater than the median (17.6 vs 8), and the maximum degree (535) is over thirty times greater than the average degree. The degree distribution $P(K = k)$ against degree k is shown in figure 3.9 on a doubly logarithmic scale. The plot follows a relatively straight line

[62] RProjects/paper_xls_output/R/make_df_correlation_between_centrality_measures

	degree	betweenness	eigenvector	closeness	transitivityNA	transitivity0	kcoreness
degree	1.000	0.902	0.879	0.851	0.196	0.375	0.986
betweenness	0.902	1.000	0.762	0.782	-0.002	0.209	0.854
eigenvector	0.879	0.762	1.000	0.969	0.330	0.462	0.906
closeness	0.851	0.782	0.969	1.000	0.292	0.430	0.872
transitivityNA	0.196	-0.002	0.330	0.292	1.000	1.000	0.265
transitivity0	0.375	0.209	0.462	0.430	1.000	1.000	0.428
kcoreness	0.986	0.854	0.906	0.872	0.265	0.428	1.000

Table 3.9: Correlation of centrality measures. Spearman's rho including only pairwise complete. Details of measures in table 3.3. All $p < 2.2 \times 10^{-16}$ other than correlation between transitivity NA and betweenness p= 0.901 using package `Hmisc` and checking with `R` stats `cor.test`.

`RProjects/db_essential_genes/R/get_corr/cor_centrality_measures` in object t
`RProjects/db_essential_genes/R/get_corr/cor_centrality_measures_performance`

in the tail of the distribution and would usually be considered to be consistent with a power law in the tail of the distribution.

Using the method described by Clauset, Shalizi and Newman (section 3.6.4) the minimum degree value (x_{min}) for which the power law distribution holds is x_{min} is 24 and the power law coefficient γ is 2.51 (table B.5)[63]. A bivariate plot of x_{min} and γ figure B.6) shows that the power law exponent increases as the value of x_{min} increases. For details on degree distribution including comparison of likelihood to other distributions see section B.1.

3.9.2 Small world property

The PSP network displays the small world property of a short average path length and high clustering coefficient compared to a random graph. Supplemental section B.1.3 contains detailed results including discussion of differences between different transitivity measures and comparisons with network small world metrics.

[63] GOF 0.19 and $n$tail (the number of nodes in the tail of the power law distribution i.e. greater than or equal to x_{min}) is 68 so there are a number of low degree nodes outside the power law portion of the degree distribution[64]

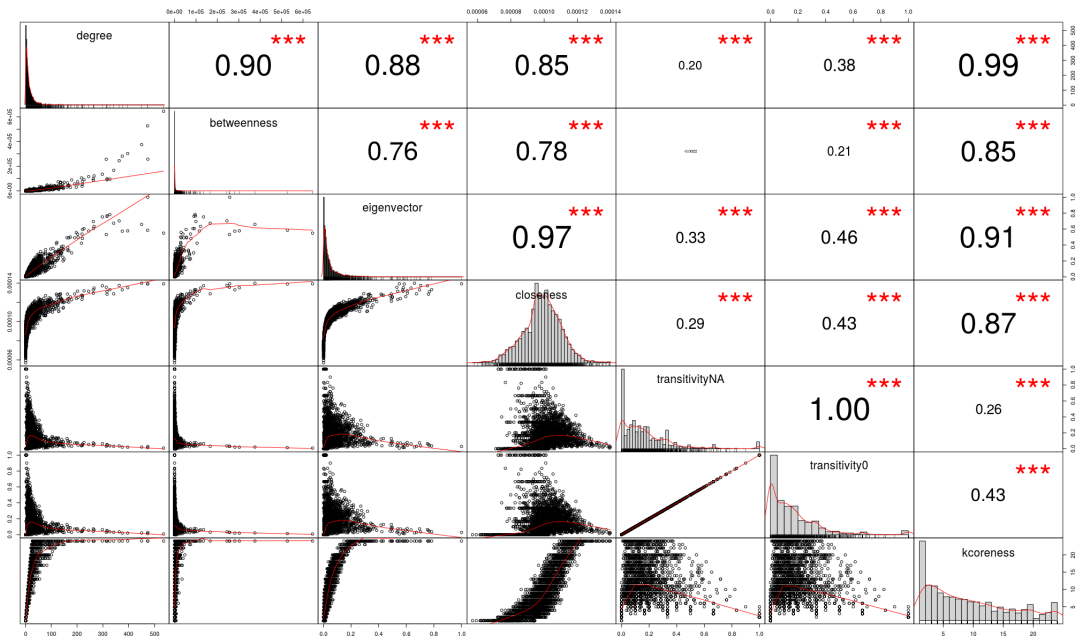


Figure 3.8: Correlation plot of centrality measures using package Performance Analytics. Deg = degree, eig= eigenvector centrality, clo = closeness centrality, bet = betweenness centrality, tra = local transitivity, kcore = kcoreness.

RProjects/db_essential_genes/R/get_corr/cor_centrality_measures_performance

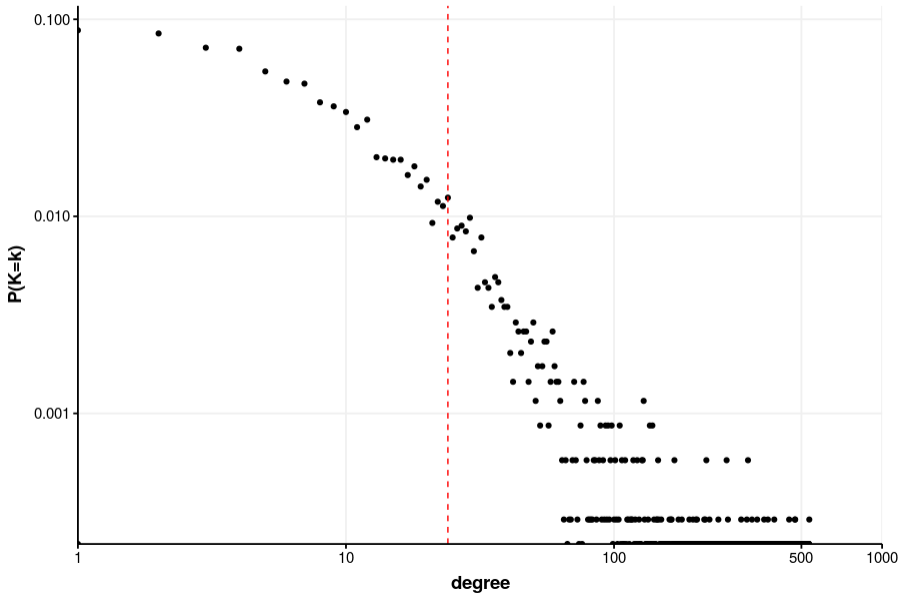


Figure 3.9: Degree distribution of post synaptic proteome. Log10 - log10 scale.

Code at RProjects/chapter3/R/fit_power_law/graphing/plot_degree_distribution_log10_theme_red_dotted_xminRProjects/gridsearch_gamma/R/plot_log_degreedist

3.9.3 Assortativity

3.9.3.1 Degree Assortativity

The PSP network is disassortative (section 3.6.6) with regard to degree (ρ_D=-0.18)[65],. This value is similar to those reported in biological networks including yeast protein-protein interaction networks (-0.156) and neural networks (-0.163) [291] and summarised in table 3.10. Amongst non biological networks it is most similar in magnitude to that of the internet in 2002[291] (r =-0.189).

A Monte Carlo estimate of the expected value of the degree assortativity for the configuration model with identical degree sequence is very close to zero: -0.00003 (10,000 iterations ; min. -0.022, max. 0.026). The degree assortativity is far from what would be expected by chance $p_{MC} < 0.0001$[66].

The degree assortativity is negative in most real world networks other than social networks (table 3.10)[291]. The random model is known to have zero correlation although low values ($r < 0.02$) may be due to chance)[307]). A standard (Barabsi-Albert) generative model for scale free networks is mildly disassortative[307] when small but also shows no degree assortativity in the limit of large n[307].

The assortativity suggests that the Barabasi-Albert model is 'incomplete' as a model of the internet [291] and by extension the PSP protein-protein interaction network and would need to be accounted for in construction probabilistic models of the network. The Callaway model of network generation (a 'grown graph' [289]) shows the positive degree correlation seen in social networks but is inaccurate as a model of other complex networks (including online social networks).

Clustering and interconnection of communities tends to lead to higher assortativity by degree and branching negative[124],[307]. Higher assortativity in social networks leads to higher modularity and clustering coefficient[307]. Community formation is more remarkable therefore in overall disassortative networks.

3.9.3.2 Assortativity other centrality measures

Table 3.11 shows that the network is most disassortative by degree. By contrast kcoreness and closeness are assortative suggesting a central area of genes (core) that are more highly connected[67]. An accurate probabilistic generative model would also need to take this into consideration.

[65] using the notation in Noldus[307]

[66] RProjects/chapter3/R/assortativity_distribution/assortativity_config

[67] RProjects/chapter3/R/assortativity_distribution/centrality/assortativity_by_centrality

network	N	*r*
internet	10697	-0.189
world wide web	269504	-0.065
post synaptic proteome	3457	-0.18
protein interactions	2115	-0.156
neural network	307	-0.163
food web	92	-0.276

Table 3.10: Degree assortativity coefficients of networks adapted from Newman 2002 [291] including post synaptic proteome

Centrality measure	assortativity (ρ)
kcoreness	0.089
Closeness	0.072
Eigenvector	-0.033
Transitivity0	-0.036
Betweenness	-0.093
Degree	-0.178

Table 3.11: Assortativity by centrality measure.
RProjects/chapter3/R/assortativity_distribution/centrality/assortativity_by_centrality

3.9.3.3 Assortativity and gene score

The guilt by association hypothesis (section 3.6.6) suggests that similar genes are linked. We can test whether genes associated with intelligence or educational attainment in the sample GWAS ore more likely to be linked considering the network as a whole by calculating the assortativity of the network by study z-score provided by MAGMA (as we are interested in correlation rather than the absolute values - see discussion).

Some genes in the PSP do not have SNPs within them that appear in the study samples and therefore not have a z-score assigned by MAGMA. These missing values are set to zero or the median value of the z-score (to ensure there is no significant difference between either strategy) as the assortativity function does not accept missing values [68].

[68] RProjects/chapter3/R/assortativity_distribution/z_score/assortativity_z.R'

Study	Missing values	Assortativity_0	$\text{Assortativity}_{\text{median}}$	$p_{BS\ 0}$	$p_{BS\ \text{median}}$
$\text{Intelligence}_{\text{Replication}}$	148	0.003	0.0032	0.78	0.79
$\text{Education}_{\text{Replication}}$	172	-0.006	-0.0064	0.20	0.18
$\text{Education}_{\text{Discovery}}$	156	0.001	-0.0002	0.67	0.59
$\text{Intelligence}_{\text{Discovery}}$	156	-0.004	-0.0038	0.31	0.33

Table 3.12: Assortativity of PSP graph vertices by study Z score. PSP graph members without corresponding Z score marked as missing values had z set to either 0 (Assortativity_0 or the median z score $\text{Assortativity}_{\text{median}}$for missing set to zero `RProjects/centrality/R/get_gene_scores/assortativity_with_z_scores` . All values not significantly different ($p > 0.05$) than bootstrap estimate of assortativity (`RProjects/chapter3/R/assortativity_distribution/bootstrap/bootstrap_zscore_assortativity_studies`

3.9.3.4 Assortativity with z score results

There is no evidence that across the PSP nodes are assortative or disassortative by the level of association of gene Z scores for studies of intelligence or educational attainment.

The results are shown in table 3.12, missing values have been assigned value 0 (Assortativity_0) or the median z score ($\text{Assortativity}_{\text{median}}$) as the nominal assortativity function in igraph requires values for all vertices. Assortativity scores tend to me modest (as score of approximately 0.15 is often found for degree assortativity). These results are close to zero and do not differ from what could be expected by chance

This suggests that nodes with high or low z scores in the population cohorts are no more likely to have links to one another *globally* across the network. It does not exclude areas within the graph of high assortativity nor distinct areas that containing high scoring genes connected by link nodes.

3.9.3.5 Assortativity with murine LTP

The ten genes sets of mammalian phenotypes associated with LTP or learning (section 3.6.9) show modest but significant assortative behaviour compared to a null model in contrast to the assortativity for population gene scores in the educational attainment and intelligence studies (table 3.12). As nominal categories the modularity (Q) is reported in table 3.13. A Monte Carlo estimate of the modularity of 10,000 random samples of vertices of the same size as the gene set is used to calculate a one tailed p

MP	Name	n	Q	max Q_{MC}	μ_Q MC	p
MP0001463	abnormal spatial learning	272	0.045	0.022	-0.001	< 0.0001
MP0001468	abnormal temporal memory	215	0.029	0.021	-0.001	< 0.0001
MP0001469	abn context condition	181	0.030	0.020	-0.001	< 0.0001
MP0001473	reduced long term potentiation	172	0.025	0.021	-0.001	< 0.0001
MP0001898	abnormal LTDep	102	0.063	0.024	-0.001	< 0.0001
MP0002062	abn assoc learning	339	0.042	0.021	-0.001	< 0.0001
MP0002062	abormal associative learning	339	0.042	0.018	-0.001	< 0.0001
MP0002063	abnormal learning memory conditioning	742	0.053	0.020	-0.001	< 0.0001
MP0004859	abnormal synaptic plasticity	77	0.049	0.025	-0.001	< 0.0001
MP0014114	abnormal cognition	744	0.053	0.020	-0.001	< 0.0001

Table 3.13: Nominal assortativity for mammalian phenotypes associated with LTP or learning. p is one tailed estimate Monte Carlo estimate of assortativity of same number of random vertices. p is proportion of random $q >= Q$. Q is modularity as a measure of nominal assortativity. max Q_{MC} is the maximum modularity found on Monte Carlo estimates (Samples:10000.),μ_Q MC is the sample mean of modularity for the 10,000 Monte Carlo estimates.

RProjects/chapter3/R/assortativity_distribution/nominal_murine/2_Feb_murine_LTP_assortativity_CI

value. For 10,000 iterations none of the Monte Carlo values of Q were greater than those in the mouse phenotypes. This suggests that the genes implicated in mammalian models of learning and LTP tend to link to one another more than would be expected by chance while the significant genes in the intelligence and educational attainment GWAS do not connect more than would be expected by chance.

Essential genes are assortative. Nominal assortativity is greater for genes having five or more appearances in the DEG (Q =0.118) than when testing using any appearance in the DEG (Q =0.0445)[69].

Assortativity (scalar) using pLI is $r = 0.081$[70]. The assortativity is more remarkable considering that high degree nodes are associated with high pLI (table 3.35) but are disassortative by degree (table 3.11).

[69] RProjects/chapter3/R/assortativity_distribution/DEG_assortatiivity/assortativity_DEG.R

[70] RProjects/chapter3/R/assortativity_distribution/DEG_assortatiivity/assortativity_DEG

3.9.4 Centrality and murine LTP

The mean centrality measures for gene sets associated with murine models of learning were compared to the PSP (Wilcoxon rank test) with correction for multiple comparisons.

No statistically significant difference from the full PSP was seen for degree (figure 3.10). Betweenness centrality was increased in these gene sets (figure 3.11) and was significantly elevated (table 3.14), after correction for multiple comparisons, in gene sets representing abnormal spatial learning (p=0.002, p adjusted BH 0.02) and abnormal cognition (p=0.003, p adjusted BH 0.026) . Closeness centrality was generally reduced (figure 3.12, table 3.15) but abnormal long term depression was the only significant phenotype after correction for multiple comparisons ($p = 0.002$, p_{BH}=0.018). Eigenvector centrality was also reduced (figure 3.13), the difference between mean eigenvector centrality for the gene set abnormal synaptic plasticity ($p = 0.002$;pBH=0.021) and abnormal long term depression ($p = 0.002$; pBH=0.021) were significant after correction for multiple comparisons. There was no significant difference for kcoreness or transitivity.

The overall picture is of a reduction in closeness centrality and eigenvector centrality with an increase in betweenness centrality. This would be consistent with vertices which tend to be bottlenecks despite being towards the periphery of the network.

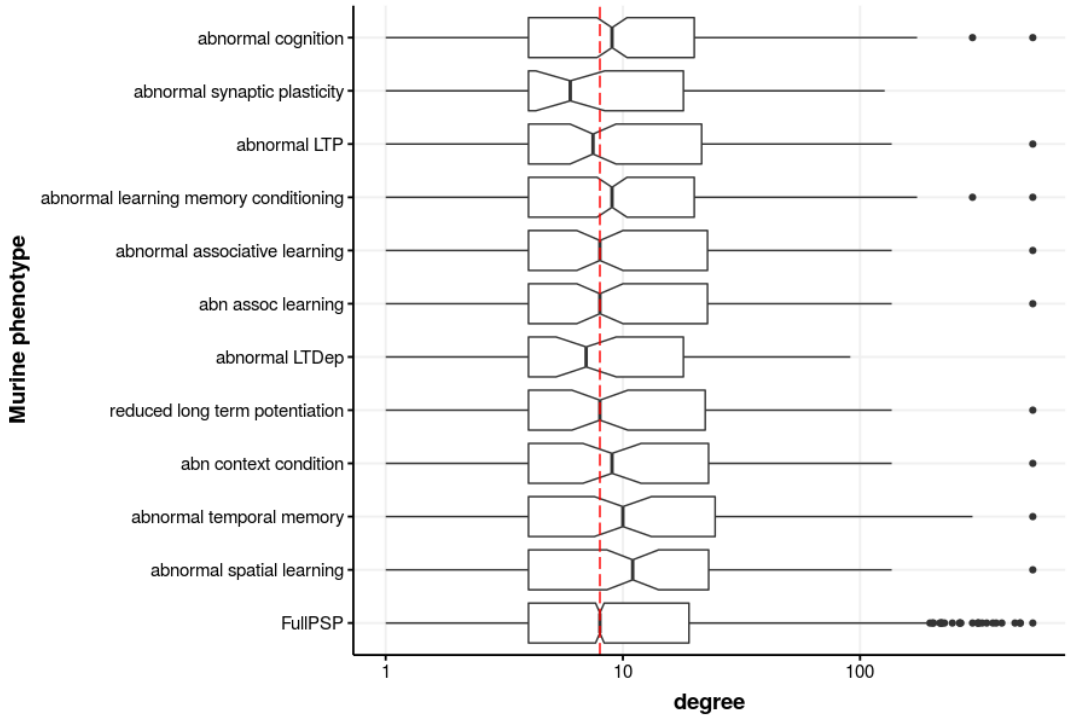


Figure 3.10: Box plot of Degree centrality distribution for differenet gene sets affecting Murine LTP or learning. Log 10 scale for degree. No statistically significant difference other than minor change in long term depression.

RProjects/chapter3/R/murine_centrality/general_murine_boxplot/by_measure/degree_centrality_boxplot_murine_ltp_general_degree_add_table

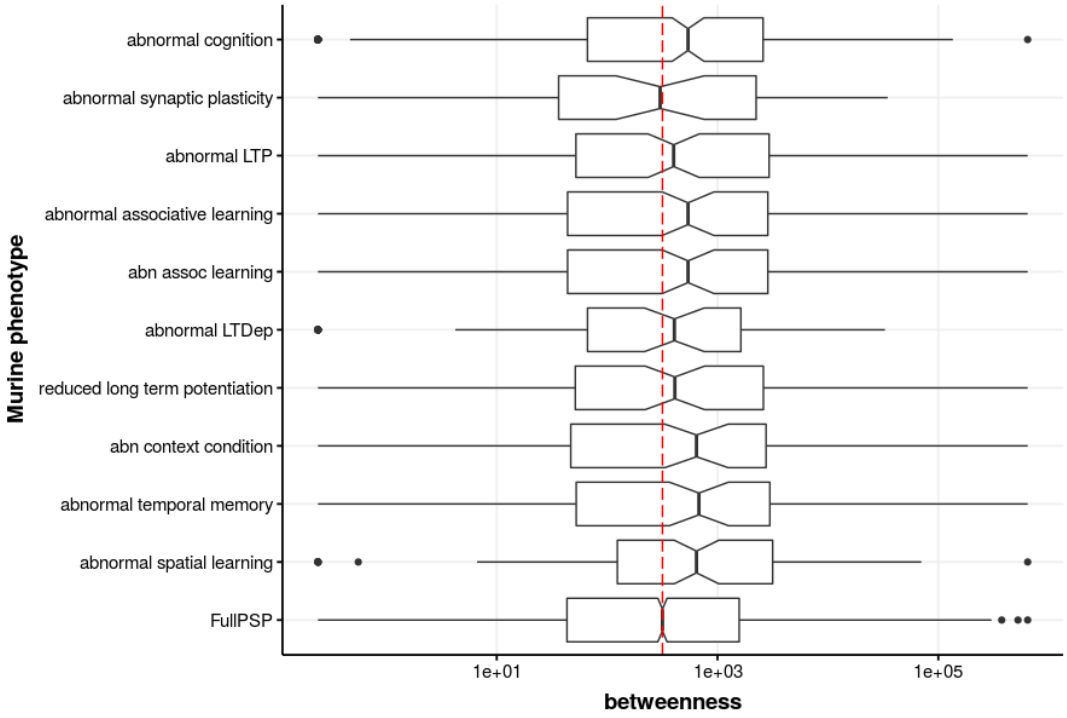


Figure 3.11: Box plot of betweenness centrality distribution for differenet gene sets affecting Murine LTP or learning. Log 10 scale for betweenness

RProjects/chapter3/R/murine_centrality/general_murine_boxplot/rough_centrality_boxplot_murine_ltp_general_closeness.R'

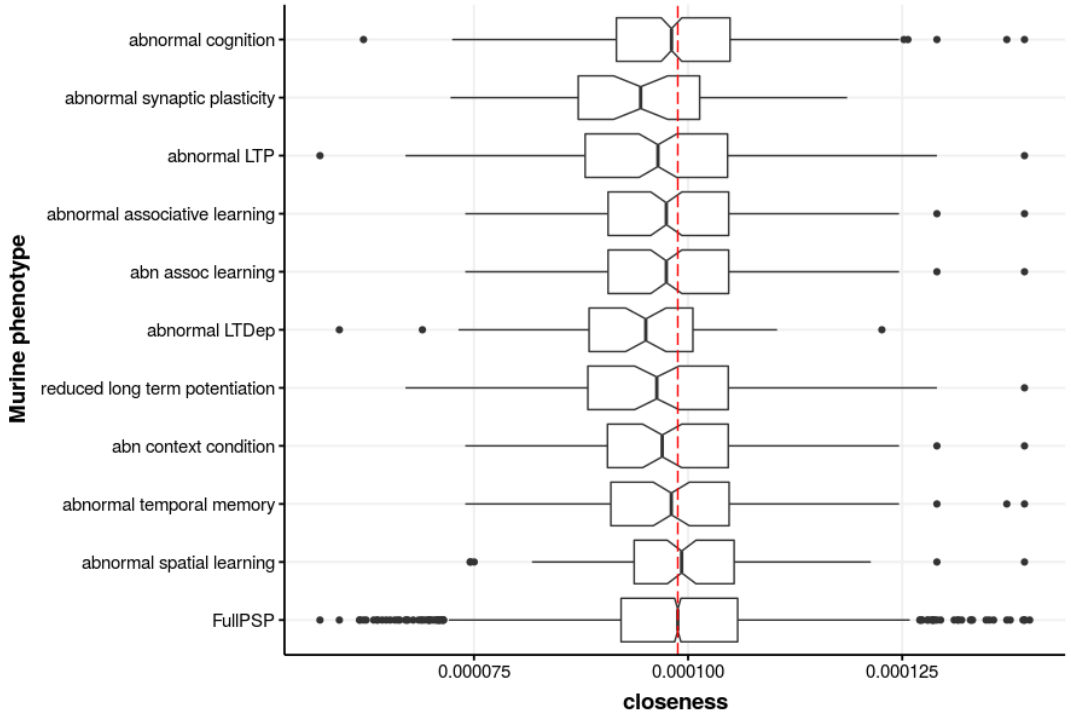


Figure 3.12: Box plot of Closeness centrality distribution for differenet gene sets affecting Murine LTP or learning. Closeness not logarithimic

RProjects/chapter3/R/murine_centrality/general_murine_boxplot/by_measure/degree_centrality_boxplot_murine_ltp_general_closeness_add_table

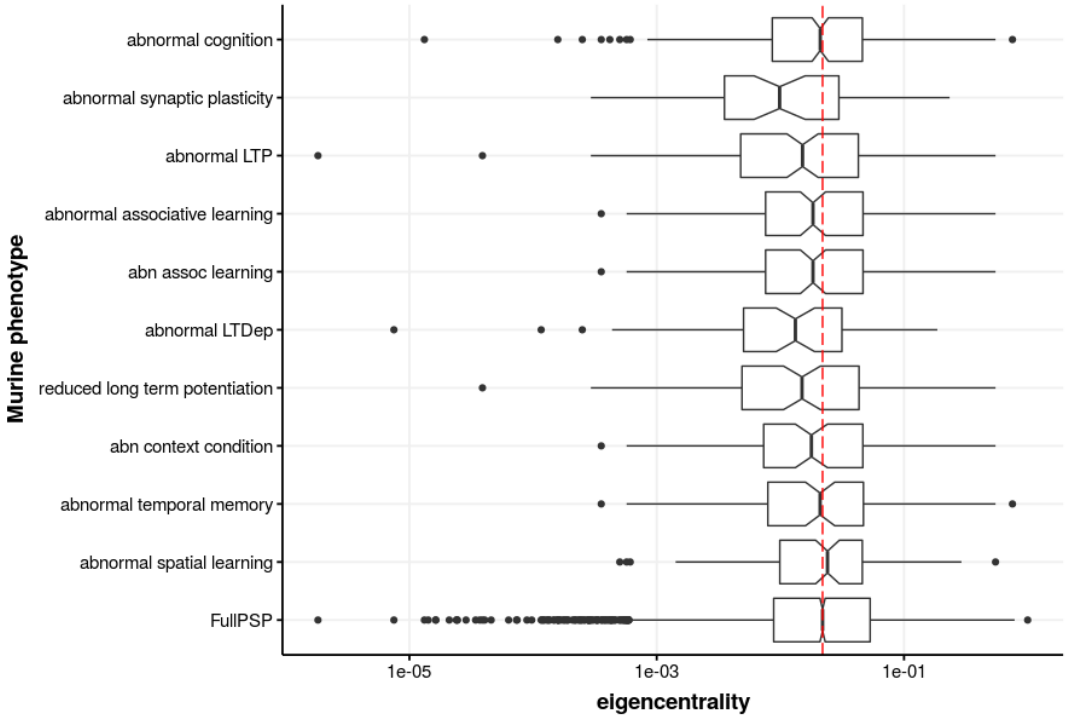


Figure 3.13: Box plot of Eigenvector centrality distribution for differenet gene sets affecting Murine LTP or learning. Log10

RProjects/chapter3/R/murine_centrality/general_murine_boxplot/by_measure/degree_centrality_boxplot_murine_ltp_general_eigenvector_add_table

	phenotype	p	p adj	p signif	method
1	abnormal spatial learning	0.002	0.020	**	Wilcoxon
2	abnormal temporal memory	0.042	0.340	*	Wilcoxon
3	abn context condition	0.114	0.570	ns	Wilcoxon
4	reduced long term potentiation	0.363	1.000	ns	Wilcoxon
5	abnormal LTDep	0.750	1.000	ns	Wilcoxon
6	abn assoc learning	0.058	0.410	ns	Wilcoxon
7	abnormal associative learning	0.058	0.410	ns	Wilcoxon
8	abnormal LTP	0.271	1.000	ns	Wilcoxon
9	abnormal synaptic plasticity	0.946	1.000	ns	Wilcoxon
10	abnormal cognition	0.003	0.026	**	Wilcoxon

Table 3.14: Mean betweenness centrality in gene sets with murine cognitive phenotypes. Comparison of mean betweenness Wilcoxon test compared with full PSP. P adjusted using Benjamini-Hochberg procedure

	phenotype	p	p adj	p signif	method
1	abnormal spatial learning	0.608	1.000	ns	Wilcoxon
2	abnormal temporal memory	0.539	1.000	ns	Wilcoxon
3	abn context condition	0.273	1.000	ns	Wilcoxon
4	reduced long term potentiation	0.088	0.610	ns	Wilcoxon
5	abnormal LTDep	0.002	0.018	**	Wilcoxon
6	abn assoc learning	0.233	1.000	ns	Wilcoxon
7	abnormal associative learning	0.233	1.000	ns	Wilcoxon
8	abnormal LTP	0.047	0.370	*	Wilcoxon
9	abnormal synaptic plasticity	0.008	0.070	**	Wilcoxon
10	abnormal cognition	0.565	1.000	ns	Wilcoxon

Table 3.15: Closeness murine phenotypes. Wilcoxon test of mean compared with all PSP genes. P adjusted using Benjamini-Hochberg procedure

	phenotype	p	p.adj	p.signif	method
1	abnormal spatial learning	0.926	1.000	ns	Wilcoxon
2	abnormal temporal memory	0.592	1.000	ns	Wilcoxon
3	abn context condition	0.321	1.000	ns	Wilcoxon
4	reduced long term potentiation	0.053	0.370	ns	Wilcoxon
5	abnormal LTDep	0.002	0.021	**	Wilcoxon
6	abn assoc learning	0.229	1.000	ns	Wilcoxon
7	abnormal associative learning	0.229	1.000	ns	Wilcoxon
8	abnormal LTP	0.012	0.097	*	Wilcoxon
9	abnormal synaptic plasticity	0.002	0.021	**	Wilcoxon
10	abnormal cognition	0.338	1.000	ns	Wilcoxon

Table 3.16: Mean eigenvector centrality for murine phenotypes. Wilcoxon test compared with full PSP. P adjusted using Benjamini-Hochberg Procedure

3.10 Characterisation of central genes

To investigate whether central genes share common properties I carried out gene ontology enrichment of the top and bottom 10% of genes for each centrality measures. Gene ontology enrichment was carried out using the 3,457 genes in the PSP as background so the results show enrichment *compared to the rest of the PSP*.

The most central genes shared a number of common features regardless of the centrality measure. The ten most enriched terms for high degree nodes in terms of degree are shown for biological process table C.2 , cellular component table C.3 and molecular function (table C.3).

Enriched biological processes, table C.2, include terms related to the cell cycle associated with low FDR q values. Molecular function terms include RNA binding and those related to transcription (table C.1). Cellular component terms, table C.3, shows enrichment for focal adhesion, cell junction terms, ribonucleoprotein and cytoskeletal fibre.

ID	Biological Process	n	n PSP	p	q FDR B H
GO:0007049	cell cycle	139	489	1.98×10^{-36}	1.25×10^{-32}
GO:0022402	cell cycle process	119	414	7.10×10^{-31}	2.25×10^{-27}
GO:0000280	nuclear division	101	319	1.43×10^{-29}	3.02×10^{-26}
GO:0048285	organelle fission	103	337	9.65×10^{-29}	1.53×10^{-25}
GO:0000278	mitotic cell cycle	95	296	3.44×10^{-28}	3.10×10^{-25}
GO:0140014	mitotic nuclear division	95	296	3.44×10^{-28}	3.10×10^{-25}
GO:1903047	mitotic cell cycle process	95	296	3.44×10^{-28}	3.10×10^{-25}
GO:0044403	symbiotic process	106	364	1.23×10^{-27}	9.33×10^{-25}
GO:0051726	regulation of cell cycle	100	329	1.33×10^{-27}	9.33×10^{-25}
GO:0044419	interspecies interaction between organisms	106	368	3.48×10^{-27}	2.20×10^{-24}

Table 3.17: **High degree genes. Biological Process - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Biological process tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category (e.g. MF,CC,BP): 1584

Eigenvector centrality (table C.17) results in particularly strong enrichment for

ID	Molecular Function	n	n PSP	p	q FDR B H
GO:0031625	ubiquitin protein ligase binding	51	117	8.81×10^{-22}	9.12×10^{-19}
GO:0044389	ubiquitin-like protein ligase binding	52	123	1.88×10^{-21}	9.74×10^{-19}
GO:0008134	transcription factor binding	58	156	1.53×10^{-20}	5.26×10^{-18}
GO:0019904	protein domain specific binding	87	329	3.13×10^{-19}	8.11×10^{-17}
GO:0003723	RNA binding	120	554	5.07×10^{-19}	1.05×10^{-16}
GO:0044877	protein-containing complex binding	111	513	2.27×10^{-17}	3.91×10^{-15}
GO:0003720	telomerase activity	21	30	5.04×10^{-15}	7.00×10^{-13}
GO:0003964	RNA-directed DNA polymerase activity	22	33	5.41×10^{-15}	7.00×10^{-13}
GO:0034061	DNA polymerase activity	23	37	1.23×10^{-14}	1.41×10^{-12}
GO:0019900	kinase binding	78	325	1.50×10^{-14}	1.55×10^{-12}

Table 3.18: **High degree genes. Molecular function - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Molecular Function tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category (e.g. MF,CC,BP): 186

RNA binding (Molecular Function GO:0003723, q FDR B H 7.01×10^{-76}) as does kcoreness (table C.24 q FDR=2.68×10^{-55}). Transitivity shows far less enrichment than the other centrality terms possible reflecting its weaker correlation. It is interesting to note however that it seems quite specific with a large *proportion* of the genes involved in mitochondrial translation and ribosomal function identified (Molecular function table C.21; 29 of 103 genes in PSP identified, fold change 9.4).

Gene ontology enrichment was carried out using the ToppGene package and the PSP as a background set. These are therefore sets of genes that differ from the background enrichment of the PSP. Given that fact it is remarkable how many terms are enriched even after correcting for multiple comparisons (186 for example in the top 10% of degree for molecular function table C.1) and 1,548 in Biological Process table C.2.

By contrast fewer terms are enriched in low centrality gene sets. For degree only two terms in cellular component were enriched (table C.4). In part this is due to the fact that the distribution of centrality functions such as degree, betweenness and eigen-

ID	Cellular Component	n	n PSP	p	q FDR B H
GO:1990904	ribonucleoprotein complex	81	303	2.75×10^{-18}	2.36×10^{-15}
GO:0015630	microtubule cytoskeleton	102	447	1.01×10^{-17}	4.33×10^{-15}
GO:0036464	cytoplasmic ribonucleoprotein granule	33	74	8.92×10^{-15}	2.55×10^{-12}
GO:0035770	ribonucleoprotein granule	33	78	5.84×10^{-14}	1.25×10^{-11}
GO:1902494	catalytic complex	86	392	1.07×10^{-13}	1.84×10^{-11}
GO:0030055	cell-substrate junction	62	244	7.12×10^{-13}	1.02×10^{-10}
GO:0005925	focal adhesion	61	239	9.13×10^{-13}	1.12×10^{-10}
GO:0000974	Prp19 complex	19	31	3.66×10^{-12}	3.92×10^{-10}
GO:0099513	polymeric cytoskeletal fiber	95	482	4.16×10^{-12}	3.96×10^{-10}
GO:0005815	microtubule organizing center	60	246	1.21×10^{-11}	1.04×10^{-9}

Table 3.19: **High degree genes. Cellular Component - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Biological process tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category (e.g. MF,CC,BP): 159

vector centrality is right skewed. Closeness which is bounded in the region 0-1 and the most approximately normally distributed provides more enrichment (table C.15 and table C.14). These are genes one would expect on the periphery of the network and the synapse, those of the receptors themselves, intrinsic components of the plasma membrane and components of the extra cellular matrix. Kcoreness which has a relatively uniform distribution over a small range also has greater enrichment of low centrality terms(table 3.20). Low eigenvector centrality genes also have common themes related to the plasma and post synaptic membrane (table C.19 and table C.20)

3.10.0.1 Low degree

The only gene ontology class enriched was cellular component (table C.4) but the enriched terms are common in other low vaules of centrality measures in that it is enriched for components of the plasma membrane. Pathway enrichment shows enrichment of the extracellular matrix and ECM proteins which is found with other low centrality measure genes. The results for high and low degree genes appear qualitatively

ID	Pathway	n	n PSP	p	q FDR B H
M5889	Ensemble of genes encoding extracellular matrix and extracellular matrix-associated proteins	37	91	1.50×10^{-10}	2.24×10^{-7}
M5884	Ensemble of genes encoding core extracellular matrix including ECM glycoproteins, collagens and proteoglycans	17	31	8.53×10^{-8}	6.38×10^{-5}
M3008	Genes encoding structural ECM glycoproteins	13	25	6.70×10^{-6}	3.34×10^{-3}
132956	Metabolic pathways	74	353	6.99×10^{-5}	2.05×10^{-2}
82989	Glycerophospholipid metabolism	12	26	7.07×10^{-5}	2.05×10^{-2}
M5885	Ensemble of genes encoding ECM-associated proteins including ECM-affilaited proteins, ECM regulators and secreted factors	20	60	9.51×10^{-5}	2.05×10^{-2}
M9131	Glycerophospholipid metabolism	11	23	9.57×10^{-5}	2.05×10^{-2}
1268710	Post-translational modification: synthesis of GPI-anchored proteins	7	11	1.96×10^{-4}	3.66×10^{-2}

Table 3.20: **Low kcoreness genes. Pathway - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Pathway tested gene set is 10th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category (e.g. MF,CC,BP): 8

different even compared to other post synaptic protein genes (ie cell cycle processes verus ECM). The network architecture seems to be related to function.

ID	Cellular Component	n	n PSP	p	q FDR B H
GO:0031226	intrinsic component of plasma membrane	51	331	1.564×10^{-5}	7.462×10^{-3}
GO:0005887	integral component of plasma membrane	47	307	4.228×10^{-5}	1.008×10^{-2}

Table 3.21: ToppGene GO: Cellular Component. 10 centile cwpsp.txtp = p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP.

Pathway table C.5

ID	Pathway	p	q FDR B H	n	n PSP
M5889	Ensemble of genes encoding extracellular matrix and extracellular matrix-associated proteins	5.689×10^{-6}	5.831×10^{-3}	19	91
1269590	Adenylate cyclase activating pathway	9.603×10^{-5}	4.798×10^{-2}	4	5
1269923	Transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds	2.551×10^{-4}	4.798×10^{-2}	5	10
1272485	Aldosterone synthesis and secretion	2.896×10^{-4}	4.798×10^{-2}	9	34
1269930	Metal ion SLC transporters	3.277×10^{-4}	4.798×10^{-2}	3	3

Table 3.22: ToppGene Pathway. 10 centile cwpsp.txtp = p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP

RProjects/chapter3/R/sig_magma_genes/central_genes_toppgene/betweenness/5_0_clip_centrality_bet_hig_mf

	Essential		
PSP	Essential	Non-Essential	Total by PSP
PSP	2,328	1,129	3,457
not PSP	6,664	9,306	15,970
Total by essential	8,992	10,435	19,427

Table 3.23: Contingency table showing essential genes where essentialness is appearing once in the Database of Essential Genes (DEG). Total genes (19427). Fisher's exact test p$< 2.2 \times 10^{-16}$. OR 2.88 95% CI 2.66 3.12; OR 2.92. Results checked Mar

RProjects/db_essential_genes/R/get_essential_genes/db_2020/findforfive/get_db_essential_greater_five_results_xtable_transpose

3.11 Essentialness and centrality

3.11.0.1 The PSP is over-represented in essential genes

Genes in the post synaptic proteome are over-represented amongst essential genes as defined in the DEG (section 3.6.10.1). The contingency table 3.23 shows whether genes are in the PSP and if they appear once in the DEG. The PSP had 2,328 genes that appeared at least once in the DEG (n genes in PSP = 3457 ; 67.3%) compared with 6,664 non PSP genes (41.7% of 15,970 non synaptic genes appearing in NCBI 37).

Using the stricter definition of essentialness, appearing in five or more DEG studies (which aims to represent the estimated 10% of human genes that are essential; section 3.6.10.1) n = 728 (21.1%) of PSP genes appear in five or more experiments

	Essential		
PSP	Essential	Non-Essential	Total by PSP
PSP	728	2,729	3,457
not PSP	1,450	14,520	15,970
Total by essential	2,178	10,435	19,427

Table 3.24: Contingency table showing essential genes where essentialness is appearing five or more times in the Database of Essential Genes (DEG). Total genes (19427). Fisher's exact test p$< 2.2 \times 10^{-16}$. OR 2.67 ;95% CI 2.42 - 2.95; OR 2.67. Results checked Mar
RProjects/db_essential_genes/R/get_essential_genes/db_2020/findforfive/get_db_essential_greater_five_results_xtable_transpose

in the DEG. By contrast the figure for non PSP genes is *n*=1,450 (9.99%) (OR 2.67 Fisher's exact test table 3.24).

Fisher's exact test for over representation in synaptic genes for any appearance in the DEG (OR 2.879; 2.663-3.115,p$< 2.2 \times 10^{-16}$,table 3.24) and more than five appearances in the DEG (OR 2.671; 2.418-2.949) are shown in table (table 3.25).

PSP genes appear to be significantly more likely to be essential (table 3.25).

Study	Odds ratio	95% CI Lower	95%CI Upper	p
Any DEG vs none	2.879	2.663	3.115	1.968×10^{-166}
Greater than 5 DEG v none	2.671	2.418	2.949	6.162×10^{-79}

Table 3.25: Fishers exact test for over representation of synaptic genes in DEG (database of essential genes). First is any appearance in DEG vs genes not in DEG. Second is those in five or more (see methods)Confidence intervals and significance for over representation of PSP proteins in Database of Essential Genes (DEG)
source= RProjects/db_essential_genes/R/get_essential_genes/db_2020/findforfive/get_db_essential_greater_five_results_xtable_transpose

The PSP genes with higher centrality measures are more likely to be essential, evidence in humans for centrality lethality effects (section 3.4) . Genes in five or more studies in the DEG have the highest centrality measures, then those appearing in one or more and the lowest mean centrality is seen in those PSP genes that are not in the DEG as can be seen in the boxplots in figure 3.14. Degree, eigenvector centrality and betweenness centrality are shown with a log10 scale. The difference in the median values is clear even in the logarithmic plot.

The correlation (Spearman's ρ) between node centrality and the number of studies in the DEG that the gene appears in is shown in table 3.26 . Higher centrality genes tend to appear in more of the DEG studies, the relationship is strongest for eigenvector centrality ($\rho = 0.398$) and weakest for transitivity $\rho = 0.207$. The betweenness centrality is the second smallest (transitivity often has weak effects compared to other centrality measures) $\rho = 0.243$.

Centrality Measure	S	ρ	p
Degree	1.401×10^9	0.334	1.265×10^{-61}
Eigenvector	1.266×10^9	0.398	3.479×10^{-89}
Betweenness centrality	1.591×10^9	0.243	9.620×10^{-33}
Closeness	1.341×10^9	0.362	4.825×10^{-73}
TransitivityNA	1.378×10^9	0.207	1.233×10^{-22}
kcoreness	1.336×10^9	0.365	4.081×10^{-74}

Table 3.26: Correlation of centrality measures and number of studies (max 20) found in DEG. Correlation type spearman.

RProjects/db_essential_genes/R/get_essential_genes/db_2020/findforfive/clean_up/get_db_essential_five_or_more_ttests_essential_centrality_correlation_count_centrality

3.12 Probability of loss of function intolerance (pLI)

The probability of loss of function intolerance is associated with measures of vertex centrality in the post synaptic proteome graph. The correlation is greatest for degree ($\rho = 0.230$) and lowest for transitivity ($\rho = 0.086$). Betweenness and eigenvector centrality have similar correlation to loss of function intolerance as degree (see table 3.35)

3.12.1 Distribution of pLI

The distribution of the probability of loss of function intolerance based on Gnomad v2.1.1 (section 3.6.11 is shown as a stacked histogram in figure 3.15 and as a density plot in figure 3.16. Both plots show a bimodal distribution of probability of loss of function intolerance with the probability mass centred at extreme high and low values and then a low level of probability mass between these extremes. The PSP genes (green in figure 3.15 and the yellow-brown line in figure 3.16) have a roughly symmetrical distribution with similar density at high and low probability of loss of function

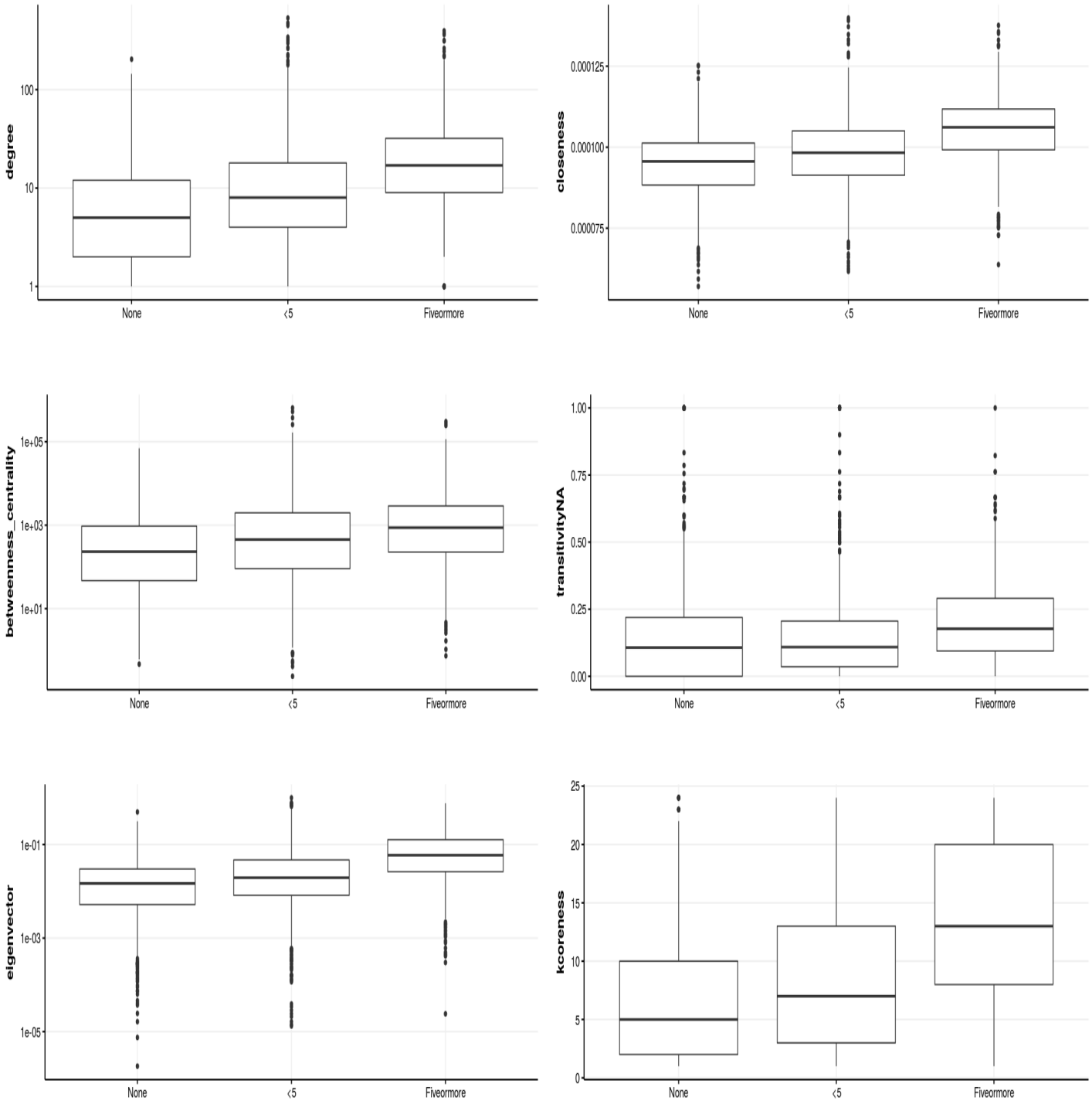


Figure 3.14: Multiplot of boxplot of centralities and DEG. On the x axis categories are non essential, Up to 5 appearances in DEG, 5 or more appearances in DEG

RProjects/db_essential_genes/R/get_essential_genes/db_2020/theme/get_db_essential_threecategoriesboxplot_multiplot_no_x_lab_essential_centrality_theme

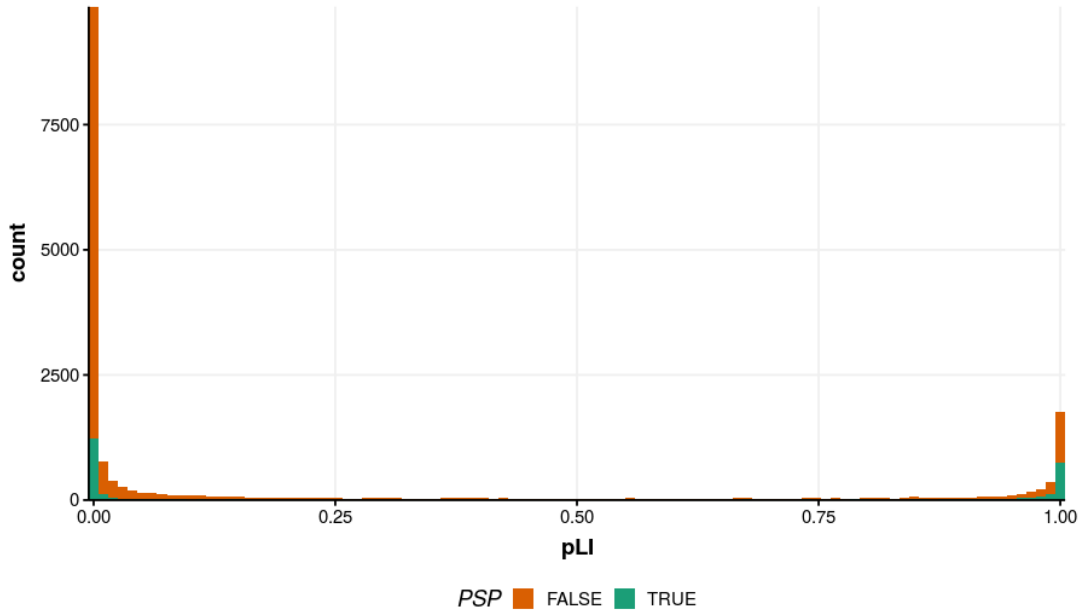


Figure 3.15: Stacked histogram showing the distribution of pLI in PSP and non PSP genes. The distribution is bimodal with most of the probability mass at extreme low and high values of pLI however the PSP genes have a more symmetrical distribution with a relatively higher peak in the high probability of loss of function intolerance range.
RProjects/db_essential_genes/R/gnomad/histogram_pli_psp/theme/histogram_pli_theme

intolerance while the peak at low levels of probability of loss of function intolerance is much higher in the non PSP genes. This reflects a greater evolutionary constraint on the PSP genes and is consistent with the findings using the database of essential genes (section 3.11.0.1) but with the additional benefit of a *measure* of selection pressure.

The median pLI is higher for the PSP genes than non PSP (0.246 vs 0.001). The median PSP is greater than the third quartile of non PSP genes (table 3.27) and it is clear that the means are significantly different (Wilcoxon rank sum test with continuity correction W=17,893,688 $p < 2.2 \times 10^{-16}$) [71].

[71] RProjects/db_essential_genes/R/gnomad/histogram_pli_psp/statistics/summary_pLI_PSP

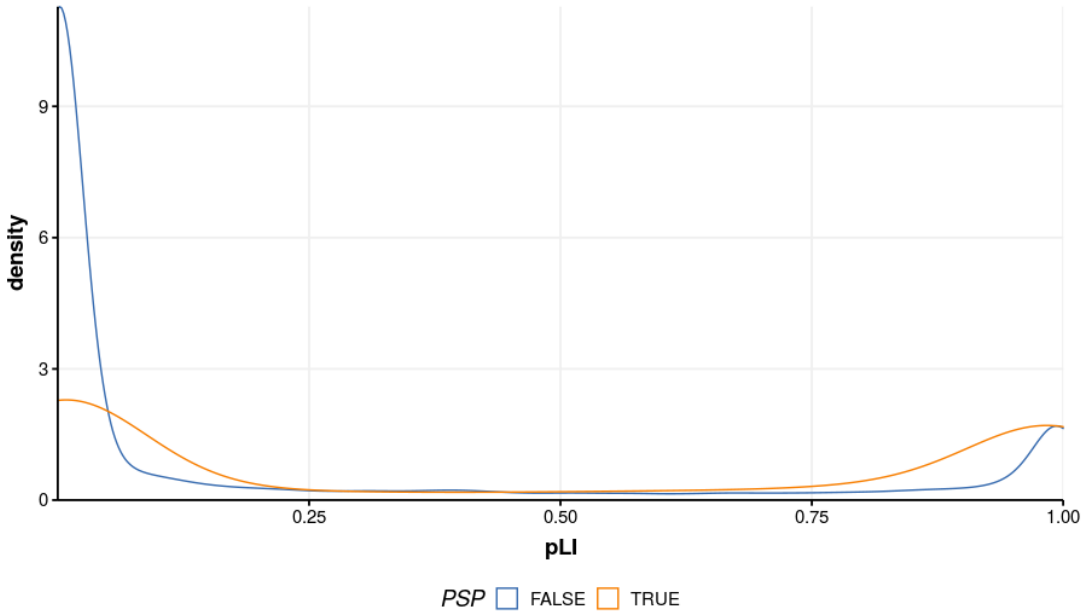


Figure 3.16: Density plot of pLI for PSP and non PSP. The distribution of probability loss intolerance is bimodal for both the PSP and non PSP genes with peaks at extreme low and high PSP values. The plot is more symmetrical for the PSP genes which shows a similar density of very high and very low probability of loss intolerance (pLI) values compared with the non PSP genes which, if they have extreme values are more likely to have low pLI

RProjects/db_essential_genes/R/gnomad/histogram_pli_psp/theme/density_pli_theme

	mean	median	sd	IQR	Q1	Q3	min	max	n
PSP	0.45	0.246	0.45	0.99	3.118×10^{-5}	0.988	1.861×10^{-164}	1	3407
non PSP	0.20	0.001	0.35	0.23	2.068×10^{-8}	0.231	8.222×10^{-118}	1	14972

Table 3.27: Probability of loss of function intolerance in genes in PSP versus those not in PSP. Wilcoxon rank sum test with continuity correction W=17,893,688, $p < 2.2 \times 10^{-16}$. No missing values. Q1 = first quartile, Q3 = third quartile, sd = standard deviation, IQR = interquartile range

RProjects/db_essential_genes/R/gnomad/histogram_pli_psp/statistics/summary_pLI_PSP

3.13 Disease enrichment for centrality measures

3.13.1 Disease enrichment

Disease enrichment is carried out using ToppGene to query the DisGenNet database. Table

Significant diseases numerous include neoplasms and viral processes. The most significant neuropsychiatric disorder is the 9th term Amyotrophic Lateral Sclerosis (C0002736; ; n in high degree 85 and n in PSP 354 p 8.15×10^{-15} FDR q 5.72×10^{-12}) (table 3.28). Amyotrophic lateral sclerosis is a neurodegenerative disorder characterised by muscular weakness. It is part of a common disease spectrum with frontotemporal dementia which is the second most enriched neurological disorder in table 3.29 (C0338451 q FDR BH 1.081×10^{-9})). The q table 3.29 is corrected by the total number of all diseases in the database (i.e not just the neurological disorders).

	ID	Name	n PSP	n set	p	q FDR B.H
1	C0042769	Virus Diseases	332	95	1.434×10^{-22}	1.124×10^{-18}
2	C0019163	Hepatitis B	290	82	6.045×10^{-19}	2.370×10^{-15}
3	C2931822	Nasopharyngeal carcinoma	294	80	2.383×10^{-17}	6.228×10^{-14}
4	C0026764	Multiple Myeloma	341	85	7.091×10^{-16}	1.390×10^{-12}
5	C0024299	Lymphoma	300	77	4.178×10^{-15}	5.720×10^{-12}
6	C0302592	Cervix carcinoma	352	85	5.658×10^{-15}	5.720×10^{-12}
7	C0033578	Prostatic Neoplasms	175	55	6.894×10^{-15}	5.720×10^{-12}
8	C0007847	Malignant tumor of cervix	315	79	6.963×10^{-15}	5.720×10^{-12}
9	C0002736	Amyotrophic Lateral Sclerosis	354	85	8.152×10^{-15}	5.720×10^{-12}
10	C1332977	Childhood Leukemia	367	87	8.194×10^{-15}	5.720×10^{-12}
11	C0079731	B-Cell Lymphomas	273	72	8.656×10^{-15}	5.720×10^{-12}
12	C0023418	leukemia	407	93	8.755×10^{-15}	5.720×10^{-12}
13	C1832661	Anopthalmia and pulmonary hyp.	298	76	9.684×10^{-15}	5.840×10^{-12}
14	C4048328	cervical cancer	340	82	2.147×10^{-14}	1.202×10^{-11}
15	C0040136	Thyroid Neoplasm	231	64	2.877×10^{-14}	1.503×10^{-11}
16	C3539878	Triple Negative Breast Neoplasms	351	82	1.510×10^{-13}	7.399×10^{-11}
17	C0476089	Endometrial Carcinoma	283	71	2.256×10^{-13}	1.040×10^{-10}
18	C0007103	Malignant neoplasm of endometrium	229	62	2.523×10^{-13}	1.099×10^{-10}
19	C0007115	Malignant neoplasm of thyroid	195	56	2.950×10^{-13}	1.141×10^{-10}
20	C0019196	Hepatitis C	362	83	3.195×10^{-13}	1.141×10^{-10}

Table 3.28: Disease enrichment using ToppGene with the PSP genes as background set for high degree (90th centile) nodes in the PSP. n PSP is the number of disease annotations in the PSP as a whole and n set is the number of disease genes in the 90th centile vertices.

RProjects/chapter3/R/toppgene_disgen/xtable/output_high_degree

	ID	Name	*n* PSP	*n* set	*p*	q FDR B.H
1	C0002736	Amyotrophic Lateral Sclerosis	354	85	8.152×10^{-15}	5.720×10^{-12}
2	C0338451	Frontotemporal dementia	119	40	4.413×10^{-12}	1.081×10^{-9}
3	C0236642	Pick Disease of the Brain	89	32	9.642×10^{-11}	1.429×10^{-8}
4	C0949664	Tauopathies	91	31	9.233×10^{-10}	1.114×10^{-7}
5	C1843013	Alzheimer disease, familial, type 3	57	22	2.097×10^{-8}	1.536×10^{-6}
6	C0524851	Neurodegenerative Disorders	484	88	3.194×10^{-8}	2.276×10^{-6}
7	C0014057	Japanese Encephalitis	30	15	6.448×10^{-8}	4.284×10^{-6}
8	C0751072	Frontotemporal Lobar Degeneration	81	26	9.075×10^{-8}	5.692×10^{-6}
9	C0040997	Trigeminal Neuralgia	17	11	1.147×10^{-7}	7.139×10^{-6}
10	C0004114	Astrocytoma	255	54	1.805×10^{-7}	1.040×10^{-5}
11	C0220756	Niemann-Pick Disease, Type C	45	18	2.235×10^{-7}	1.231×10^{-5}
12	C0020179	Huntington Disease	271	56	2.491×10^{-7}	1.347×10^{-5}
13	C0006118	Brain Neoplasms	245	52	2.930×10^{-7}	1.552×10^{-5}
14	C4551993	Amyotrophic Lateral Sclerosis, Familial	37	16	2.982×10^{-7}	1.566×10^{-5}
15	C4024896	Motor neuron atrophy	55	20	2.996×10^{-7}	1.566×10^{-5}
16	C0027765	nervous system disorder	289	56	2.308×10^{-6}	8.785×10^{-5}
17	C0002622	Amnesia	49	17	5.056×10^{-6}	1.680×10^{-4}
18	C0009241	Cognition Disorders	173	38	5.876×10^{-6}	1.927×10^{-4}
19	C0026837	Muscle Rigidity	65	20	6.135×10^{-6}	2.004×10^{-4}
20	C4552000	Episodic Kinesigenic Dyskinesia 1	53	17	1.682×10^{-5}	4.866×10^{-4}

Table 3.29: Disease enrichment for neurological disorders using the 90th centile of degree distribution compared with a background set of all PSP genes. Correction for multiple comparisons by Benjamini and Hochberg method (q FDR BH) and is corrected using full list of diseases (a filtered list is presented). Twenty most enriched gene sets shown ordered by q FDR. The number of genes annotated to the disease term in the PSP is n PSP and the number in the top 90th centile of genes by degree in PSP is n set

RProjects/chapter3/R/toppgene_disgen/xtable/output_high_degree

3.13.2 Distribution of pLI and intelligence

Sample	S	ρ	p
$\text{Intelligence}_{\text{Replication}}$	8.41×10^{11}	0.031	3.542×10^{-5}
$\text{Education}_{\text{Replication}}$	7.95×10^{11}	0.046	2.054×10^{-9}
$\text{Intelligence}_{\text{Discovery}}$	8.18×10^{11}	0.057	8.756×10^{-14}
$\text{Education}_{\text{Discovery}}$	8.01×10^{11}	0.077	6.338×10^{-24}

Table 3.30: Correlation pLI and z score for gene significant in difference samples. Spearman coefficient.

RProjects/chapter3/R/magma_gnomad/compare_Mar/0_load_magma_and_load_gnomad

Sample	t	df	CI low	CI high	r	p
$\text{Intelligence}_{\text{Replication}}$	7.526	17334	0.042	0.072	0.057	5.470×10^{-14}
$\text{Education}_{\text{Replication}}$	9.234	17098	0.056	0.085	0.070	2.893×10^{-20}
$\text{Intelligence}_{\text{Discovery}}$	10.043	17324	0.061	0.091	0.076	1.150×10^{-23}
$\text{Education}_{\text{Discovery}}$	14.230	17324	0.093	0.122	0.107	1.078×10^{-45}

Table 3.31: Correlation pLI and z score Pearson correlation coefficient. CI low and high are 95% confidence intervals for correlation

RProjects/chapter3/R/magma_gnomad/compare_Mar/0_load_magma_and_load_gnomad

There was a weak association between probability of loss of function intolerance and genes having variants associated with differences in intelligence or educational attainment. In table 3.30 there is a weak but significant correlation between z score and probability of loss of function intolerance. Genes that were more likely to be significant in these samples were more likely to be loss of function intolerant. Table 3.33 shows the correlation using the negative log10 transform of the p value which is commonly quoted in GWA studies. As these are in the same rank order as the z score these are identical and show that more significant genes tend to be are more loss of function intolerant. The relationship is strongest in the $\text{Education}_{\text{Discovery}}$ sample (ρ=0.077). Results showing the difference in relationship between synaptic and non synaptic genes in these cohorts are shown in table 3.34 [72] the relationship is stronger for synaptic

[72] RProjects/paper_xls_output/R/make_df_correlation_pLI_genescore

genes than for non synaptic genes although the effect is small and synaptic genes as shown below have a higher probability of loss of function intolerance.

The distribution of pLI is bi-modal and a non-parametric test of association is most appropriate. For comparison the Pearson correlation between pLI and gene z score is shown and highest for $\text{Education}_{\text{Discovery}}$ (r=0.107) and lowest for $\text{Intelligence}_{\text{Replication}}$(r=0.057) (table 3.31). Using the negative log 10 transform of the MAGMA p values the highest Pearson correlation was with $\text{Education}_{\text{Discovery}}$ r=0.095 (table 3.32). The $\text{Education}_{\text{Discovery}}$ sample shows the highest correlation between log10 p and pLI (ρ=0.077) (table 3.33)

Sample	t	df	low	high	r	p
$\text{Intelligence}_{\text{Replication}}$	8.161	17334	0.047	0.077	0.062	3.546×10^{-16}
$\text{Education}_{\text{Replication}}$	10.534	17098	0.065	0.095	0.080	7.214×10^{-26}
$\text{Intelligence}_{\text{Discovery}}$	9.735	17324	0.059	0.089	0.074	2.440×10^{-22}
$\text{Education}_{\text{Discovery}}$	12.544	17324	0.080	0.110	0.095	6.141×10^{-36}

Table 3.32: Correlation between pLI (probability of loss of function intolerance) and negative log10 p value. Pearson correlation. Low and high refer to 95% confidence intervals.

RProjects/chapter3/R/magma_gnomad/compare_Mar/0_load_magma_and_load_gnomad

Sample	S	ρ	p
$\text{Intelligence}_{\text{Replication}}$	8.41×10^{11}	0.031	3.542×10^{-5}
$\text{Education}_{\text{Replication}}$	7.95×10^{11}	0.046	2.054×10^{-9}
$\text{Intelligence}_{\text{Discovery}}$	8.18×10^{11}	0.057	8.754×10^{-14}
$\text{Education}_{\text{Discovery}}$	8.01×10^{11}	0.077	6.340×10^{-24}

Table 3.33: Correlation pLI and the negative log10 transform of gene p values. Spearman correlation

RProjects/chapter3/R/magma_gnomad/compare_Mar/0_load_magma_and_load_gnomad

3.13.2.0.1 Subsets The correlation between pLI and zscore is shown for subsets of the PSP, non PSP and all genes (table 3.34). The highest correlation is again in the $\text{Education}_{\text{Discovery}}$ sample (ρ=0.077 all genes, p=6.34×10^{-24}) and correlation is highest for genes in the PSP (ρ=0.106;p= 1.284×10^{-9} for $\text{Education}_{\text{Discovery}}$) The

correlation is also shown between pLI and the negative log 10 transform of the gene p value (table 3.34) which have identical coefficients (ρ) as the correlation with the z score as they will be in the same rank order.

Sample	subset	n	S	ρ	p
$\text{Intelligence}_{\text{Replication}}$	PSP	3276	5.59×10^{9}	0.047	7.362×10^{-3}
	non-PSP	14060	4.55×10^{11}	0.018	3.601×10^{-2}
		17336	8.41×10^{11}	0.031	3.542×10^{-5}
$\text{Education}_{\text{Replication}}$	PSP	3254	5.43×10^{9}	0.055	1.698×10^{-3}
	non-PSP	13846	4.25×10^{11}	0.039	5.707×10^{-6}
		17100	7.95×10^{11}	0.046	2.054×10^{-9}
$\text{Intelligence}_{\text{Discovery}}$	PSP	3270	5.36×10^{9}	0.080	4.931×10^{-6}
	non-PSP	14056	4.45×10^{11}	0.038	6.317×10^{-6}
		17326	8.18×10^{11}	0.057	8.754×10^{-14}
$\text{Education}_{\text{Discovery}}$	PSP	3270	5.21×10^{9}	0.106	1.284×10^{-9}
	non-PSP	14056	4.38×10^{11}	0.054	1.528×10^{-10}
		17326	8.01×10^{11}	0.077	6.340×10^{-24}

Table 3.34: Correlation between loss of function intolerance (pLI) and negative log10 transform of P value comparing genes in the PSP, those not in the PSP and all genes. PSP genes are the most loss of function intolerant.

Edited from `RProjects/chapter3/R/magma_gnomad/compare_Mar/0_load_magma_and_load_gnomad_subset`

3.13.2.0.2 Centrality and pLI There is a significant correlation between all centrality measures and probability of loss of function intolerance (table 3.35). The relationship is strongest for kcore and weakest for the transitivity measures. Given the right skew of some of the measures it was possible that lower values of centrality did little to distinguish between vertices therefore I calculated the effect of removing the bottom 0.2 quantile but there was very little effect other than removing any effect from $\text{transitivity}_{\text{NaN}}$[73]

[73] `RProjects/chapter3/R/centrality_ch3/centrality_and_gwastudies/0_load_magma_spearman_centrality_subset`

Centrality	S	ρ	p
kcoreness	5.07×10^9	0.230	3.30×10^{-42}
Closeness	5.11×10^9	0.225	2.83×10^{-40}
Degree	5.11×10^9	0.224	4.01×10^{-40}
Eigenvector	5.19×10^9	0.213	3.32×10^{-36}
Betweenness	5.27×10^9	0.201	2.38×10^{-32}
Transitivity_0	5.79×10^9	0.121	1.26×10^{-12}
$\text{Transitivity}_{\text{NaN}}$	4.51×10^9	0.098	4.68×10^{-8}

Table 3.35: Correlation of pLI and vertex statistics. Spearman rank correlation using for $\text{Transitivity}_{\text{NaN}}$ pairwise complete observations. Ordered by correlation coefficient. All measures of centrality are associated with greater loss of function intolerance.

RProjects/chapter3/R/centrality_ch3/centrality_and_gwastudies/0_load_magma_spearman_centrality

3.13.3 Correlation centrality measures and study p

There is no significant correlation between any of the centrality scores and the results of the Gene level MAGMA analysis for the intelligence samples (table 3.36) or the educational attainment samples (table 3.37).

3.13.4 Summary

Summary central genes are associated with loss of function intolerance. Significant genes for intelligence and cognitive ability are slightly more likely to be loss of function intolerant .The relationship is strongest for the studies with the greatest power (ie the discovery cohorts). There is no association between centrality and significance in intelligence of cognitive ability.

	Discovery			*Replication*		
Centrality	S	ρ	p	S	ρ	p
Degree	6.20×10^9	-0.03	0.04	5.91×10^9	0.02	0.20
Betweenness	6.20×10^9	-0.03	0.05	5.91×10^9	0.02	0.21
Eigenvector	6.16×10^9	-0.03	0.11	6.05×10^9	-0.00	0.93
Closeness	6.14×10^9	-0.02	0.17	6.08×10^9	-0.01	0.68
$\text{Transitivity}_{\text{NaN}}$	4.55×10^9	0.00	0.85	4.68×10^9	-0.02	0.34
Transitivity_0	5.96×10^9	0.01	0.75	6.08×10^9	-0.01	0.67
kcoreness	6.18×10^9	-0.03	0.07	5.92×10^9	0.02	0.27

Table 3.36: Correlation of significance of gene in intelligence samples and centrality measures. Spearman Correlation of correlation coefficient and negative log 10 transform of p value. No significant correlation with centrality is found.

RProjects/chapter3/R/centrality_ch3/centrality_and_gwas_p_val/combined_tables/spearman/intelligence_centrality_correlation

	Discovery			*Replication*		
Centrality	S	ρ	p	S	ρ	p
Degree	5.84×10^9	0.01	0.53	6.04×10^9	-0.01	0.69
Betweenness	5.82×10^9	0.02	0.39	6.01×10^9	-0.00	0.92
Eigenvector	5.95×10^9	-0.01	0.65	6.18×10^9	-0.03	0.07
Closeness	5.96×10^9	-0.01	0.65	6.19×10^9	-0.03	0.06
$\text{Transitivity}_{\text{NaN}}$	4.63×10^9	-0.03	0.10	4.59×10^9	-0.01	0.77
Transitivity_0	6.01×10^9	-0.02	0.31	6.05×10^9	-0.01	0.62
kcoreness	5.88×10^9	0.00	0.79	6.06×10^9	-0.01	0.52

Table 3.37: Correlation of centrality measures for genes with negative log 10 transform of p value in GWA studies of Educational attainment. No correlation is seen between centrality and enirchment with genetic variants associated with Educational Attainment.

RProjects/chapter3/R/centrality_ch3/centrality_and_gwas_p_val/combined_tables/spearman/education_centrality_correlation

3.14 Summary of chapter

We show that in the PSP nodes of very high centrality are associated with essentialness. The PSP has a high degree of essentialness compared to the rest of the genome.

High degree nodes are associated with specific phenotypes and specific gene ontology terms however the are not correlated with genetic differences that are enriched for differences in intelligence. This is not obvious before the fact and is different from other traits eg essentialness. One might hypothesis that the reason is that there is either no effect (ie it is irrelevant) or that in population studies genes with high centrality are more likely to be essential and hence have less genetic variation. We find also in one of the principle animal models of learning murine long term potentiation that genes of high importance were no more likely or less likely to be involved in long term potentiation.

We used data from the Gnomad project and DEG to further investigate whether high importance nodes were subject to selection pressure such that they proved more intolerant of non synonymous genetic changes. Nodes with high degree are associated with an increased probability of being intolerant to changes. This is seen in other measures of centrality and provides evidence from human population studies of the effects of centrality

Nodes with high betweenness centrality were involved in neuro-degenerative disorders see section 3.2.4. They were also associated with severe phenotypes in murine models including phenotypes incompatibly with viability. Low betweenness centrality was associated with abnormalities in synaptic transmission in murine models.

High eigenvector centrality is again associated with severe murine phenotypes and enrichment for RNA and cadherin binding. Disease enrichment is of neoplasms and neurodegenerative disorders.

Low eigenvector centrality (10%) were enriched for human phenotypes of seizures and ontology terms of transmitter, mouse phenotype was of abnormal synaptic transmission. The top five enriched disorders in DisGeNET were seizure related.

Despite this there was no association between node centrality measures and the significance of genetic variants for differences in intelligence and educational attainment.

There is a complex relationship between transitivity and degree for nodes with degree 3 or less the correlation is positive and results in a positive overall correlation despite a plot of the relationship appearing clearly negative. This is due to the large number of low degree nodes. When the correlation is calculated for higher degree

the relationship previously commented upon in the literature [289][74] is seen. Caution should therefore be used in determining linear correlations involving transitivity especially in smaller networks and it may be prudent to plot the correlation against minimum degree.

Gene sets from animal models of cognition have modest but significantly increased assortativity.

[74] p 335

Chapter 4

Assessment and selection of clustering algorithms

GSA is limited to curated gene sets, and it is hard to 'discover' new gene sets. A prominent network orientated GSA methods paper called for data-driven gene sets[228]. Clustering of the network using community detection algorithms can provide data-driven lists to use as gene sets.

Community specific enrichment using GSA would be a good test of the hypothesis that specific communities are associated with genetic differences related to intelligence and educational attainment.

There are many community detection methods, so careful evaluation and selection of appropriate algorithms is needed independent of GSA. This chapter assesses the performance of community detection algorithms to determine which would be most appropriate to provide communities for GSA.

The Lancichinetti–Fortunato–Radicchi is helpful in selecting the optimal algorithm when augmented with network specific tests.

The main finding of this chapter is that spectral method of clustering identifies a cluster of genes encoding proteins that form a group around glutamate receptors that have an enriched association with both intelligence and educational attainment using Gene Set Analysis. This group would not be discovered by the (standard) use of curated gene sets in Gene Set Analysis.

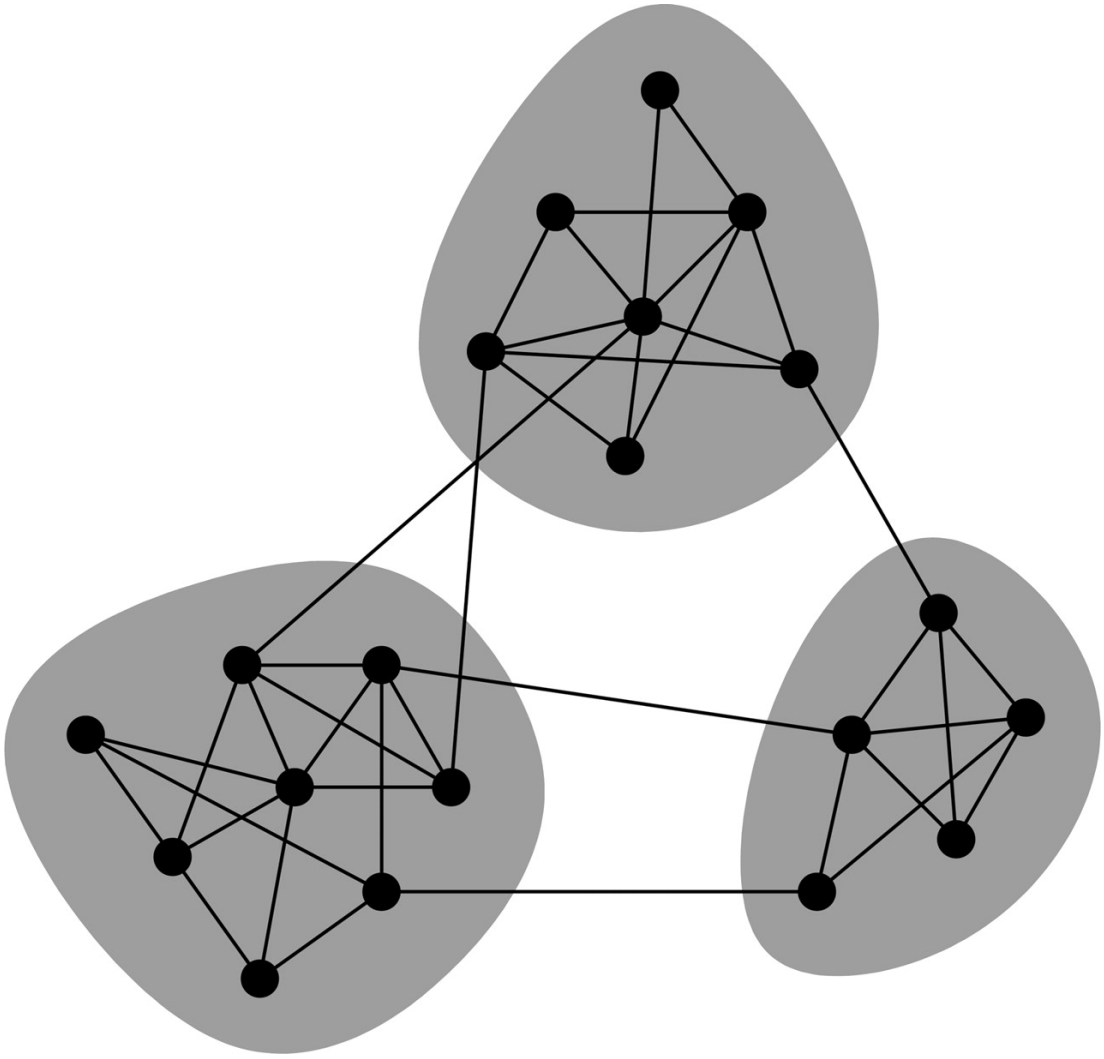


Figure 4.1: Line drawing of three communities showing dense connections within the communities (shaded) and fewer connections between communities

Creative commons for attribution see https://commons.wikimedia.org/wiki/File:Network_Community_Structure.svg

4.1 Network communities

Communities are sub-graphs of a network with dense interactions between community members and sparser connections to the rest of the network (figure 4.1), [294],[134],[146]. They are the most well studied large-scale structure in networks[294] and are missing from random networks. Examples include a group of friends in a social network. They are more likely to know one another, and so form a cohesive, densely interconnected group. Community detection can detect groupings in networks corresponding to functional units using only the pattern of interconnections[2],[452][289].

Communities are a feature of real-world networks and do not occur in random models. In the synaptic protein network, communities may represent functional modules[340],[266].

Community detection using smaller, earlier postsynaptic protein networks found communities with differential enrichment for disease and Gene Ontology terms using over-representation GSA[340],[266]. One paper used data from GWA studies along with over-representation GSA[144] of communities found in a global interaction network that included PPI. Ghiassian et al. added genes with genome-wide significant SNPs in the Phenotype Genotype integrator database (PhenGenI) to gene-disease associations from other sources to analyse seventy different disorders[350]. This analysis, therefore, did not use all the information available in summary GWAS statistics. The edges in the interaction network could come from evidence other than direct protein-protein interaction such as co-expression or metabolic enzyme coupled reactions. In addition, the network was not representative of the tissues affected by the diseases or traits under consideration[144]. Different tissues will contain different proteins, and so the protein interaction network will have different nodes and edges. Vertex (node) statistics such as degree will be different, and community detection algorithms may give different results.

The hierarchical divisive edge betweenness community detection method (Girvan–Newman algorithm) [146] has been used (section 4.3.1) to discover communities in the protein-protein interaction graph of the NRC/MASC complex (248 edges, 105 nodes)[340]. An advantage of this algorithm is that it does not need the number of communities to be determined in advance. Instead, the division of the network stops when the maximal modularity (Q) is obtained[146].

4.1.1 Modularity

Modularity (Q) was used as a measure of nominal assortativity (equation 3.20, section 3.3) the extent to which vertices that share a property are linked. A natural extension is as a measure of the quality of the division of the network into multiple groups, communities, where community membership is the property that nodes possess that maximises the modularity.

The most common method of community detection seek to maximise an objective function typically the modularity (see section 4.1.1 and Eq(B.8)) over possible partitions of the network to obtain non-overlapping groupings of the nodes[295]. Alternately a statistical model can be used to find the community allocation most likely to generate the network although modularity and maximum likelihood approaches have been shown to be equivalent[296]. Modularity maximisation is computationally hard (NP complete[56]) and a number of heuristics have been described to provide approximations to the optimal modularity[295]. Community detection algorithms also vary by being divisive or agglomerative and hierarchical (where smaller clusters are nested in larger ones) or partitional (where the best division of the network is sought without regard for hierarchy[355]).

With increasing graph size some algorithms can become unstable, give rise to improbable community sizes, or become prohibitively computationally expensive[340],[266]. McLean et al.[266] described a parallelised implementation of the edge betweenness algorithm used previously on the NRC-MASC complex[338] and found it to be prohibitively slow with large networks. With a contemporary (2016) version of the human PSD proteome (V=1,312, E=8,031) the algorithm took 14 days (20,160 minutes) and the estimated completion time for a sequential implementation was 26 days (37,889 minutes). For all edge betweenness algorithms including parallelised implementations there is a steep inflection point in running time at around 4,000 edges (figure 1 in the manuscript[266]).

One method of community detection that scales well with network size makes use of the spectral properties of the matrix representations of the network[295]. The network can be recurrently partitioned using the sign of the leading eigenvector of the modularity matrix[295]. McLean et al [266] provide an implementation of this algorithm, which we will use in this chapter, along with the random walk and edge betweenness algorithms optimised for computers with multiple cores. Examining the performance of these algorithms on an earlier network model of the human post synap-

tic density (PSD) proteome (V=1,312, E=8,031) they found that the spectral clustering method with a fine-tuning step provided the best performance, measured by running time, community size and differential functional enrichment.

In this chapter I will first try to find the best community detection algorithm for the larger (V=3,457) PSP. This will expand upon the analysis of McLean[266] et al. as I will look in particular for algorithms that produce communities well suited for gene set analysis. The size of the communities generated is important as confounding effects can occur with large gene sets[100] and there may also be an optimum size of cohesive communities [244]. On the other hand large numbers of small communities will result in sets dominated by the most significant genes and an increased penalty for multiple testing with a resultant decrease in power in analysing population GWAS-GSA. Algorithms that are prohibitively slow on networks of over 3000 nodes will not be suitable and algorithms that are 'self-tuning', lacking additional parameters or the necessity to determine community size in advance, are preferred as they simplify study design. If there are not clear default or optimal parameters settings then each different setting could (should?) be regarded as a different test increasing type I error.

Prospective algorithms in addition should be readily implementable and produce communities that 'look right' and *generally* correspond to functional units.

4.2 Community Detection

Before testing communities using GSA we have to decide on an appropriate community detection algorithm or algorithms.

4.2.1 Assessment of community detection

A number of benchmarks for community detection algorithms exist including the use of networks where we know the community allocation of the vertices ('Ground truth'[444], examples include Zachary's Karate club[452]) or the use of benchmark networks which can be simple (Girvan-Newman, 128 nodes, four communities[146]) or derived from a more realistic generative model (Lancichinetti–Fortunato–Radicchi (LFR)[230]). Although they have been used to determine the 'best' algorithm under different circumstances ([446]) they have been less widely reported to select an algorithm for a particular the context of network analyses. The reason is probably that these analyses of algorithms[446], [7] are very involved but a simple procedure is to choose the LFR

benchmark most similar to the network one is studying and assess algorithm performance on it in terms of performance measures and community size.

Looking at the performance of community detection algorithms on ground truth communities Yang and Leskovec (2015) describe the usual process of assessing community detection performance. It bears quoting at length:

"Generally, after some community detection algorithm identifies communities based on the network structure, the essential next step is to interpret the communities by identifying a common external property or a function that the members of a given community share and around which the community organizes. *For example, given a protein–protein interaction network of a cell, one first identifies communities based on the structure of the network and then examines that these communities correspond to real functional units of a cell*[444][1]."

This has been the standard approach and has the obvious disadvantage that one only finds what one expects to find. In the example above one must *know* the "real functional units of the cell". There is no chance to find new structure or organisation although it is well recognised that valuable structure can be discovered that differs from 'ground truth'[134][2].

In studies of community detection in biological networks Gene Ontology definitions[266], Gene RIF, gene lists of gene disease associations (GDA) such as OMIM [168][3], ClinVar [231][4], CTD (Comparative Toxicogenomics Database) [98][5], dbGAP [406] DisGeNET [335] [334] and GWAS catalog[431] have been used to assess whether the community detection algorithm is returning 'natural' communities in the belief that a community enriched for these terms represents a functional unit, an entity that differs from a random selection of the elements that make up the network.

There are a number of assumptions implicit in this including: that Gene Ontology or disease terms are precisely mapped, that a module has one function and that we know or can enumerate all functional units in a tissue (based on Gene Ontology or another repository).

One problem with this approach is that these lists of genes do not correspond closely with the findings of large scale population genetic studies. For example, in DisGeNET over 1,000 genes are associated with schizophrenia from curated sources

[1] italics mine
[2] p17
[3] (application [144])
[4] application [258] ref 46
[5] application [20]

and the total number of genes found in the database is 1,700[335]. This represents approximately 8.5% of protein encoding genes in the genome [126]. This contrasts with 341 genes found in the Psychiatric Genomics Consortium–Schizophrenia Workgroup (PGC–SCZ) genome wide association study [360]. These genes include those found in animal models or drug targets or may be found in diseases associated with schizophrenia. Using such a heterogenous group may lead to a set of genes with a limited connection to schizophrenia appearing to have a significantly enriched association.

There is also the issue of the definition of disease. DisGeNET disease refers to 'actual diseases, disease symptoms and abnormal phenotypes that are observed as disease manifestations as well as normal traits and phenotypes that are currently explored in large scale Genome Wide Association studies (GWAS)[335]'. Statements made about diseases and network topology such as where disease genes are within the interactome [23],[144] may refer to a number of different entities using Gene Disease Association. It may be better to limit a network analysis to a specific diseases or set of diseases and to differentiate between genetic variation associated the disease, gene expression in a specific phase of the disorder or other factors such as the genes targeted by common therapies. Finally all of the genes in these lists are scored equally in over-representation analyses whether they have a very strong association with the disorder or a weak one. Using information from GWA catalog to populate a gene disease association list does not provide the full polygenic signal from a group of genes found in summary GWA data.

One way to assess clustering algorithms utility is to see whether they discover communities that are associated with a disease or trait in population studies where the disease or trait is operationalised as much as possible (GWAS)[67][410]. By using competitive Gene Set analysis using GWA study data as implemented in programs such as MAGMA[239] or PASCAL[228] maximum information is retained from summary GWAS data and we may also discover new insights into the subject of the GWA study.

We will still need some criteria to decide upon which algorithm to use in this GSA paradigm. Before turning to this we have to introduce the community detection methods.

4.3 Choice of community detection algorithm

Given the number of community detection algorithms that are described it is likely none is optimal[7]. The Girvan-Newman (GN) benchmark [146] is a standardised benchmark with 128 nodes and four communities but is regarded as being too simple[7]. One limitation is that communities and degree are similar and the more demanding Lancichinetti–Fortunato–Radicchi (LFR)[230] allows scale free community size and degree(section 4.4). The other main way to assess algorithms is to compare communities with networks in which the 'ground truth' (or metadata associated with the nodes) is known[6][243].

Algorithms that perform well on small data sets often do not scale to larger networks[146],[266],[446]. In addition we would like to partition the network into communities that are both topologically robust but make sense biologically. It is hard to be prescriptive about what this means it is essentially to reject the implausible. For example, We have an intuition that the PSP network consists of more than three components, if it were divided by an algorithm into three components we would expect these would in turn have identifiable subdivisions. Biological plausibility[338] has been used to choose between algorithms and set parameters. At larger scale the best I can do is to reject the implausible or impracticable (see below).

The goal of this chapter is to combined Gene set analysis (GSA) with community detection to provide data generated gene sets. GSA can be biased by very large and very small groups[100] and there are arguments that community sizes in real world networks tend to follow power law distributions (or at least heavily right skewed distributions)[230] or conform to an optimum size[243]. A reasonable place to start would be community detection methods that give 'good enough' communities on the PSP network.

An ideal community detection algorithm would therefore

- Perform well on standard benchmarks (LFR)
- Perform well on the PSP network [7]
- Have communities of 'reasonable' size
- Have a reasonable run-time

[6] leaving aside for now the question of what constitutes ground truth
[7] section 4.4.0.3

- Have an available reliable implementation
- Not have a large number of parameters

The algorithms considered were those that are available in the widely used igraph package which have been previously used to measure algorithm performance [446],[388].

4.3.0.0.1 Implementations The clustering of the PSP network used to provide gene sets to analyse the GWA study samples were performed by Dr. Colin McLean (see study design - section 5.1.1). The clustering to assess performance on benchmarks was carried out by me. I did not carry out Geodesic edge betweenness on the benchmark networks as the algorithm was already clearly impracticable and took a substantial amount of time to run on the ECDF (Edinburgh Compute and Data Facility) Linux Compute Cluster (Eddie) when performed on the PSP graph[8]. Community detection on an earlier pilot study led to the decision to limit detection algorithms to those implemented in igraph or the CDM suite that had been developed within our group. Markov clustering was carried out after the initial assessment of algorithms[9].

4.3.1 Girvan–Newman algorithm

Communities can be discovered by progressively removing links between dissimilar nodes. The Girvan–Newman algorithm[146] calculates the betweenness of all edges in the network. The edge with highest betweenness is then removed and the process is repeated. Over time the removal of these 'bottleneck' edges breaks the network into highly connected subgraphs (divisive hierarchical community detection). The betweenness needs to be recalculated at each stage and the optimum partition (the level in the dendrogram to stop dividing) is that which maximises modularity. This algorithm has been used to to find communities in the MASC complex[340].

An implementation is provided in igraph but does not scale to networks of the scale of the PSP as it has complexity of $O(m^2n)$[298]. Even an efficient parallelised implementation took a prohibitively long time; for a network with 1108 nodes and 4691 edges on a single core the algorithm took 9690 minutes (approximately 6.7 days)[266].

[8] `https://www.ed.ac.uk/information-services/research-support/research-computing/ecdf/high-performance-computing`

[9] they were not either implemented in igraph or developed 'in house'. Markov clustering was assessed as it was used in ref [144]

4.3.2 Spectral clustering

Graphs can be divided into communities using the spectrum of the eigenvalues and eigenvectors of the matrices that represent them[292]. Spectral methods have been described for modularity maximisation, statistical inference and normalised graph cut partitioning and have been shown to be equivalent [10][295].

Community allocation is assigned on the basis of the sign of the principal eigenvector of the modularity matrix (or the second eigenvector of the Laplacian). Spectral community detection with a fine tuning step was implemented using the CDM suite provided by the first author of [266]. The theory underlying spectral clustering can be found in [295] and supplemental section B.4.1. Details of the implementation and fine tuning step can be found in McLean et al.[266]. The C++ implementation is provided in the supplementary material.

4.3.3 Louvain

The Louvain method,the university of the authors of the original paper. is a hierarchical agglomerative community detection method using modularity maximisation as the objective function. It functions very well at extreme scale and is fast[47].

The algorithm works by first assigning nodes into communities. Then those communities are in turn combined into communities composed of communities. The process stops when maximal modularity is achieved. It alternatively referred to as multilevel (see for example[446]).

4.3.3.1 Methods

Louvain community detection was carried out using igraph Version 1.2.4.2 using the `cluster_louvain` command. The Louvain implementation requires no parameters to be supplied other than an optional weights argument for weighted networks and this simplifies study design. The igraph implementation returns communities discovered at the level of hierarchy that maximises the modularity. The communities objects returned by the `cluster_louvain` method were stored as vertex attributes in a .gml file provided by Dr C.McLean and converted into a .gmt file (GSA input file) using a custom R script[11]. This procedure was used for all of the algorithms that were used to provide gene sets for GSA.

[10] at least in the case of two groups

[11] `RProjects/paper_xls_output/R/magma_synapse/make_synapse_gmt.R`

4.3.4 Infomap

Infomap, described by Rosvall, is one of many algorithms conceptually based on the properties of a random walk across as network[362]. In such a random walk the walker will tend to remain in communities once in them (see also walktrap section 4.3.6). Infomap uses information theory to find an optimal community structure and aims to minimise the descriptive length of an encoding of the community struture. The hypothetical random walk is given a binary encoding with movements in and out of communities given specific codes and movement within communities given codes. These do not have to be unique because the information already provided about the location in the walk can identify where the walk is in terms of a node or community. The goal is then to encode community entry and exit in such a way to minimise the descriptive length. The traversal of the actual community structure will be best described with a labelling that corresponds as close as possible to the community structure but there is also a cost to intra group movement and a tendency to produce shorter labels. Infomap was implemented using igraph for R with default arguments (n.b trials=10) as increasing the number of trials only led to a minimal increase in the modularity.

4.3.5 Spinglass

Community detection can be formulated as the solution of an infinite Potts spin glass where each node is coupled and the spin of each node in the ground state corresponds to the community. The Hamiltonian is minimised using simulated annealing[220]. The energy of the ground state can be interpreted as a quality function.

The parameter γ which controls the balance between present and missing edges can also control the size of communities and their nesting eg with communities in higher gamma domains being nested within those in lower (see Reichardt[355]). With γ=1 the solution should be similar to maximising the modularity[116].

There are a number of tunable parameters for the implementation of the spinglass algorithm implementation in igraph in addition to γ there are: spins (default 25 - upper limit for number of communities); starting, stopping and cooling temperatures and update rules.

4.3.6 Walktrap

The walk trap algorithm is an agglomerative community detection algorithm uses the properties of random walks to define distances between nodes and discover communities. At the stationary distribution the probability of being on a node is proportional to its degree. At earlier times nodes within communities are more likely to be visited by a random walk originating within that community. Short random walks (the default in igraph is four) should therefore stay in the community. The algorithm determines the random walk probability density scaled by the node degree[341]. The The optimum level to stop merging the dendrogram is modularity by default in igraph. It has computational complexity of $O(n^3)$[92] or $O(n^2 logn)$ on a sparse network[446].

4.3.7 Fast greedy

The fast greedy algorithm (abbreviated as fc) implemented in igraph is the Clauset Newman Moore (CNM) algorithm[12] and is a hierarchical agglomerative algorithm. Communities are merged if the merge results in the greatest gain in modularity[76]. The algorithm starts with each node in a single community and achieves efficiency by updating a matrix of ΔQ_{ij} the change in modularity when communities i and j are merged. Its worst case running time is $O(md\ \log n)$ where m and n are the number of vertices and edges as before and d is the depth of the dendrogram representing the merges of communities.

4.3.8 Label propogation

This heuristic algorithm runs in near linear time but has no guarantee that a similar clustering will be found (to a greater extent than other algorithms). In the implementation in igraph for R and python very disparate values of the modularity are found and often the modularity is very low[347]. Each time the function is called in igraph for Python the modularity of the division is usually close to 0 although some are higher. With a number of iterations and keeping the partition with the highest modularity and acceptable division is achieved.

The label propagation method is somewhat similar to algorithms such as k nearest neighbours and k means clustering. At the start each node belongs to a community consisting entirely of that node. In each round of the algorithm the label changes to

[12] https://igraph.org/r/doc/cluster_fast_greedy.html

that of the majority of its neighbours. Ties are broken.

4.3.9 Markov Cluster Algorithm

The Markov Cluster Algorithm (MCL) is also conceptually based on random walks over the network with alternating patterns of expansion and contraction. After loops have been added the network is rendered as a stochastic matrix where the transition probability between two nodes of length n is the stochastic matrix raised to the nth power. Vertices in similar communities should have a high probability of transitioning to one another and this effect is made more pronounced by contraction which does something. Markov clustering was implemented using the default command line arguments of Version 14.137 running on Ubuntu 16.04 (from `https://micans.org/mcl/`)[113].

4.3.10 Leading eigenvector community detection - LEC

A simple implementation of the leading eigenvector algorithm of Newman is provided as part of igraph. The principal difference between this and the spectral method implemented by McLean et al. is that there is no fine tuning step (there are also a number of differences in the calculation of the eigenvalue spectrum (see [266]).

4.3.11 Stochastic block model

The stochastic block model is a generative model of network and community structure that can incorporate both community structure and other features such as core periphery structure. It assumes that the network is generated from a set of group assignments g and a matrix of probabilities P where $P_{i,j}$ is the probability of an edge between two nodes if the nodes are in group i and j.

4.4 Lancichinetti–Fortunato–Radicchi (LFR) benchmark

The Lancichinetti–Fortunato–Radicchi (LFR) benchmark[230] generates a network with a degree and community size distribution parameterised as a power law distribution[229] to accord with their observed prevalence in real world networks (see [346],[320]). It is intended to be more challenging than the Girvan-Newman (GN) model which uses four communities[146]. A mixing parameter μ, the proportion of

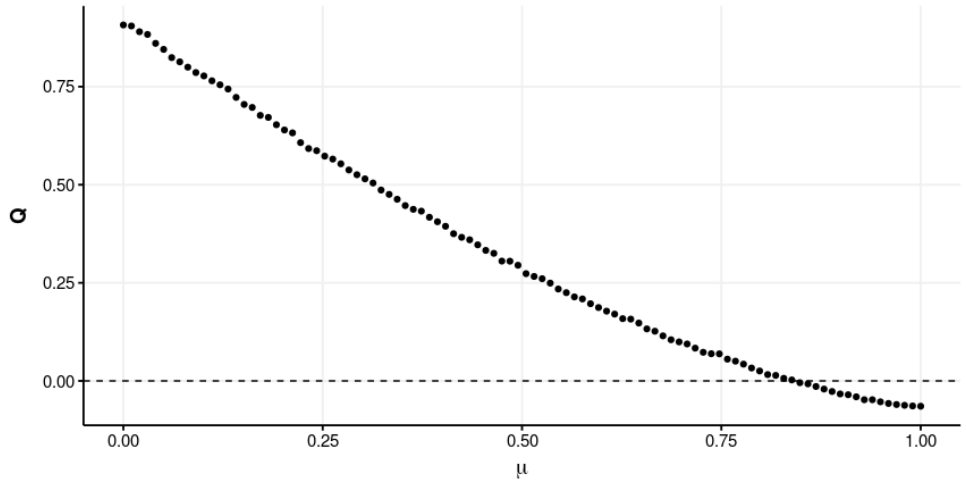


Figure 4.2: Plot of modularity with parameter *mu* in LFR benchmark. Using LFR 3457 40617 benchmark

Plot RProjects/fortunato/R/plot/new_plot_qandmu.R Calculations /home/grant/Projects_/Python/src/Qandmu.ipynb

edges found between communities, controls the amount of noise in the network and hence the difficulty of the task for the community detection algorithm. With increasing values of μ the modularity of the ground truth partition decreases (figure 4.2).

4.4.0.1 Methods

The LFR benchmark networks were generated using the Python package NetworkX (version 2.4)[165]. The NetworkX implementation is almost identical to the authors implementation[13]. In addition to the power law exponent of the degree distribution either the minimum or average degree can be provided as arguments to generate the desired degree distribution[14].

The algorithm does not converge for all values of parameters and the distribution of values of μ obtained (figure 3 in[230]) is bi-modal distribution with peaks $P(\mu)$ around 0.1 and 0.5[15].

Using the default parameter values of the authors' implementation resulted in a network where it was difficult to discriminate between algorithms even at $\mu = 0.5$ as performance was almost universally good.

[13] the differences are that rewiring is done at the end which is held to be equivalent and the community allocation is treated as an integer

[14] https://networkx.org/documentation/networkx-2.1/reference/algorithms/generated/networkx.algorithms.community.community_generators.LFR_benchmark_graph.html

[15] The authors of the NetworkX method note that 'a somewhat narrower parameter set is likely to successfully produce a graph' which has also been my experience

Setting the minimum and maximum degree using the matching values of the PSP and degree exponent of 2.51 (section 3.9.1) produces a network with roughly equal vertices and edges (V:3,457, E:4,736), a mean degree of 2.74, and maximum degree 119 (degree assortativity -0.09, global transitivity 0.012). This is clearly a poor model for the PSP.

If instead, using the same degree exponent, one sets the mean degree parameter equal to the PSP mean (rather than setting minimum or maximum degree) the network is much more like the PSP although there are no nodes of degree less than seven . These make up 1,445 vertices of the PSP graph equivalent to 41.7% of the vertices. This may be because the power law distribution does not hold in the low degree region of the degree distribution. This LFR benchmark network has 3,457 nodes and 40,617 edges. The modularity was 0.431 for ground truth community allocations (μ=0.375). The graph is available in .gml format[16]. I will refer to this at the 'PSP like' network.

I have used an LFR benchmark close to the PSP as algorithms performance is known to vary depending on parameters such as network size and average degree[426].

The Potts model[354] and modularity optimisation by simulated annealing[163] were assessed in the paper introducing the LFR benchmark[230]. The Louvain and Infomap algorithms were shown to perform best out of a number of algorithms on the LFR benchmark in a subsequent paper[229]. A variety of spectral method was assessed on this occasion, the method of Donnetti and Munoz, which assigns nodes to clusters based on their proximity in an embedding in Euclidean space using a few of the eigenvectors of the Laplacian matrix[111]. The networks produced using the parameters from this paper[229] are quite dissimilar to the PSP[17]. The LFR benchmark has limitations, in addition to the difficulty in generated particular networks it is known to produce similar sized communities which differ from real world networks[316],[7].

Networks generated using the LFR benchmark have been reported to have low transitivity and zero degree correlation in common with many other generative models and so do not completely capture the structure of complex networks [316],[315]. However using the 'PSP like' LFR network it is interesting to note the degree assortativity is r=-0.16 (PSP r=-0.18) table 3.10) and the transitivity C=0.0415 (PSP=0.067 section B.1.3.1) and C_{WS} =0.0810[18].

Given these limitations of the LFR benchmark, I have also used the performance of algorithms on the PSP, as measured by modularity and plausibility of group size, to

[16] `git@github.com:14760/ipynotebooks_phd.git`

[17] the average degree is 20, the maximum degree 50, the exponent of the degree distribution is -2 and that of the community size distribution is -1.'[229]

[18] `/home/grant/Projects_/Python/venvs/Create_LFR_similar_to_PSP.ipynb`

determine the most appropriate algorithms to use.

4.4.0.1.1 Measuring performance Normalised mutual information (NMI), derived from information theory, has been used as a measure of the quality of partitions of a network[92] and lies in the range 0-1. The algorithms were assessed on the LFR benchmark by comparing the detected communities to the ground truth partition using the `compare_communities` method in igraph for Python. The mutual information, adjusted for the expected value of the mutual information (adjusted NMI)[414], was calculated using scikit-learn[1]. Normalised mutual information may not accurately reflect performance when there are a large number of communities[13],[134].

Normalised and adjusted normalised mutual information were used to compare the similarity of the communities discovered in the PSP. Here the communities found by different algorithms are compared against one another rather than ground truth.

4.4.0.2 LFR Performance

The performance of the igraph community detection algorithms, Markov Cluster algorithm and the Spectral Clustering Algorithm (CDM)[266] on the LFR (PSP like) benchmark graph are shown in table 4.1. A sizeable difference is seen between the NMI and Adjusted NMI, particularly for the Markov Clustering partition (NMI=0.653, Adj NMI=0.295), supporting the utility of including this measure[414].

The Markov Cluster algorithm method (adjusted NMI (aNMI)=0.127; Q=0.127) and leading eigenvector method (aNMI=0.134; Q=0.187) perform poorly as does the fast greedy algorithm despite good performance on modularity (aNMI=0.457 ; Q =0.368). Label propagation finds a good quality partition (aNMI= 0.8; Q=0.416).

Table 4.2 shows the distribution of community sizes detected. The Markov Cluster algorithm generates a large number of communities ($n = 1,621$). We expect divisive algorithms to give rise to some small communities (as agglomerative ones do large ones) but only 42.4% of nodes are in communities larger than fifteen (the default minimum gene set size for GSEA) compared with 97.98% for the spectral CDM algorithm and 99.65% for label propagation (table 4.3). The fast greedy algorithm discovers few communities ($n = 8$) which are large (median size 407 vs ground truth 57) as does the leading eigenvector (LEC) method (median size=593). Walktrap, Infomap, and label propagation have median group sizes close to that of ground truth. The spectral algorithm has low median community size (8) but once the numerous ($n = 34$) but small communities smaller than GSEA default have been removed is a good match to

Method	Modularity	NMI	Adjusted NMI
Ground truth	0.431	1.000	1.000
Infomap	0.432	0.974	0.971
Spinglass gamma1	0.429	0.914	0.887
Spinglass gamma2	0.433	0.931	0.882
Louvain	0.432	0.913	0.870
Walktrap	0.423	0.893	0.870
Label prop	0.416	0.835	0.800
Spinglass gamma5	0.379	0.810	0.796
Spectral	0.392	0.782	0.762
Fast greedy	0.368	0.564	0.457
Markov clustering	0.127	0.653	0.295
Leading eigenvector method	0.187	0.190	0.134

Table 4.1: Performance of algorithms measured by modularity, Normalised Mutual Information (NMI) and adjusted NMI on the LFR Benchmark using the 'PSP' like artifical network (V=3,457, E= 40,6017)

RProjects/python_benchmarks/R/format_LFR_Q_removesbm.R

ground truth (72 vs ground truth 57 see figure 4.3; excluding 2.02% of nodes). The ground truth has a minimum community size of 16 which may favour agglomerative algorithms.

The comparison of community sizes to ground truth with small groups ($<$15 members) removed is shown in figure 4.3. Some of the partitions discovered have slightly higher modularity than the ground truth[134]. Both Louvain and Spectral have similar community size to ground truth in figure 4.3 although Louvain is a little large (median 110 vs gt 57) as is Spinglass with $\gamma = 5$ (min size=98).

Infomap has the best (0.971) performance on aNMI, the rest of the algorithms have acceptable performance between 0.762 (Spectral) and 0.887 (Spinglass gamma1). Fast greedy, MCL and LEC perform too poorly to be considered further.

	n	Min.	1Q.	Median	Mean	3Q.	Max.
Ground truth	36	16	37.75	57.0	96.0	107.75	558
Markov Cluster	1621	1	1.00	1.0	2.1	1.00	148
Spectral	66	1	1.00	8.5	52.4	68.50	518
Louvain	24	18	60.25	110.0	144.0	204.25	573
Infomap	39	2	29.00	50.0	88.6	96.50	568
Label prop	26	12	40.25	64.5	133.0	177.25	793
Fast greedy	8	205	371.00	407.0	432.1	502.25	633
Spinglass gamma 1	24	18	79.50	115.0	144.0	180.50	564
Spinglass gamma 2	25	60	76.00	99.0	138.3	176.00	481
Spinglass gamma 5	25	98	113.00	129.0	138.3	156.00	208
LEC (Leading eigenvector)	5	287	448.00	593.0	691.4	905.00	1224
Walktrap	33	15	35.00	55.0	104.8	139.00	633

Table 4.2: Group sizes compared with ground truth LFR 3457 PSP like benchmark

RProjects/python_benchmarks/R/make_table_gt_lfr_3457_removeSBM.R

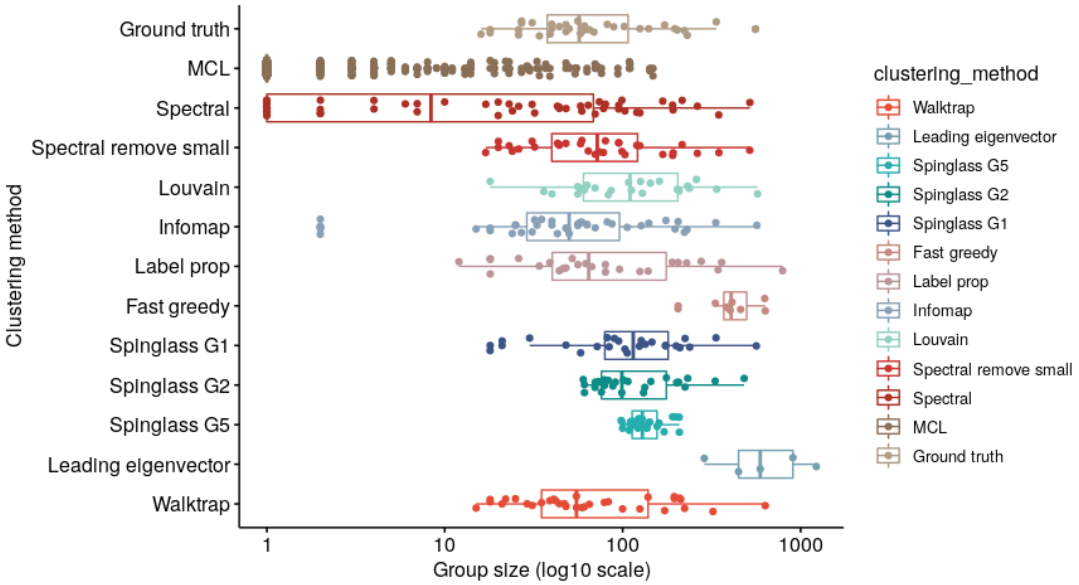


Figure 4.3: Boxplot of sizes of communities detected for various clustering methods on a 3457 node 40617 edge benchmark LFR benchmark graph. Label propagation not included as was not one of the clusterings supplied to me

RProjects/python_benchmarks/R/group_size_ggpubr/lfr_3457_nosizefilter.R

	Min	1Q	Median	Mean	3Q	Max	n	coverage	n comm	n max comm
Ground truth	16	37.75	57	96.03	107.75	558	3457	100.00	36	36
MCL	18	25.00	39	54.33	71.00	148	1467	42.44	27	1621
Spectral	17	40.25	72	105.84	121.25	518	3387	97.98	32	66
Louvain	18	60.25	110	144.04	204.25	573	3457	100.00	24	24
Infomap	15	32.75	56	95.86	111.25	568	3451	99.83	36	39
Label prop	18	44.00	67	137.80	190.00	793	3445	99.65	25	26
Fast greedy	205	371.00	407	432.12	502.25	633	3457	100.00	8	8
SG1	18	79.50	115	144.04	180.50	564	3457	100.00	24	24
SG2	60	76.00	99	138.28	176.00	481	3457	100.00	25	25
SG5	98	113.00	129	138.28	156.00	208	3457	100.00	25	25
LEC	287	448.00	593	691.40	905.00	1224	3457	100.00	5	5
Walktrap	15	35.00	55	104.76	139.00	633	3457	100.00	33	33

Table 4.3: Group sizes found with community detection methods compared with ground truth community size in the LFR 3457 ‘PSP like benchmark’ filtered with communities smaller than fifteen removed. n is number of genes in communities with over fifteen members, coverage is the percentage of PSP genes in communities with over fifteen members, n comm is number of communities and n max comm is the number of communities if no size filtering is used. Summary statistics (1Q=1st quartile, median, mean, 3q=3rd quartile) refer to the size of the communities discovered.

4.4.0.3 PSP Performance

Examination of the performance of the community detection algorithms on the PSP network shows that the maximum modularity (Q=0.34) is found in the Louvain algorithm and Spinglass ($\gamma = 1$).

Five algorithms (Spinglass ($\gamma = 1$ and 2), Louvain, Spectral and fast greedy) have a modularity $>= 0.3$ (table 4.4). The group sizes and modularity of the above algorithms on the PSP graph are shown in table 4.4 and figure 4.5. Fast greedy finds eight communities with a high median degree (407) would be undesirable for GSA. Spinglass ($\gamma = 1$) has eighteen groups, four of these are very large (figure 4.4) and less suitable for GSA. Spin glass ($\gamma = 1$) has median group size close to a 'good' level of 100 members[243]. Spectral CDM has a lower median group size but this is mostly due to small numbers of very small groups. Table 4.5 and figure 4.6 shows a summary of the group sizes the effect of filtering out small groups that might have biased assessments using GSA[100] using fifteen the default lower bound for GSEA. Only 2.5% of genes are excluded from the spectral clustering method but the median is much closer to 100 now, there is only one large group.

Table 4.4 shows that the Geodesic algorithm and walktrap have an improbably large number of communities (2,566 and 562) and are not considered further.

The Infomap algorithm is popular, is fast, and is reported to perform well the LFR benchmark[230], [289] and performed best in section 4.4.0.2. On the PSP network it discovers 171 communities ($Q = 0.28$) with a maximum size of 1,228 (35.5% of the PSP). A recent comparisons of algorithms implemented in igraph[444],[446] showed Infomap's performance deteriorated with higher noise on the LFR benchmark. Algorithms that performed well in section 4.4.0.2) gave rise to communities of unrealistic sizes with on the PSP network. Importantly it is difficult to reproduce closely the degree structure of the PSP network using the LFR benchmark. A resolution limit exists for community detection algorithms[134],[289] and is demonstrated for Infomap in section 4.4.1.

The walk trap algorithm and Geodesic algorithm produce many communities with low mean community size and have suboptimal modularity. The igraph implementation of spectral clustering (LEC) divides the network into three groups all larger than 1000 nodes.

Louvain clustering discovers 14 communities in the PSP with modularity 0.339.

The largest community is 539 genes in size (module 8) and the minimum commu-

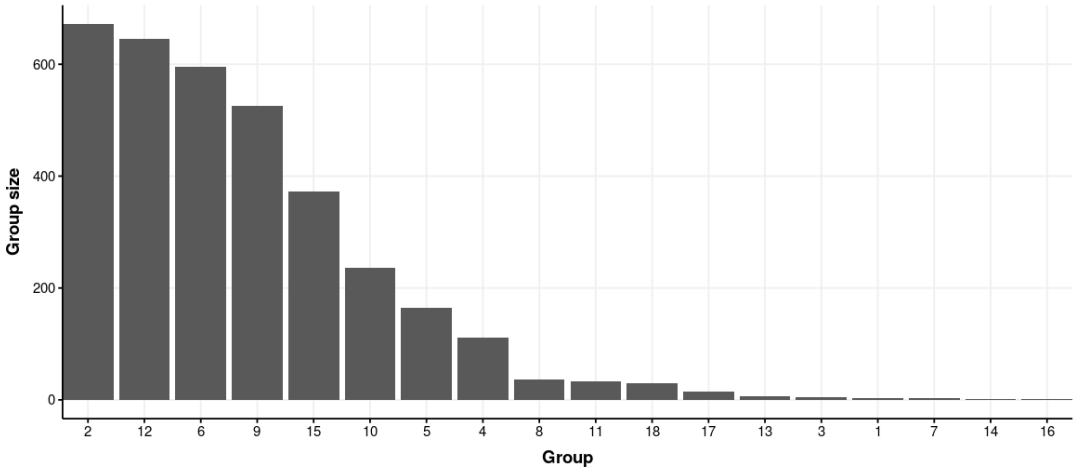


Figure 4.4: Group size Spinglass γ=1

nity size is 21 genes (module 12). The median module size is 190.5 and the mean is 246.9. The distribution of community sizes is shown in the boxplot in figure 4.12.

The algorithms that perform adequately well on all measures are the spinglass algorithms (although $\gamma = 1$ and 5 had problematic group sizes), Louvain and spectral.

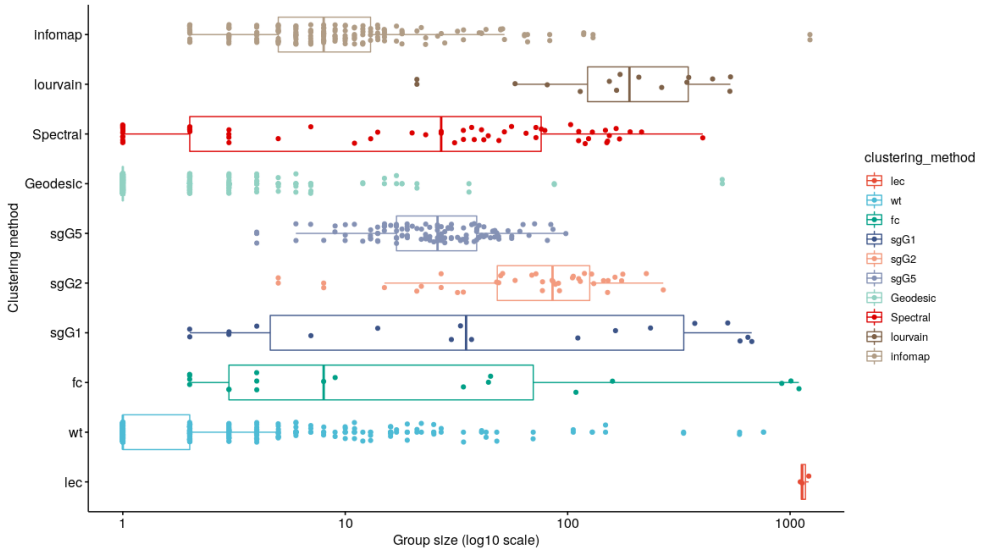
The adjusted NMI suggest the Spectral is most similar to Spinglass, Infomap and louvain. Figure 4.7 shows a dendrogram using the adjusted NMI as a similarity measure. The dendrogram and distances are calculated in R using `dist` ("Euclidean") and `hclust` (method="complete"). The three poorly performing algorithms can be seen clearly together.

Louvain is most similar to spinglass Gamma1 and Spectral to Spinglass with gamma 2 and 5. Walktrap and infomap both conceptually related to random walks are close together. Table 4.6 shows the Normalised Mutual Information between community allocation on the PSP for the graphs generated for the study (those carried out by Colin McLean). The NMI for lec (leading eigenvector method in igraph) and fast greedy are poor. The NMI adjusted for chance is shown in table 4.7.

Spinglass with γ=1 gave rise to large groups and with $\gamma = 5$ the groups seemed particularly large and similar in size, all were large compared to ground truth on LFR[19]. The number of potential parameters also mitigated against the choice of spinglass.

These results accord with recent assessments of community detection methods[446],[289],[134]. I chose to use the spectral method as it had performed well across all measures, had good performance and differential enrichment on PSP like networks[266] and we had

[19] from the point of view of study design large being worse than numerous small which can be excluded



Rplot

Figure 4.5: Boxplot of the sizes of communities detected in the post synaptic proteome (PSP) network using different community detection methods. The boxplot shows all communities and is not filtered by community size. Scale of response (group size) scaled by log10.

RProjects/fortunato/R/calculate_PSP_Group_sizes_boxplot.R

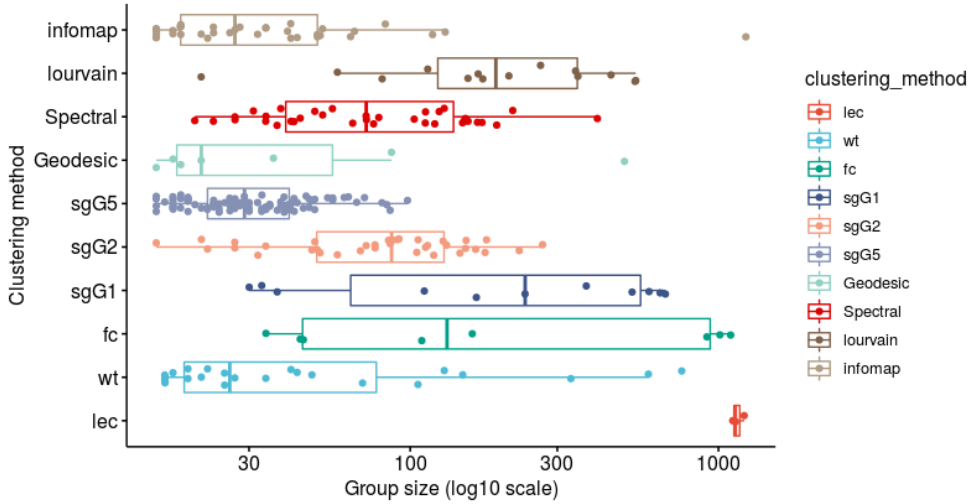


Figure 4.6: Boxplot of the sizes of communities detected in the post synaptic proteome network using different community detection methods. The boxplot shows all communities and is filtered by community size. Only communities greater than or equal to fifteen members. Scale of response (group size) scaled by log10. Spectral clustering with small groups remmoved has a median close to the ideal of size 100

RProjects/fortunato/R/calculate_PSP_Group_sizes_boxplot_add_sizefilter.R

method	Q	n	min	max	median	mean
Leading eigenvector (LEC)	0.25	3	1111	1212	1134.0	1152.3
walktrap	0.26	562	1	760	1.0	6.2
louvain	0.34	14	21	539	190.5	246.9
Spinglass gamma1	0.34	18	2	672	35.0	192.1
Spinglass gamma2	0.33	38	5	269	85.5	91.0
Spinglass gamma5	0.26	118	4	98	26.0	29.3
infomap	0.28	171	2	1228	8.0	20.2
spectral	0.30	65	1	405	27.0	53.2
geodesic	0.10	2566	1	496	1.0	1.3
fast greedy (fc)	0.31	19	2	1010	8	181.9

Table 4.4: Modularity (Q) and group sizes PSP. Louvain, spinglass and spectral have the best modularity. The fast greedy algorithm has a good modularity but contains an implausibly large group $n = 1010$

expertise in the implementation of the algorithm in our group. In addition the pattern of group size (figure 4.12) was similar to a power law distribution.

The Louvain algorithm was very attractive with a superior modularity to the spectral algorithm on the PSP (0.34 vs 0.30), but the communities were larger than one would prefer with Louvain (double median in LFR, median 190 PSP) and therefore slightly less suited to GSA (see [100] for the effects of large groups or small groups for GSA).

4.4.0.3.1 Other performance measures Conductance (supplemental section B.4.2) is a measure of the cohesiveness of communities[134]. The lower the value the more isolated the community is from the rest of the graph and the measure lies in the range 0-1.

However the conductance is correlated with group size. We can however identify particularly cohesive communities within a partition. For the smaller spectral clustering groups median conductance was 0.7188. As communities get larger the conductance is lower (figure 4.8)[20] and the correlation is moderate and statistically

[20] `RProjects/fortunato/R/calculate_fortunato_measures_name_input.R`

method	min	mean	median	IQR	max	n_c	n_{nodes}	coverage
Leading Eigenvector (LEC)	1111	1152.3	1134	50.5	1212	3	3457	1.000
walktrap	16	106.6	26	60.5	760	24	2559	0.740
Fast greedy fc	34	426.8	134	896.2	1095	8	3414	0.988
Spinglass gamma1	30	311.1	236	486.0	672	11	3422	0.990
Spinglass gamma2	15	95.7	87	79.0	269	36	3444	0.996
Spinglass gamma5	15	33.8	29	18.5	98	95	3215	0.930
Geodesic	15	98.6	21	44.0	496	7	690	0.200
Spectral	20	96.3	72	99.0	405	35	3369	0.975
louvain	21	246.9	190	224.5	539	14	3457	1.000
infomap	15	68.7	27	32.0	1228	37	2541	0.735

Table 4.5: Group size of communities detected in the PSP for different algorithms excluding communities with less than 15 members. n_c is number of communities

RProjects/fortunato/R/summary_table/calculate_PSP_Group_sizes_boxplot_add_sizefilter_summary_table.R

significant when calculated using Spearmans method in R $\rho = -0.62$ S=11,577 $p = 6.765 \times 10^{-15}$.

The same can be seen for the larger Louvain groups (figure 4.9 plotting group 7 in red). Median conductance=0.58; $\rho = -0.736$, S=790, p=0.0038.

The community statistics for the communities detected in the PSP using the spectral clustering method are shown in table B.21 for Louvain see table B.24 and infomap table B.20.

4.4.1 Benchmark: Stochastic block model Modularity and Group size

Before turning to testing the spectral algorithm with GSA it is worth looking at the difference in performance on the LFR benchmark and PSP. To investigate this I analysed their performance on a block stochastic network model with two planted communities and increasing levels of background noise. Here I present the results of the infomap and Louvain method (infomap as it performed well in the benchmark and Louvain as it had relatively good performance on the PSP and has a short running time). There are resolution limits to community detection[289][21] and one can see how these algorithms differ as they encounter increasing noise.

[21] p512

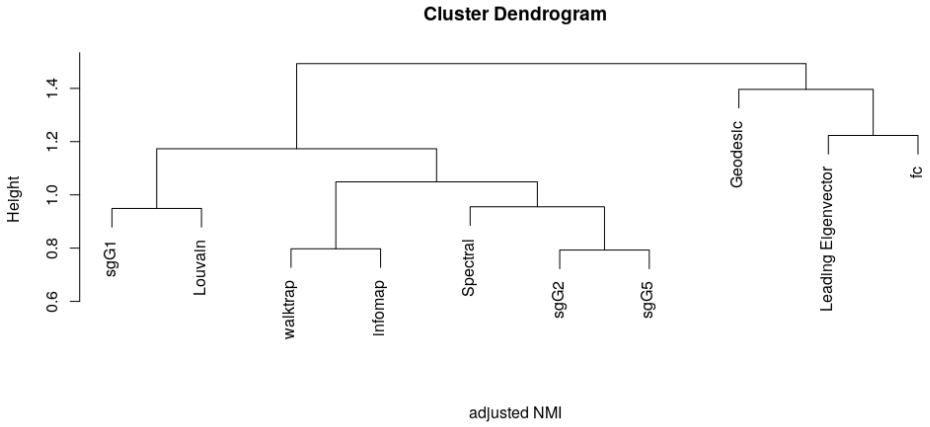


Figure 4.7: Dendrogram of distances between clustering modalities using adjusted nmi

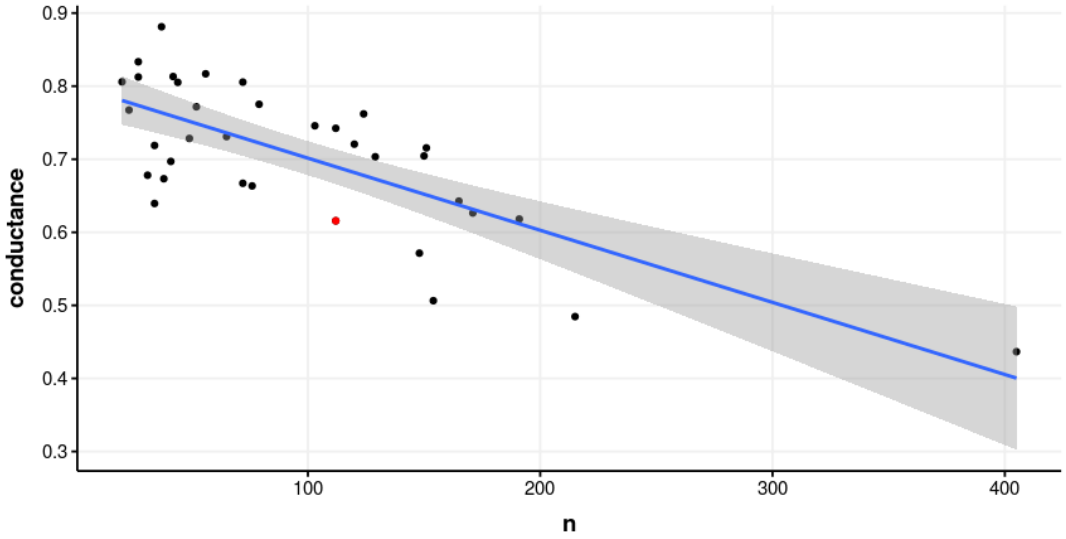


Figure 4.8: Relationship between conductance and group size. Spectral clustering method. Communities of size less than 15 removed. Linear regression line shown to illustrate trend. Red point is community 5 which has a low conductance (cohesive) for its size.

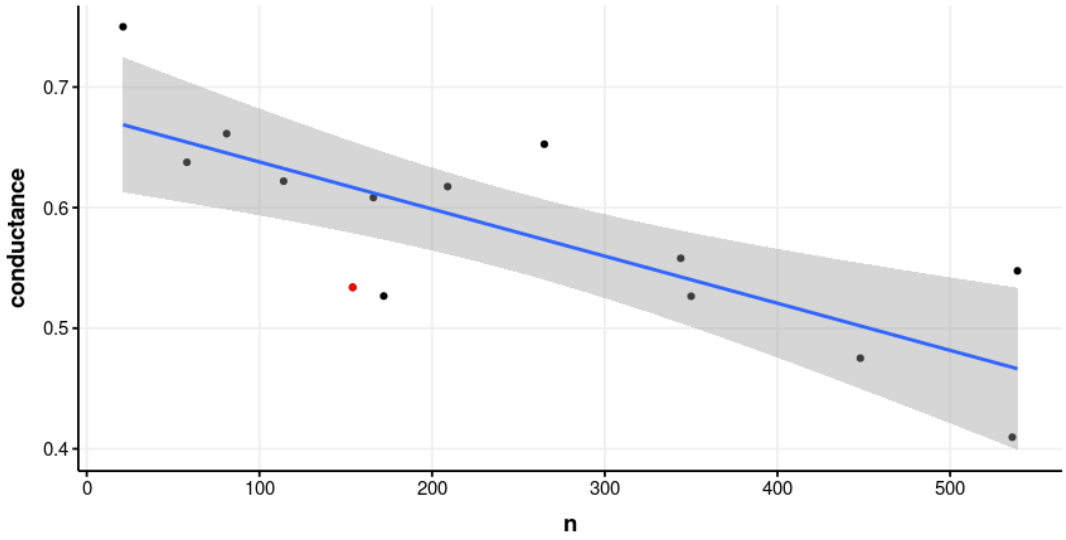


Figure 4.9: Relationship between conductance and group size. Louvain clustering method. Communities of size less than 15 removed. Regression line shown to illustrate trend. Red point is community 7 which has a low conductance (cohesive) for its size.

RProjects/fortunato/R/plot/conductance_groupsize_lou.R

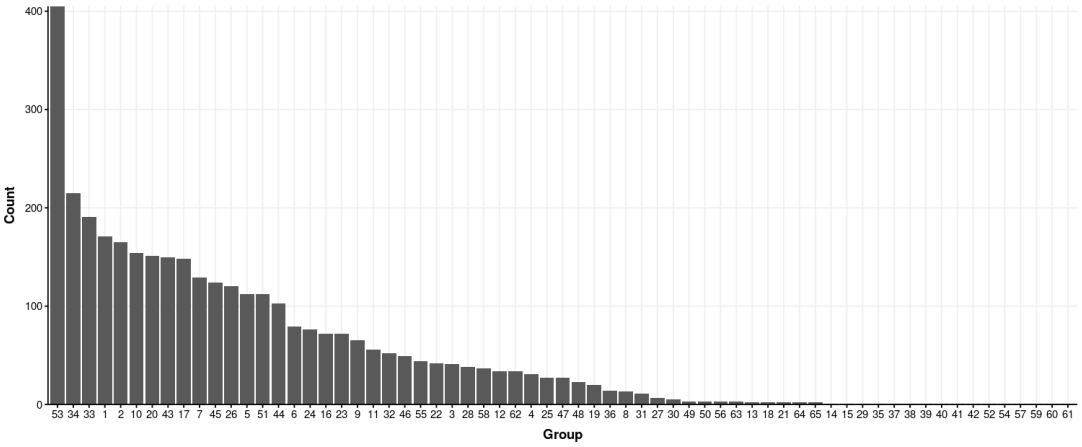


Figure 4.10: Distribution of group sizes spectral clustering

RProjects/fortunato/R/plot/barplot_spectral.R

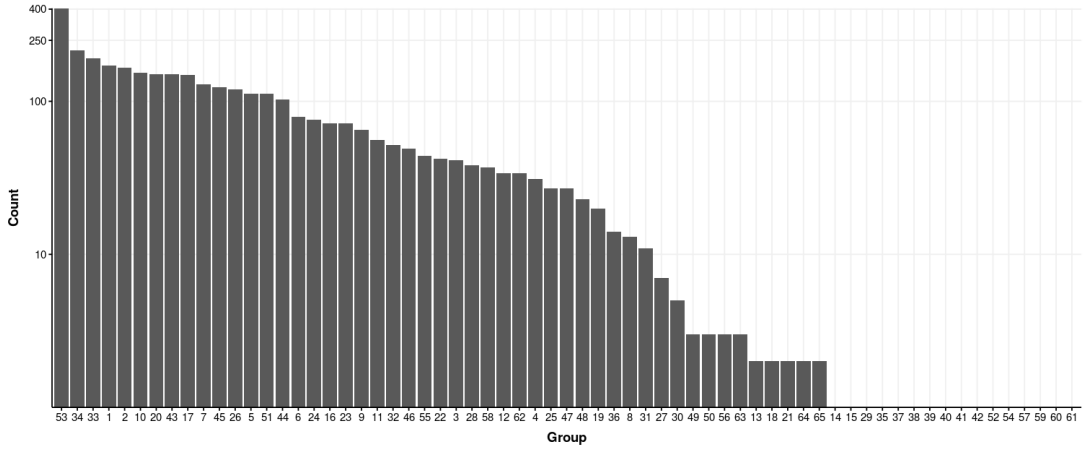


Figure 4.11: Distribution of group sizes spectral clustering. Log 10 scaled y axis

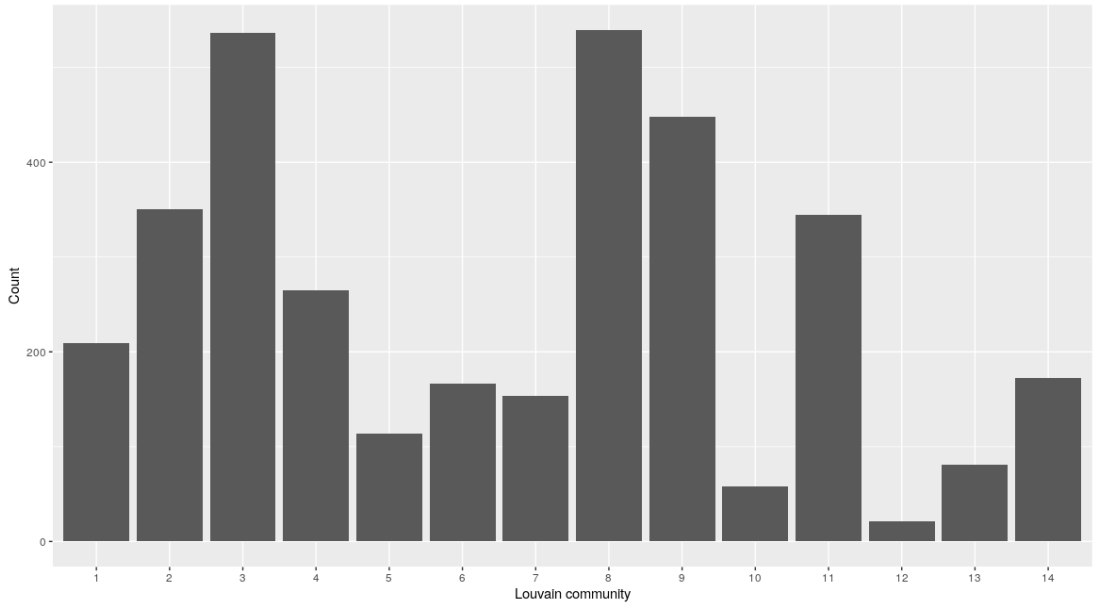


Figure 4.12: Barplot of the size of communities from the PSP discovered using the Louvain clustering method implemented in igraph

code RProjects/graph2community/R/plots/louvain_size.R

	LEC	WT	FG	sgG1	sgG2	sgG5	Spectral	Geodesic	Louvain	infomap
LEC	1.000	0.172	0.185	0.160	0.132	0.128	0.191	0.235	0.165	0.145
WT	0.172	1.000	0.291	0.366	0.467	0.520	0.456	0.661	0.399	0.631
FG	0.185	0.291	1.000	0.327	0.268	0.256	0.280	0.341	0.243	0.324
sgG1	0.160	0.366	0.327	1.000	0.400	0.357	0.340	0.422	0.366	0.417
sgG2	0.132	0.467	0.268	0.400	1.000	0.595	0.443	0.578	0.423	0.537
sgG5	0.128	0.520	0.256	0.357	0.595	1.000	0.482	0.689	0.412	0.620
Spectral	0.191	0.456	0.280	0.340	0.443	0.482	1.000	0.587	0.356	0.494
Geodesic	0.235	0.661	0.341	0.422	0.578	0.689	0.587	1.000	0.462	0.663
Louvain	0.165	0.399	0.243	0.366	0.423	0.412	0.356	0.462	1.000	0.445
infomap	0.145	0.631	0.324	0.417	0.537	0.620	0.494	0.663	0.445	1.000

Table 4.6: Comparison of community detection methods using the PSP. Pairwise Normalised Mutual Information (NMI) comparing communities found on PSP. LEC=Leading eigenvector method, wt = walktrap, fg = fast greedy CMN, sgG1-5 spin glass with Gamma 1,2 or 5, Geodesic - geodesic edge betweenness, Louvain - Louvain (multilevel)

I have calculated the normalised mutual information (NMI - not adjusted for chance) for the communities detected compared to ground truth. The test graph, generated using `sample_sbm` in igraph for R has 1000 vertices and a preference matrix where the probability of an edge is 0.1 if the nodes are in the same group and the chance of an edge if the nodes are not in the same group (background rate) varies between 0.001 and 0.1 in 0.001 steps.

Similar results were presented by [446] using the LFR and changing the mixing parameter but the SBM allows demonstration of the exact probability rather than the mixing parameter which I find more intuitive[22].

Figure 4.13 shows perfect performance for the infomap algorithm as measured by NMI until the background probability reaches p=0.017. The NMI then falls to zero as infomap assigns all of the nodes to a single community. By contrast with the Louvain algorithm (figure 4.14) the performance is not as good as infomap in the region where the background p is less than 0.017 but above 0.017 the Louvain algorithm continues to perform tolerably well until over 0.05. Infomap performs very well (perfectly) when there is no 'noise' but fails completely as background noise increases. We see a very similar picture in the PSP.

[22] it can also be implemented along with the clustering algorithms in igraph

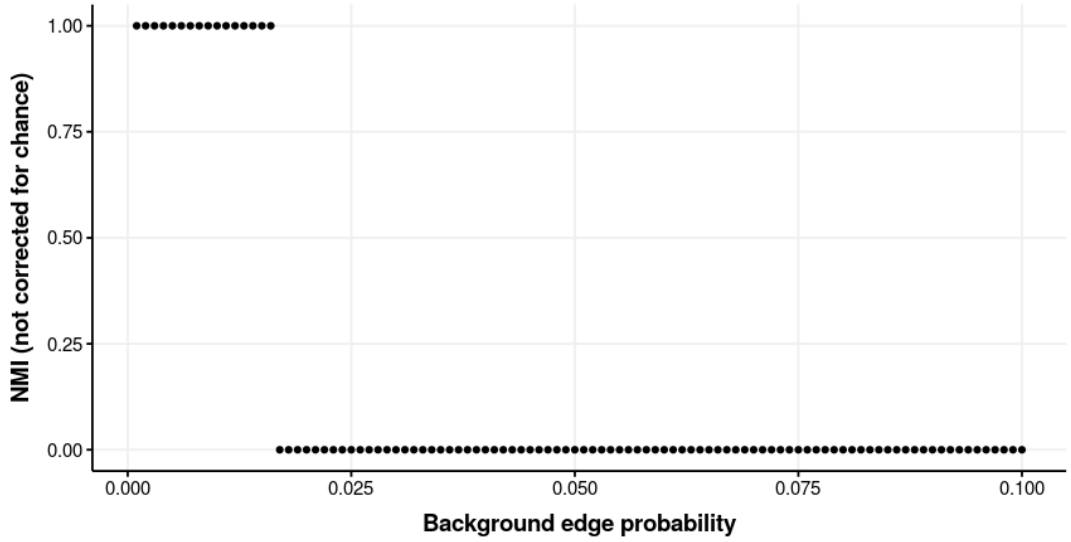


Figure 4.13: Diagnostics for infomap. Block stochastic model of two groups one 30% of nodes other 70% of nodes. Probability of edge between nodes in group p=0.1. x axis shows the probability of edge between nodes when they are in different groups (background probability). At 0.1 on x axis would be identical with probability in group. Graph shows dramatic decline in performance when background probability reaches p=0.017.

RProjects/group_sizes/R/sbm_benchmark/two_groups_infomaprds.R

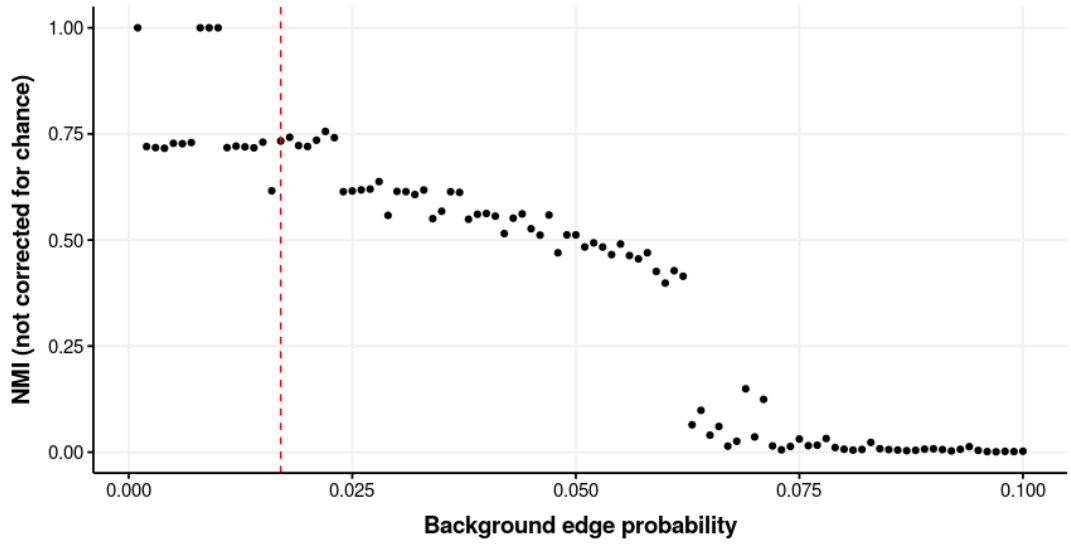


Figure 4.14: Diagnostics for Louvain clustering. Block stochastic model of two groups one 30% of nodes other 70% of nodes. Probability of edge between nodes in group p=0.1. x axis shows the probability of edge between nodes when they are in different groups (background probability). There is a gradual decline in the value of NMI ion contrast to the infomap figure 4.13. Red dashed line represents background probability p=0.017 at which infomap performance dramatically decreases. In [446] for LFR

RProjects/group_sizes/R/sbm_benchmark/two_groups_louvainrds.R

	LEC	WT	FG	sgG1	sgG2	sgG5	Geodesic	Spectral	Louvain	Infomap
LEC	1.00	0.07	0.15	0.12	0.08	0.07	0.02	0.12	0.12	0.08
WT	0.07	1.00	0.13	0.19	0.29	0.27	0.12	0.27	0.22	0.47
FG	0.15	0.13	1.00	0.28	0.18	0.14	0.04	0.19	0.19	0.20
sgG1	0.12	0.19	0.28	1.00	0.31	0.23	0.06	0.25	0.33	0.28
sgG2	0.08	0.29	0.18	0.31	1.00	0.45	0.08	0.40	0.35	0.43
sgG5	0.07	0.27	0.14	0.23	0.45	1.00	0.10	0.33	0.28	0.41
Geodesic	0.02	0.12	0.04	0.06	0.08	0.10	1.00	0.09	0.06	0.15
Spectral	0.12	0.27	0.19	0.25	0.40	0.33	0.09	1.00	0.28	0.37
Louvain	0.12	0.22	0.19	0.33	0.35	0.28	0.06	0.28	1.00	0.32
Infomap	0.08	0.47	0.20	0.28	0.43	0.41	0.15	0.37	0.32	1.00

Table 4.7: Comparison of community detection methods using the PSP. Pairwise adjusted normalised Mutual Information (NMI) comparing communities found on PSP. Adjusted NMI for PSP is shown only for the clustering methods used by Dr McLean (see 5.1.1).

/home/grant/Projects_/Python/src/adjusted_nmi.ipynb

Chapter 5

Topology based Gene Set Analysis

5.1 Topology based GSA

Having selected the spectral clustering algorithm, I sought to determine whether any of the communities were enriched for variants associated with intelligence or educational attainment.

5.1.1 Study design Topology based GSA

To test the hypothesis that there are network communities enriched for genes associated with intelligence or educational attainment, we carried out the following procedures (the overall study design is shown in figure 5.1).

The communities found by the spectral clustering algorithm were used as gene sets. The discovery and replication GWA study samples used to test the association of centrality measures with educational attainment and intelligence (sections 2.2.1 and 2.2.2) were again used.

Following the recommendations of Wilmot and Mooney, we used more than one enrichment method, utilising GSEA to provide a non-parametric enrichment analysis [393],[282] (section 1.5.1). GSA using MAGMA methods were identical to section 2.5 except using spectral communities as gene sets. Gene set enrichment analysis was carried out using the command line interface for the Java implementation of GSEA (GSEA - 3.0). To test the GSEA analysis pipeline we ensured we could replicate the findings of Hill et al[177].

Community detection was carried out by Dr Colin McLean (section B.5) and recorded as vertex properties in a `gml` file (section 2.1.1.1). Community detection was carried

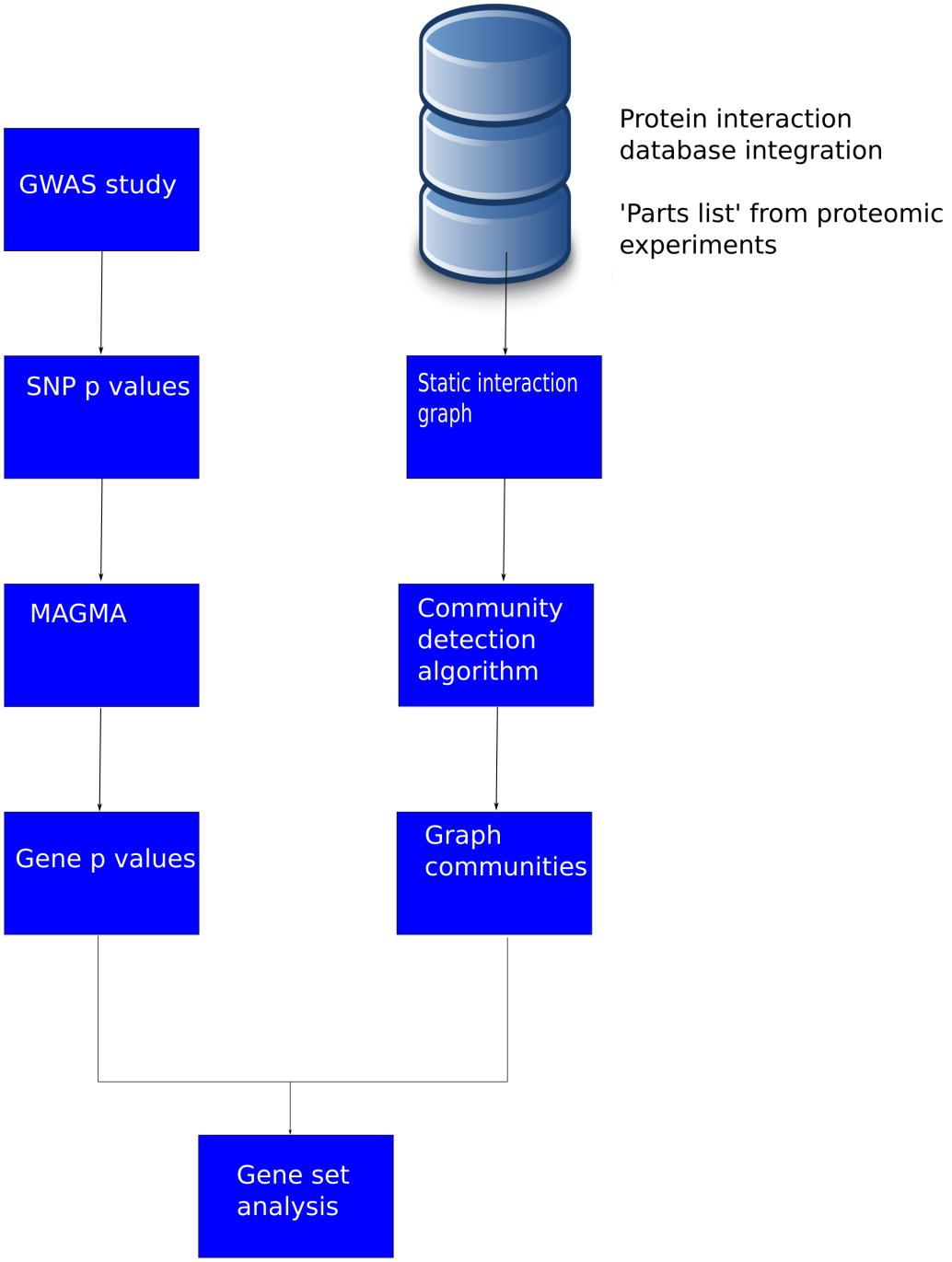


Figure 1: Study design

Figure 5.1: Study design

out using the CDS package as described in [266] using ECDF, the distributed computing facility provided by the University of Edinburgh. All other analysis was performed on an Intel® Core™ i7-7700K CPU @ 4.20GHz 8 core workstation running Ubuntu 16.04 LTS 64 bit.

Dr McLean carried out the community detection before the summary data was obtained, and was therefore blind to the results of the gene level tests (the right hand portion of figure 5.1). We used the non-overlapping communities generated by the spectral community detection algorithm as gene sets stored in `.gmt` file format (each row a gene set). We represented each protein in the community by the Entrez ID representing the gene encoding that protein.

The communities in the `gml` file were converted to `gmt` format using a custom R script. Output files and logs from both GSEA and MAGMA GSA are included in the supplementary material.

We excluded communities smaller than the default lower limit gene size for the Gene Set Enrichment Analysis software, GSEA (n=15). Divisive algorithms are known to often produce numerous small communities, and small gene sets can result in the signal being dominated by a single gene. Gene based statistics have already been calculated (section 2.3) for the discovery and replication samples. Gene set analysis (see 1.5.1) was carried out using MAGMA GSA using the gene level output as input, and GSEA was performed using MAGMA gene p values -log10 transformed, and formatted as `.rnk` files.

As we had no reason to believe which, if any, communities would be enriched, we identified a set of genes (community) to be of interest if it attained nominally significant enrichment ($p < 0.05$) on both MAGMA GSA and GSEA (see discussion section 6.3).

These gene sets were then tested in the $\text{Intelligence}_{\text{Replication}}$ and $\text{Education}_{\text{Replication}}$ samples, with correction for multiple testing using the Bonferroni method. Finally, the analysis pipeline was re-run using a different method of both gene and pathway scoring (Pascal) to test whether the findings were specific to a particular analysis platform[228]. The methods for Pascal are identical to section 2.3.6.1.

5.1.2 Gene ontology enrichment

To investigate any significant community modules we used gene ontology enrichment analysis (see section 2.1.3)[271].

We tested for over representation of Gene Ontology terms in significant communi-

ties (sets) for each of the major Gene Ontology clades, using Fisher's exact test with corrections for multiple comparisons using the FDR method. These were calculated using ToppGene[73] and PANTHER (section 2.1.4.1)[272], using the 3,457 genes in the PSP as the background set (section 2.1.4.2). The ToppGene enrichment package was also used to perform enrichment analysis of significant communities for murine and human phenotypes, again using the PSP as the background set[73].

Evolutionary constraint in significant communities was assessed using the probability of loss of function intolerance using Gnomad v2.1.1 (section3.6.11).

5.1.3 Results

5.1.3.1 Spectral community detection

Sixty-five postsynaptic graph communities were detected using the spectral clustering algorithm with fine-tuning step; the partition's modularity (Q) was 0.30, indicating a robust community architecture. The communities ranged from 1 to 405 genes (mean 53.2, median 27, standard deviation 73). We excluded those communities with less than 15 members, removing 35 communities and 88 genes (2.5% of the PSP). Thirty-five communities with a total of 3,369 genes remained with mean community size 96.1, median 72 and standard deviation 76.8 (see section 4.4.0.3 and table 4.5). The communities detected are shown in figure 5.3. The interactions of the communities are shown in figure 5.4 where each node represents a community.

5.1.3.2 Gene set enrichment results

5.1.3.2.1 Educational attainment Four of 35 communities showed greater than nominally (p <0.05) significant enrichment using $\text{MAGMA}_{\text{GSA}}$ in the $\text{Education}_{\text{Discovery}}$ sample (community 5, p_{MAGMA}=0.002, number of genes=106 ; community 53 n=405, p_{MAGMA}=0.015 ; community 22 n=42, P_{MAGMA}=0.035; community 47, P_{MAGMA}= 0.035 n=47) (table B.22). All communities other than 53 were among those ($n = 7$) showing nominally significant enrichment using GSEA (5,11,20,22,24,33,47) and so communities 5 (p_{GSEA}= 0.004), 22(p_{GSEA}= 0.014) and 47 (p_{GSEA}= 0.017) were tested using the replication samples. p_{GSEA} for community 53 was 0.29 (table 5.3).

5.1.3.2.2 Intelligence In the $\text{Intelligence}_{\text{Discovery}}$ sample, four gene sets showed nominally significant enrichment using $\text{MAGMA}_{\text{GSA}}$ (community 10, p_{MAGMA}=0.008, n=146; community 5 p_{MAGMA}=0.014, n=106; community 22, p=0.046 n=42; cluster

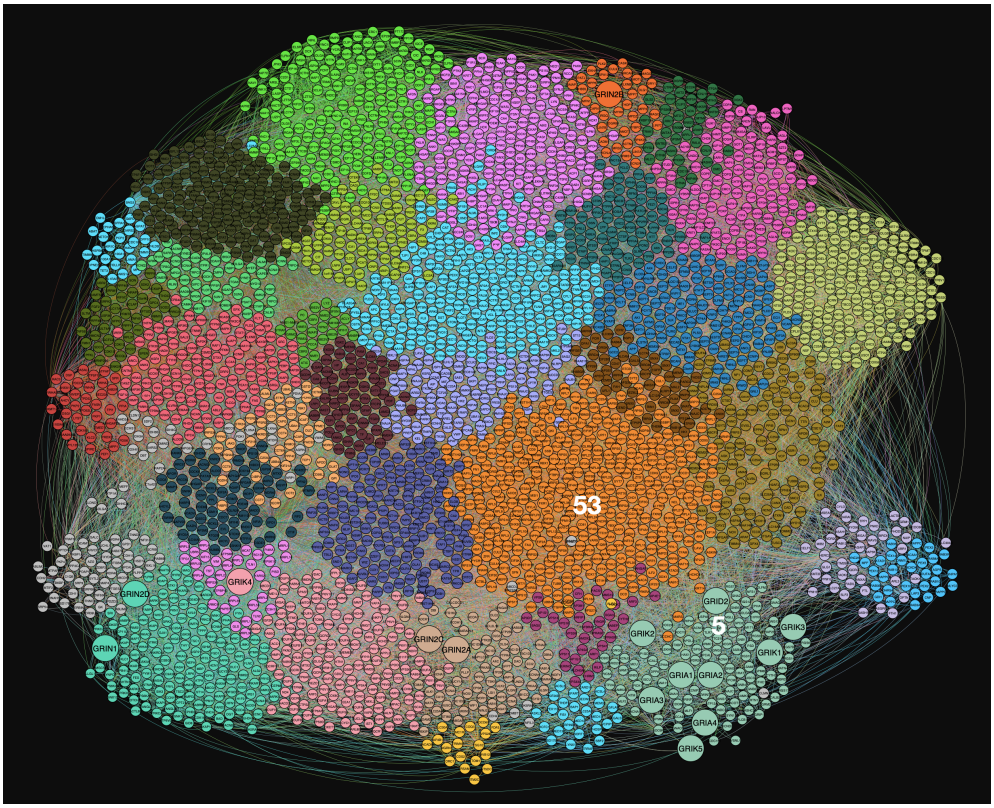


Figure 5.2: PSP visualised in Gephi. Group 5 and 53 labelled. Glutamate receptors are highlighted by using larger node size

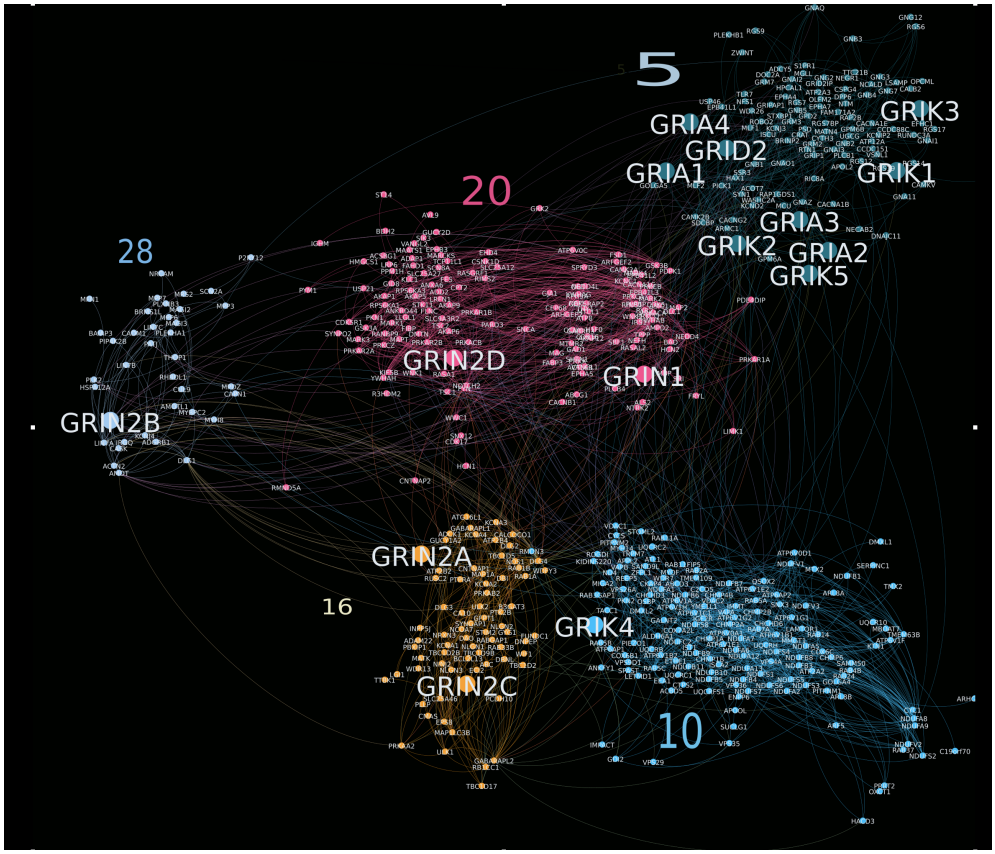


Figure 5.3: Distribution of glutamate receptors across spectral groups and interconnections

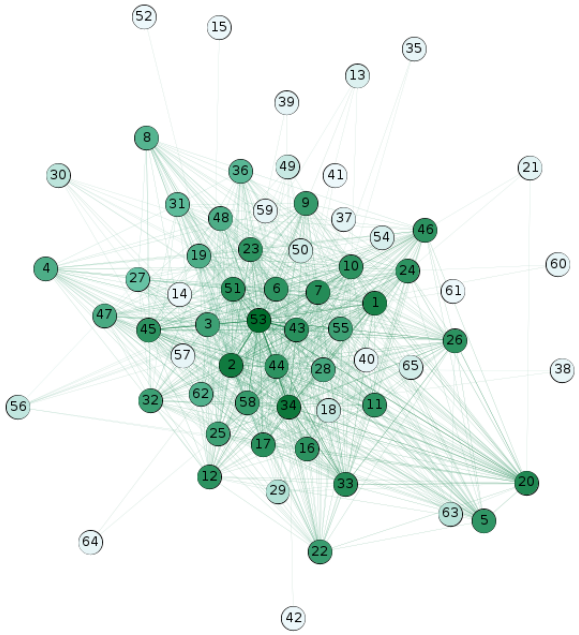


Figure 5.4: Interaction of spectral groups in the PSP. Each node represents a community. Edge thickness is proportional to the edges between communities. Shading (deeper green) is proportional to community size

53 p=0.050, n=386). However, only community 5 in the $\text{Intelligence}_{\text{Discovery}}$ sample showed at least nominally significant enrichment on GSEA (table 5.4) and MAGMA GSA ($p_{\text{MAGMA}} <0.05$; table 5.1), and was analysed in the replication sample.

5.1.3.2.3 Replication Community 5 was confirmed to have an enriched association with educational attainment in the $\text{Education}_{\text{Replication}}$ sample p_{MAGMA}=0.01548, p_{MAGMA} 0.045 Bonferroni ; Community 5 genes also were enriched in the $\text{Intelligence}_{\text{Replication cohort}}$ (p_{MAGMA} = 0.0084 , table 5.1 ; p_{GSEA}=0.003, table 5.4)

The analysis was repeated using gene scoring and the pathway enrichment function in Pascal [228]. The results from the $\text{Intelligence}_{\text{Discovery}}$ and $\text{Intelligence}_{\text{Replication}}$ samples are shown in table 5.5. Community 5 showed significant enrichment in all four cohorts using Pascal's gene scoring and GSA analysis methods. $\text{Education}_{\text{Discovery}}$ p=0.00007, $\text{Intelligence}_{\text{Discovery}}$ p=0.0001; $\text{Education}_{\text{Replication}}$ p=0.0003, $\text{Intelligence}_{\text{Replication}}$ p=0.0001; alpha Bonferroni=0.0015 table 5.5 and table 5.6. The chi squared is reported as the authors state in an '*extreme case one gene might reach scores so high that it precludes detection of pathways not containing that gene when the empirical sampling strategy is used*'[228]. Community five was the most enriched set in all samples tested in Pascal.

	Discovery					*Replication*				
Community	*n*	β	β STD	SE	*p*	*n*	β	β STD	SE	*p*
1	161	-0.03	-0.00	0.07	0.639	162	0.05	0.01	0.07	0.207
2	153	0.10	0.01	0.08	0.093	153	-0.07	-0.01	0.07	0.867
3	39	0.22	0.01	0.16	0.082	39	0.09	0.00	0.14	0.260
4	29	-0.17	-0.01	0.18	0.821	30	-0.36	-0.01	0.15	0.991
5	106	0.22	0.02	0.10	**0.014**	106	0.21	0.02	0.09	**0.009**
6	76	-0.07	-0.00	0.12	0.715	77	0.09	0.01	0.10	0.170
7	125	0.14	0.01	0.09	0.052	125	-0.08	-0.01	0.08	0.840
9	61	-0.01	-0.00	0.12	0.545	61	0.23	0.01	0.11	0.016
10	146	0.19	0.02	0.08	**0.008**	146	0.02	0.00	0.07	0.375
11	55	0.14	0.01	0.13	0.145	55	-0.11	-0.01	0.11	0.848
12	32	-0.01	-0.00	0.16	0.522	32	0.18	0.01	0.15	0.113
16	68	0.00	0.00	0.12	0.494	68	0.21	0.01	0.10	0.022
17	144	0.02	0.00	0.08	0.417	144	-0.00	-0.00	0.07	0.506
19	20	0.21	0.01	0.18	0.124	20	0.10	0.00	0.15	0.262
20	144	0.13	0.01	0.09	0.070	144	0.13	0.01	0.07	0.039
22	42	0.24	0.01	0.14	**0.047**	42	-0.06	-0.00	0.13	0.669
23	67	-0.08	-0.00	0.12	0.741	67	-0.17	-0.01	0.10	0.954
24	74	0.10	0.01	0.12	0.191	74	0.18	0.01	0.10	0.032
25	26	0.01	0.00	0.19	0.487	26	-0.03	-0.00	0.16	0.582
26	117	0.10	0.01	0.09	0.130	116	0.06	0.00	0.08	0.245
28	35	-0.04	-0.00	0.17	0.603	35	0.34	0.01	0.15	0.011
32	50	-0.05	-0.00	0.15	0.635	50	0.10	0.01	0.12	0.200
33	180	0.05	0.00	0.07	0.270	180	0.10	0.01	0.06	0.062
34	210	0.06	0.01	0.07	0.216	211	-0.09	-0.01	0.06	0.937
43	137	0.02	0.00	0.08	0.389	139	-0.03	-0.00	0.07	0.656
44	101	0.02	0.00	0.09	0.434	101	-0.00	-0.00	0.08	0.505
45	118	0.09	0.01	0.09	0.139	119	0.03	0.00	0.07	0.335
46	47	0.07	0.00	0.12	0.283	47	0.06	0.00	0.10	0.264
47	23	0.22	0.01	0.20	0.132	24	-0.05	-0.00	0.18	0.619
48	22	0.03	0.00	0.20	0.445	22	-0.05	-0.00	0.21	0.592
51	110	-0.06	-0.00	0.09	0.750	110	0.13	0.01	0.08	0.038
53	386	0.08	0.01	0.05	0.050	387	0.08	0.01	0.04	0.024
55	42	0.22	0.01	0.15	0.069	42	0.04	0.00	0.12	0.371
58	36	-0.10	-0.00	0.16	0.738	36	-0.01	-0.00	0.14	0.543
62	34	-0.06	-0.00	0.15	0.656	34	-0.18	-0.01	0.14	0.905

Table 5.1: MAGMA GSA Intelligence β is the regression coefficient of the gene set. SE is the standard error of the regression coefficient. β STD is, the semi-standardised regression coefficient, corresponding to the predicted change in Z-value given a change of one standard deviation in the predictor gene set gene covariate. Note that the number of genes refers to the number of genes tested in MAGMA which may be less than the number of genes in the community, as some genes may have no assigned SNPs.

/RProjects/chapter_community_detection/R/magma_gsa/intelligence_gsa_table.R

2021 /RProjects/chapter_community_detection/R/magma_gsa/rawread/intelligence_gsa_new_raw_table.R

	Discovery					*Replication*				
Community	*n*	β	β STD	SE	*p*	*n*	β	β STD	SE	*p*
1	161	0.10	0.01	0.08	0.104	160	0.07	0.01	0.07	0.169
2	153	0.09	0.01	0.08	0.120	153	0.17	0.02	0.07	0.012
3	39	-0.21	-0.01	0.16	0.905	39	-0.02	-0.00	0.14	0.556
4	29	-0.30	-0.01	0.19	0.944	29	-0.12	-0.00	0.15	0.789
5	106	0.29	0.02	0.10	**0.002**	106	0.20	0.02	0.09	**0.015**
6	76	0.01	0.00	0.12	0.454	76	0.10	0.01	0.10	0.162
7	125	0.01	0.00	0.09	0.470	123	-0.09	-0.01	0.08	0.867
9	61	0.10	0.01	0.12	0.191	61	0.03	0.00	0.11	0.402
10	146	0.09	0.01	0.08	0.133	146	0.06	0.01	0.07	0.206
11	55	-0.02	-0.00	0.14	0.570	55	-0.05	-0.00	0.11	0.665
12	32	0.11	0.00	0.16	0.244	32	0.19	0.01	0.16	0.112
16	68	-0.14	-0.01	0.12	0.868	68	0.06	0.00	0.11	0.282
17	144	-0.02	-0.00	0.08	0.605	144	-0.05	-0.00	0.07	0.754
19	20	0.01	0.00	0.19	0.475	20	0.02	0.00	0.19	0.464
20	144	0.00	0.00	0.09	0.489	142	0.22	0.02	0.08	0.002
22	42	0.27	0.01	0.15	0.035	41	0.09	0.00	0.13	0.250
23	67	-0.00	-0.00	0.12	0.509	67	0.12	0.01	0.11	0.125
24	74	0.10	0.01	0.12	0.188	74	0.12	0.01	0.11	0.122
25	26	-0.51	-0.02	0.20	0.995	26	0.12	0.00	0.17	0.238
26	117	0.08	0.01	0.10	0.209	116	-0.09	-0.01	0.09	0.841
28	35	-0.13	-0.01	0.17	0.783	35	0.09	0.00	0.15	0.289
32	50	0.17	0.01	0.15	0.130	49	-0.24	-0.01	0.14	0.961
33	180	0.10	0.01	0.08	0.097	180	0.12	0.01	0.07	0.049
34	210	0.02	0.00	0.07	0.400	209	-0.02	-0.00	0.06	0.620
43	137	-0.04	-0.00	0.08	0.669	137	0.01	0.00	0.08	0.448
44	101	-0.02	-0.00	0.10	0.563	100	-0.01	-0.00	0.08	0.562
45	118	0.03	0.00	0.09	0.374	118	0.04	0.00	0.08	0.294
46	47	0.12	0.01	0.13	0.169	47	0.11	0.01	0.12	0.167
47	23	0.37	0.01	0.20	0.035	23	-0.05	-0.00	0.21	0.599
48	22	0.07	0.00	0.21	0.374	22	0.07	0.00	0.22	0.367
51	110	0.01	0.00	0.09	0.466	110	-0.04	-0.00	0.08	0.710
53	386	0.11	0.02	0.05	0.015	381	0.06	0.01	0.04	0.085
55	42	0.05	0.00	0.15	0.378	42	-0.16	-0.01	0.15	0.855
58	36	-0.12	-0.01	0.16	0.767	35	-0.05	-0.00	0.16	0.629
62	34	-0.14	-0.01	0.16	0.808	34	-0.14	-0.01	0.14	0.826

Table 5.2: MAGMA GSA Educational attainment β is the regression coefficient of the gene set. SE is the standard error of the regression coefficient. β STD is the the semi-standardized regression coefficient, corresponding to the predicted change in Z-value given a change of one standard deviation in the predictor gene set / gene co-variate. Note that the number of genes refers to the number of genes tested in MAGMA which may be less than the number of genes in the community, as some genes may have no assigned SNPs.

RProjects/chapter_community_detection/R/magma_gsa/education_gsa_table.R

	Discovery			*Replication*		
Community	n	ES	p	n	ES	p
24	73	0.61	0.002	73	0.58	0.004
33	179	0.54	0.002	179	0.52	0.003
5	105	0.58	0.006	105	0.52	0.015
47	22	0.70	0.009			
20	143	0.55	0.015	141	0.52	0.005
22	41	0.62	0.018			
11	54	0.58	0.024			
55				41	0.64	0.003
9				60	0.59	0.006
43				136	0.52	0.014

Table 5.3: Nominally significant gene sets using GSEA ($p < 0.05$) Educational Attaintment. Discovery and replication cohorts. ES = Enrichment score. n = number of genes in gene set, p= nominal p value

RProjects/chapter_community_detection/R/gsea/format_sig_educational.R

	Discovery			*Replication*		
Community	n	ES	p	n	ES	p
5	105	0.60	0.001	105	0.54	0.003
53	385	0.51	0.004			
43	136	0.54	0.013			
34	209	0.51	0.023			
3	38	0.60	0.038			
20				143	0.51	0.009
33				179	0.48	0.021
26				115	0.50	0.033

Table 5.4: Nominally significant gene sets using GSEA ($p < 0.05$) Intelligence Discovery and Replication samples. ES = Enrichment score. n = number of genes in gene set, p= nominal p value

RProjects/chapter_community_detection/R/gsea/format_sig_intelligence.R

	Discovery		*Replication*	
Community	χ^2 p value	emp p value	χ^2 p value	emp p value
5	0.00005	0.00011	0.00005	0.00012
10	0.00118	0.00842	0.00215	0.00238
53	0.00551	0.00674	0.05289	0.04662
7	0.01457	0.03852	0.23362	0.17917
33	0.02126	0.06174	0.03314	0.03936
26	0.02152	0.03371	0.09655	0.12936
20	0.02629	0.05821	0.00142	0.00239
34	0.03021	0.07349	0.48926	0.45609
28	0.06082	0.06805	0.18416	0.23682
43	0.06633	0.03141	0.37601	0.41863
55	0.08403	0.13622	0.44314	0.47030
22	0.09465	0.07006	0.32361	0.26264
3	0.09514	0.07290	0.13478	0.11655
19	0.09847	0.05388	0.88967	0.86107
58	0.12710	0.10808	0.26938	0.19509
2	0.13497	0.16725	0.76900	0.73094
45	0.13872	0.15079	0.43880	0.38630
12	0.14885	0.21203	0.13265	0.17448
11	0.15258	0.13086	0.90882	0.89701
24	0.17389	0.16971	0.04770	0.06090
46	0.18991	0.18483	0.20068	0.22571
47	0.26565	0.24708	0.34319	0.29384
1	0.29315	0.32094	0.21352	0.24257
9	0.31205	0.38279	0.07316	0.09553
23	0.33463	0.39840	0.64103	0.66756
32	0.33661	0.41270	0.17894	0.21654
44	0.36507	0.39839	0.57021	0.53741
17	0.37087	0.31369	0.35273	0.31447
25	0.44107	0.44977	0.44278	0.45825
51	0.45466	0.56525	0.07424	0.08153
62	0.65120	0.63921	0.47485	0.49011
48	0.65913	0.65999	0.82988	0.82570
16	0.73722	0.70116	0.08981	0.11902
6	0.76799	0.74957	0.39325	0.31039
4	0.80237	0.78159	0.96353	0.96408

Table 5.5: Intelligence samples Pascal GSA Discovery and Replication samples ordered by χ^2 p value. emp = empirical p value method.

	Discovery		*Replication*	
Community	χ^2 p value	emp p value	χ^2 p value	emp p value
5	0.00006	0.00007	0.00101	0.00026
20	0.00009	0.00064	0.00019	0.00214
2	0.00010	0.00197	0.02859	0.05484
33	0.00033	0.00351	0.00774	0.03373
47	0.01040	0.03107	0.29248	0.30947
24	0.01135	0.00904	0.00121	0.00124
53	0.01668	0.03576	0.44041	0.57615
22	0.01690	0.04507	0.17748	0.17328
1	0.04775	0.04565	0.19103	0.20056
11	0.04848	0.02619	0.24541	0.10527
10	0.04947	0.12340	0.14220	0.12294
34	0.06379	0.09816	0.02931	0.08681
32	0.07427	0.14669	0.88655	0.88639
26	0.12693	0.11476	0.29690	0.21301
51	0.12879	0.19808	0.06750	0.10246
44	0.13406	0.21763	0.12548	0.22032
46	0.16189	0.17636	0.37360	0.40041
7	0.20011	0.23541	0.44014	0.47564
4	0.21800	0.27362	0.94031	0.93664
12	0.22076	0.24324	0.12664	0.19635
45	0.22085	0.33226	0.34137	0.41257
55	0.23156	0.31491	0.04923	0.05854
9	0.23518	0.32986	0.03420	0.07557
19	0.29122	0.33875	0.37543	0.39652
3	0.31886	0.27426	0.34172	0.35703
28	0.33989	0.36836	0.19262	0.26534
23	0.39185	0.47784	0.58355	0.61850
6	0.42214	0.46701	0.70931	0.70462
17	0.42623	0.47928	0.60287	0.57664
16	0.43234	0.51029	0.31117	0.42090
43	0.47397	0.47877	0.00273	0.00416
48	0.49223	0.46679	0.55516	0.56118
58	0.71790	0.74793	0.62110	0.65671
62	0.96246	0.96134	0.19788	0.25513
25	0.96330	0.96085	0.64162	0.65989

Table 5.6: Educational attainment samples Pascal GSA Discovery and Replication samples ordered by χ^2 p value. emp = empirical p value method.

5.1.4 Gene ontology analysis

Community 5 is significantly enriched for terms related to glutamate receptor activity and the heterotrimeric g protein complex, including the G protein coupled serotonin receptor. Community 5 has a significant proportion of these terms whether one compares against the rest of the genome or relative to other synaptic groups. 21 of the 27 genes annotated by the Gene Ontology Consortium as glutamate receptor activity (GO:0008066 MF) are found in the PSP. Of these, 12 are found in group 5 (57.1%). 23 of the 37 genes encoding proteins annotated heterotrimeric g protein (GO:0005834) are found in the PSP. Of these, 16 are found in group 5 (69.5% of PSP GO:0005834).

ToppGene and PANTHER over-representation test was carried out for each of the GO clades, Biological Process, Molecular Function, Cellular Component with the genes in community five compared with the 3.457 PSP genes. ToppGene ORA results ordered by FDR are shown for Molecular Function (table 5.7), Biological Process (table 5.8) and Cellular Component (table 5.9).

The biological process (BP) with the lowest FDR q value was G protein coupled receptor signalling activity, GO:0007186 n=38, n PSP=189, FDR q 2.36×10^{-18}. The same term was top in PANTHER although the absolute enrichment level was different (GO:0007186: FDR q=$2.16 \times 10-8$, fold enrichment 4.23). The second lowest was glutamate receptor signalling pathway GO:0007215 n=17, n PSP= 43 ,FDR 3.46×10^{-12} (PANTHER, also second term GO 0007215: FDR q 7.23 x 10-07, fold change 12.62). These two terms were used for subgroup analysis (section 5.1.6).

Glutamate receptor activity GO:0008066 was also the most over represented molecular function term using both enrichment methods: ToppGene n =12, n PSP=21, 6.96×10^{-11} (PANTHER FDR 7.58 x 10-07, fold change 16). The most enriched cellular component was GTPase complex GO:1905360: n=16 , n PSP=22, FDR q=6.40×10^{-18} (PANTHER FDR 8.75×10^{-8}, fold change 18.23, also first). This was followed by heterotrimeric G-protein complex:GO:0005834 n= 16 , n PSP= 22, FDR q=6.40×10^{-18} (PANTHER FDR $q = 1.75 \times 10^{-7}$, fold change 18.23).

5.1.5 Murine and Disease Phenotypes in Community 5

The genes within group 5 have a strongly-enriched association with murine phenotypes related to synaptic transmission (see table 5.10 for enrichment with PSP background).

The fifth term is abnormal long term potentiation (MP:0002207); 15 of 117 terms in PSP (12.8% of terms, 3.79 fold change compared with PSP, FDR q=2.46×10^{-3}).

ID	Name	n	n PSP	p	q FDR B.H
GO:0008066	glutamate receptor activity	12	21	1.86×10^{-13}	6.96×10^{-11}
GO:0016595	glutamate binding	12	23	8.11×10^{-13}	1.52×10^{-10}
GO:0004970	ionotropic glutamate receptor activity	9	16	2.89×10^{-10}	3.60×10^{-8}
GO:0016597	amino acid binding	12	40	2.10×10^{-9}	1.96×10^{-7}
GO:1904315	transmitter-gated ion channel activity	9	20	3.80×10^{-9}	2.84×10^{-7}
GO:0005230	extracellular ligand-gated ion channel	9	21	6.47×10^{-9}	3.00×10^{-7}
GO:0022835	transmitter-gated channel activity	9	21	6.47×10^{-9}	3.00×10^{-7}
GO:0022824	transmitter-gated ion channel activity	9	21	6.47×10^{-9}	3.00×10^{-7}
GO:0031821	G protein-coupled serotonin RB	6	7	7.23×10^{-9}	3.00×10^{-7}
GO:0099529	neurotransmitter receptor activity PS	9	22	1.06×10^{-8}	3.98×10^{-7}

Table 5.7: Community 5 GO: Molecular Function enrichment using ToppGene with PSP ($n = 3,457$) genes as background. 79 significant terms, showing 10 for illustration. PS = post synapse, RB = receptor binding

RProjects/chapter_community_detection/R/group_enrichment/read_toppgene5/enrichMF.R

Other terms are related to abnormal emotion, anxiety, hyperactivity and fear related behaviour, showing an effect from genes on both LTP and complex behaviour (see [158],[159]).

Community five has an enriched association with genes that have been reported to be implicated in the aetiology of common neuropsychiatric disorders. Using ToppGene to perform disease enrichment, using terms in DisGeNet, and the PSP as background, the most enriched disorder is schizophrenia (C0036341 n=29, n PSP=291, FDR q=2.28×10^{-3}). This comprises approximately 10% of the terms annotated in the PSP. Depressive disorder also shows an enriched association (table 5.11).

5.1.6 Subgroup analysis

Two large components can be identified in community five using gene ontology enrichment analysis: heterotrimeric g proteins (of which G-protein coupled serotonin receptor complex is a subset) and glutamate receptors. These are strongly associated with human disease phenotypes and murine models of cognitive ability (figure 5.5).

As the community gene set has been discovered using only network topology, Gene Ontology terms can be used to provide a finer dissection of community properties. For example, all of the members of a community that are or are not enriched for a term can

ID	Name	n	n PSP	p	q FDR B.H
GO:0007186	G protein-coupled receptor signaling pathway	38	189	9.89×10^{-22}	2.36×10^{-18}
GO:0007215	glutamate receptor signaling pathway	17	43	2.89×10^{-15}	3.46×10^{-12}
GO:0051339	regulation of lyase activity	18	66	7.30×10^{-13}	5.52×10^{-10}
GO:0098916	anterograde trans-synaptic signaling	42	414	1.16×10^{-12}	5.52×10^{-10}
GO:0007268	chemical synaptic transmission	42	414	1.16×10^{-12}	5.52×10^{-10}
GO:0099537	trans-synaptic signaling	42	420	1.91×10^{-12}	7.60×10^{-10}
GO:0007188	adenylate cyclase-modulating G protein-c	16	54	3.76×10^{-12}	1.09×10^{-9}
GO:0099536	synaptic signaling	42	429	3.98×10^{-12}	1.09×10^{-9}
GO:0045761	regulation of adenylate cyclase activity	17	63	4.09×10^{-12}	1.09×10^{-9}
GO:0031279	regulation of cyclase activity	17	65	7.15×10^{-12}	1.71×10^{-9}

Table 5.8: Community 5 GO: Biological Process 84 significant terms

RProjects/chapter_community_detection/R/group_enrichment/read_toppgene5/enrichMF.R

be identified.

The top two Biological Process terms were used as they occurred in the same rank order in Panther as in ToppGene. Of the 106 genes in community 5 tested using MAGMA, 16 were members of GO:0007215 glutamate signalling pathway and 38 were members of GO:0007186 G protein-coupled receptor signalling pathway. There is some overlap between these terms as their union has 45 members detectable in MAGMA. Gene set files (`gmt`) were prepared with only Glutamate signalling pathway genes, only G protein-coupled receptor signalling pathway genes, their 45 genes union and the 61 genes that did not belong to these groups. These were then tested using MAGMA GSA to test the hypothesis that the enrichment in population studies was due to either glutamate signalling pathway genes or G protein-coupled receptors.

We find that the genetic variations in G protein receptors and glutamate receptors do not have an enriched association with educational attainment or human cognitive ability, as one might expect based on animal phenotypes; rather, the remaining genes in community 5 after these have been removed have increased enrichment for educational attainment or intelligence.

In the intelligence discovery sample the remaining 61 genes were more enriched $p = 0.00069$, than group 5 as a whole $p = 0.014$ (20.3 x increase). Intelligence replication showed a decrease in enrichment from 0.0086 to 0.017, having removed glutamate and g protein pathway genes, but glutamate and g protein were not significant on their own ($p = 0.11$). In the Education Discovery sample there was almost no difference in enrichment between community 5 (p=0.00196) and the remaining genes (p=0.00209)

ID	Name	n	n PSP	p	q FDR B.H
GO:1905360	GTPase complex	16	22	3.44×10^{-20}	6.40×10^{-18}
GO:0005834	heterotrimeric G-protein complex	16	22	3.44×10^{-20}	6.40×10^{-18}
GO:0098797	plasma membrane protein complex	35	215	1.01×10^{-16}	1.25×10^{-14}
GO:0031234	extrinsic component of cytoplasmic side	17	59	1.37×10^{-12}	1.28×10^{-10}
GO:0030425	dendrite	41	435	3.60×10^{-11}	2.23×10^{-9}
GO:0097447	dendritic tree	41	435	3.60×10^{-11}	2.23×10^{-9}
GO:0097060	synaptic membrane	29	243	2.53×10^{-10}	1.34×10^{-8}
GO:0019898	extrinsic component of membrane	24	174	6.13×10^{-10}	2.85×10^{-8}
GO:0036477	somatodendritic compartment	44	560	2.41×10^{-9}	9.85×10^{-8}
GO:0045211	postsynaptic membrane	22	157	2.65×10^{-9}	9.85×10^{-8}

Table 5.9: Group 5 GO: Cellular Component 78 significant terms

RProjects/chapter_community_detection/R/group_enrichment/read_toppgene5/enrichCC.R

but there was not enrichment of glutamate pathway or g proteins, suggesting that it is the residual genes that are the source of the signal. In the education replication sample the remaining genes were $p = 0.0093$ and community 5 as a whole was p =0.0147. There was no enrichment of g proteins or glutamate.

Removing the g protein and glutamate pathway receptor genes however removes the enrichment signal for murine LTP phenotypes[1]. There is no significant enrichment for any Gene Ontology terms using the PSP as background. These genes appear to have little in common other than being neighbours and interaction partners with g-protein coupled signalling receptor pathway proteins and glutamate receptors. .

Findings for larger datasets are shown in figure 5.22

5.1.6.1 Heterotrimeric G proteins

Heterotrimeric G proteins are made up of alpha, beta and gamma subunits and act as the transducers of G protein coupled receptors, and link receptor activation to second messengers[277](see section 1.2.2).

30 genes from community 5 are are involved in GTPase activity: GO:0003924, total 796 genes. It contains 6 G-protein alpha subunits binding genes GO:0001965 (RGS19 regulator of G protein signaling 19, RGS7 , RGS12, RGS14, GRIA1 glutamate ionotropic receptor AMPA type subunit 1, RIC8A RIC8 guanine nucleotide exchange factor A) and six of the seven elements of G protein-coupled serotonin re-

[1] /home/grant/RProjects/chapter_community_detection/data/toppgene/group5subgroupremoved/group5subgroupremoved.txt

ID	Name	n	n PSP	p	q FDR B.H
MP:0001970	abnormal pain threshold	12	64	1.48×10^{-6}	2.28×10^{-3}
MP:0020870	decreased thigmotaxis	9	40	6.96×10^{-6}	2.46×10^{-3}
MP:0002572	abnormal emotion/affect behaviour	30	377	7.30×10^{-6}	2.46×10^{-3}
MP:0001362	abnormal anxiety-related response	20	194	8.67×10^{-6}	2.46×10^{-3}
MP:0002207	abnormal long term potentiation	15	117	1.00×10^{-5}	2.46×10^{-3}
MP:0001399	hyperactivity	24	268	1.11×10^{-5}	2.46×10^{-3}
MP:0002065	abnormal fear/anxiety-related behaviour	22	232	1.12×10^{-5}	2.46×10^{-3}
MP:0003635	abnormal synaptic transmission	30	391	1.56×10^{-5}	2.99×10^{-3}
MP:0001968	abnormal touch/ nociception	12	82	2.20×10^{-5}	3.75×10^{-3}
MP:0009745	abnormal behavioural response to xenobiotic	14	113	3.06×10^{-5}	4.70×10^{-3}

Table 5.10: Group 5 Mouse Phenotype 28 significant terms

RProjects/chapter_community_detection/R/group_enrichment/read_toppgene5/enrich_murine.R

ceptor binding (GO:0031821) GNA11 (G protein subunit alpha 11), GNAI1, GNAI3, GNAO1, GNAQ, GNAZ and 6 of 22 G protein beta-subunit binding activity genes GNG2, GRIA1, RGS7, GNG12, GNG3, GNG7coupled.

Previous research into the role of G proteins in intelligence concluded they have no effect[177]. These findings support this but show they are in close approximation to genes enriched for effect with intelligence and educational attainment. Importantly, G protein coupled receptors make up an estimated 30-44% of drug targets[61].

5.1.7 Genetic Constraint

I used the probability of loss of function intolerance (pLI) measure in the Gnomad v2.1.1 database as a measure of genetic constraint (section 3.6.11).

The median pLI (0.72) in community five is higher than the PSP in general (0.246) (table 5.14). Figure 5.6 shows mean pLI in community 5, community 5 genes in GO:0007186 G protein-coupled receptor signalling pathway, community five genes in GO:0007215 glutamate signalling pathway, the residual genes in group 5 after these have been removed, and the PSP and Non PSP. Median pLI for the same categories is shown in 5.7. Genetic constraint is clearly higher in the genes in the glutamate signalling pathway than community five, and community five is under greater constraint than the rest of the PSP. The PSP as a whole is under greater constraint than non PSP genes.

Genetic constraint had a modest correlation with intelligence and educational at-

ID	Name	n	n in PSP	p	q FDR B.H
C0036341	Schizophrenia	29	291	1.16×10^{-8}	2.42×10^{-5}
C0443306	Muscle Spasticity	6	10	1.80×10^{-7}	1.27×10^{-4}
C0746408	mass lesion	5	6	1.82×10^{-7}	1.27×10^{-4}
C0033953	Psychosexual Disorders	3	3	3.33×10^{-5}	8.68×10^{-3}
C0020594	Hypoactive Sexual Desire Disorder	3	3	3.33×10^{-5}	8.68×10^{-3}
C0036902	Sexual Arousal Disorder	3	3	3.33×10^{-5}	8.68×10^{-3}
C0016722	Frigidity	3	3	3.33×10^{-5}	8.68×10^{-3}
C0029261	Orgasmic Disorder	3	3	3.33×10^{-5}	8.68×10^{-3}
C0011570	Mental Depression	25	340	4.37×10^{-5}	1.01×10^{-2}
C0011581	Depressive disorder	25	344	5.33×10^{-5}	1.11×10^{-2}

Table 5.11: Disease enrichment of community 5 using TopppGene for DisGeNet lookup. Significant disease terms=24. ID= concept unique identifier for UMLS (unified medical language system

tainment (section 3.13.2) and the association was greatest for the PSP.

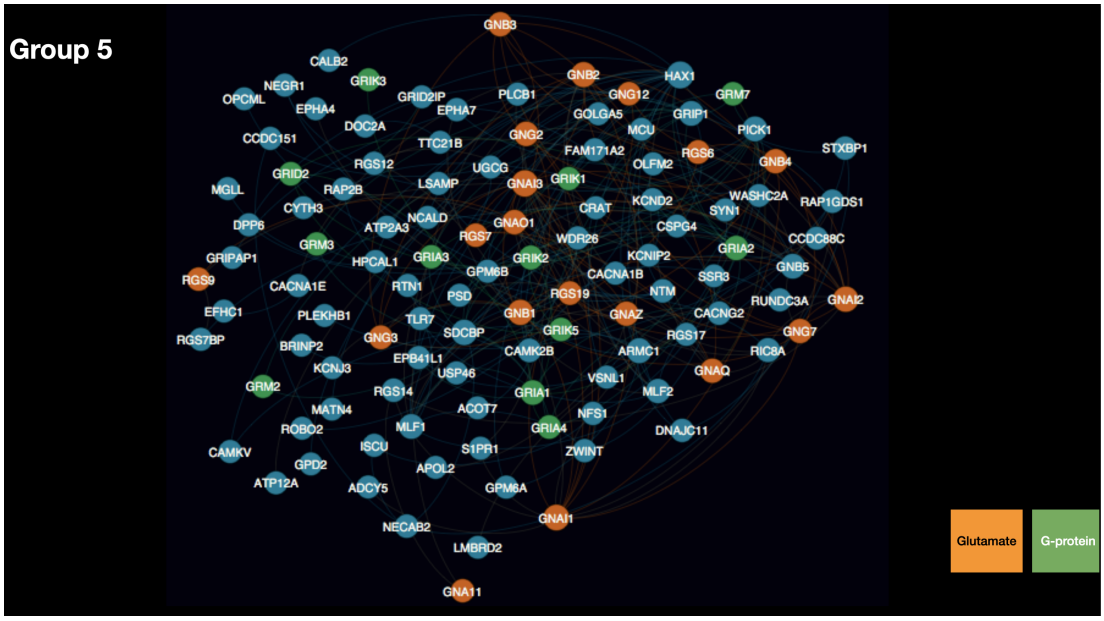


Figure 5.5: Group 5 showing glutamate receptors and G proteins. Glutamate receptors are shown based on GO:008066 rather than Biological Process as shown in the text.

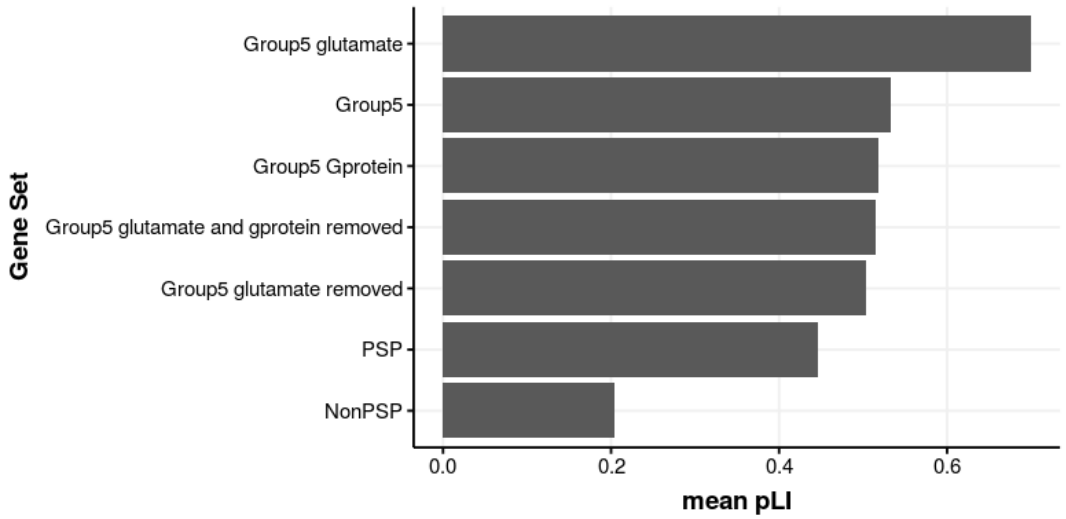


Figure 5.6: Mean pLI subgroups

RProjects/chapter_community_detection/R/pLI_groups/plot/plot_pli_group5mean.R

	Discovery					*Replication*				
Community	*n*	β	β STD	SE	*p*	*n*	β	β STD	SE	*p*
remaining	61	0.41	0.02	0.13	0.00069	61	0.24	0.01	0.11	0.01752
glutamate	16	0.51	0.02	0.28	0.03098	16	0.34	0.01	0.25	0.08540
gprotein	38	-0.22	-0.01	0.17	0.90231	38	0.19	0.01	0.15	0.09936
community 5	106	0.22	0.02	0.10	0.01403	106	0.21	0.02	0.09	0.00860
gprotein & glutamate	45	-0.07	-0.00	0.16	0.67780	45	0.16	0.01	0.14	0.11741

Table 5.12: MAGMA GSA Intelligence subgroup analysis. Glutamate is genes in group 5 that appear in GO:0007215 Glutamate signalling pathway, gprotein is genes in group 5 found in GO:0007215 Glutamate signalling pathway. Remaining refers to the proteins not in the gprotein or glutamate setsβ is the regression coefficient of the gene set. SE is the standard error of the regression coefficient. β STD the the semi-standardized regression coefficient, corresponding to the predicted change in Z-value given a change of one standard deviation in the predictor gene set / gene covariate

RProjects/chapter_community_detection/R/magma_gsa/rawread/intelligence_sub_group_gsa_new_raw_table.R

	Discovery					*Replication*				
Community	*n*	β	β STD	SE	*p*	*n*	β	β STD	SE	*p*
remaining	61	0.37	0.02	0.13	0.00209	61	0.28	0.02	0.12	0.00934
glutamate	16	0.29	0.01	0.28	0.14826	16	0.02	0.00	0.25	0.46703
gprotein	38	0.12	0.01	0.17	0.24631	38	0.10	0.00	0.15	0.26420
community 5	106	0.29	0.02	0.10	0.00196	106	0.20	0.02	0.09	0.01477
gprotein & glutamate	45	0.17	0.01	0.16	0.14945	45	0.08	0.00	0.14	0.28350

Table 5.13: MAGMA GSA Education subgroup. Glutamate is genes in group 5 that appear in GO:0007215 Glutamate signalling pathway, gprotein is genes in group 5 found in GO:0007215 Glutamate signalling pathway. Remaining refers to the proteins not in the gprotein or glutamate setsβ is the regression coefficient of the gene set. SE is the standard error of the regression coefficient. β STD the the semi-standardized regression coefficient, corresponding to the predicted change in Z-value given a change of one standard deviation in the predictor gene set / gene covariate

RProjects/chapter_community_detection/R/magma_gsa/rawread/education_sub_group_gsa_new_raw_table.R

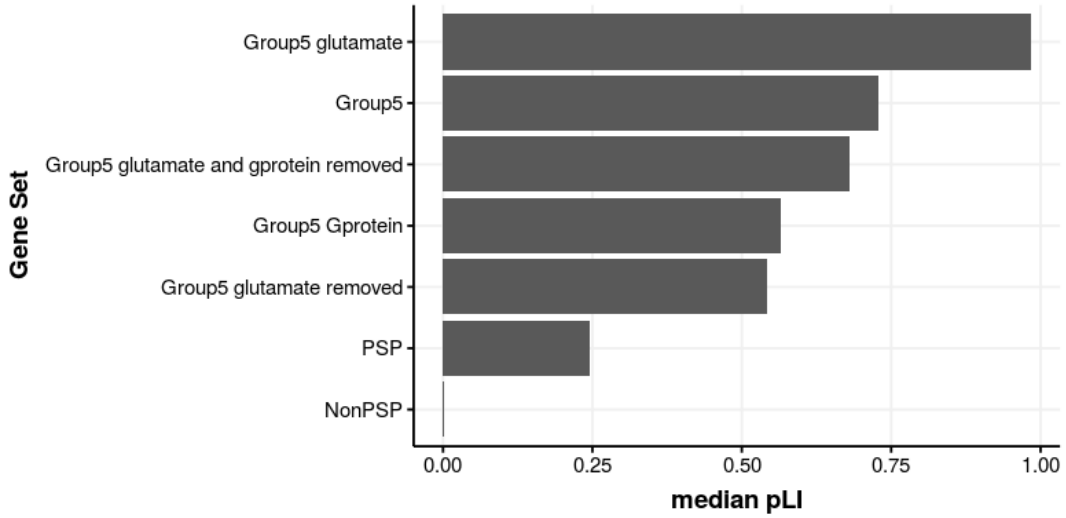


Figure 5.7: Median pLI subgroups

RProjects/chapter_community_detection/R/pLI_groups/plot/plot_pli_group5median.R

	n	mean pLI	median	IQR	min	Q1	Q3	max
community5 glutamate	17	0.699	0.983	0.874	2.236×10^{-13}	1.245×10^{-1}	0.999	1.000
community5	112	0.533	0.728	0.981	2.704×10^{-26}	5.619×10^{-3}	0.986	1.000
community5 Gprotein	38	0.519	0.566	0.949	1.125×10^{-14}	1.114×10^{-2}	0.960	1.000
C5 glutamate and gpr removed	66	0.514	0.679	0.984	2.704×10^{-26}	4.928×10^{-4}	0.984	1.000
community5 glutamate removed	95	0.503	0.542	0.958	2.704×10^{-26}	1.479×10^{-3}	0.960	1.000
PSP	3407	0.447	0.246	0.988	1.861×10^{-164}	3.118×10^{-5}	0.988	1.000
Non PSP	14972	0.205	0.001	0.231	8.222×10^{-118}	2.068×10^{-8}	0.231	1.000

Table 5.14: Probability of loss of function intolerance (pLI) for members of community 5 belonging to different gene ontology biological processes. C5= community 5, gpr= gprotein receptor. For details see section 5.1.7

	Community	mean pLI	median pLI	IQR	*n* genes
1	55	0.613	0.966	0.989	43
2	24	0.620	0.931	0.997	75
3	43	0.615	0.886	0.998	144
4	53	0.566	0.827	0.993	402
5	5	0.533	0.728	0.981	112
6	33	0.516	0.666	0.999	188
7	20	0.543	0.652	0.984	147
8	16	0.490	0.581	0.985	72
9	34	0.489	0.519	0.998	211
10	26	0.489	0.502	0.995	119
11	23	0.460	0.393	0.998	72
12	32	0.419	0.296	0.947	52
13	2	0.448	0.279	0.970	158
14	11	0.406	0.146	0.922	56
15	6	0.379	0.116	0.985	78
16	46	0.416	0.104	0.985	48
17	44	0.399	0.091	0.974	102
18	1	0.437	0.083	0.996	170
19	7	0.334	0.076	0.784	128
20	12	0.326	0.063	0.764	34
21	28	0.405	0.051	0.992	37
22	22	0.388	0.046	0.990	42
23	58	0.366	0.043	0.933	35
24	3	0.396	0.040	0.999	41
25	10	0.318	0.026	0.747	153
26	47	0.463	0.025	1.000	26
27	17	0.313	0.018	0.796	145
28	51	0.364	0.009	0.909	111
29	45	0.324	0.007	0.884	121
30	25	0.225	0.006	0.187	27
31	19	0.233	0.004	0.546	20
32	9	0.283	0.002	0.855	62
33	4	0.292	0.001	0.631	31
34	48	0.161	0.001	0.059	23
35		0.205	0.001	0.231	14972
36	62	0.235	0.000	0.191	34

Table 5.15: Spectral communities ordered by median pLI . Communities smaller than 15 filtered out for presentation. Community five has the fifth highest level of median constraint.

RProjects/chapter_community_detection/R/pLI_groups/pLI_across_groups.R

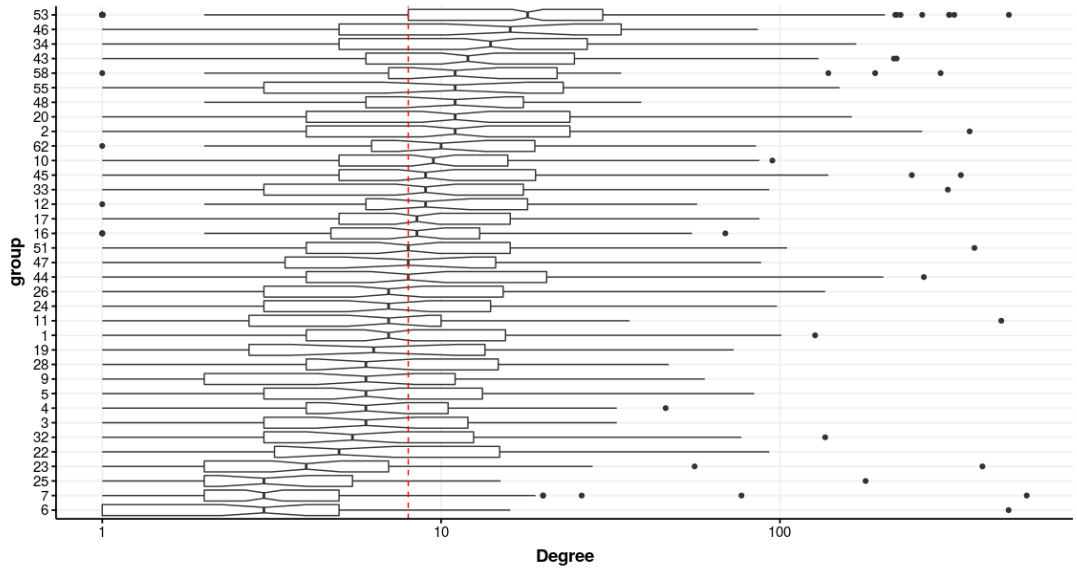


Figure 5.8: Box plot of the degree distribution of communities detected by the spectral clustering algorithm. communities with less than fifteen members omitted. Red dashed line is median degree of nodes in PSP. Degree is shown on log10 scale.

RProjects/chapter_community_detection/R/group5/degree/plot_histogram_degreebygroup.R

5.1.8 Centrality measures for community 5

The cohesive module that showed enrichment in the discovery and replication cohorts (community 5) had significantly lower mean closeness centrality (table B.25[2]) than the other communities and had the lowest eigenvector centrality (table B.26), which we have previously noted to be strongly correlated with being essential (table 3.26). Mean kcore was significantly lower (mean kcore=6.62) than the rest of the PSP after accounting for multiple comparisons (table B.28). Betweenness and degree centrality were not significantly different from other communities (table 5.16). At a community level, community 5 possessed the network characteristics that are less associated with essentiality despite the high level of pLI (figure 5.6). Community 5 has a higher ranked pLI (table 5.15), but its mean is not significantly different to the PSP as a whole (Wilcoxon rank sum test, W = 174,757, p value = 0.3335).

5.1.9 Louvain MAGMA GSA

The Louvain algorithm performed well in benchmarks and results for MAGMA GSA are included here.

[2] RProjects/chapter_community_detection/R/group5/degree/plot_histogram_closenessbygroup.R

Centrality measure	μ_{k_C}	Median	$\delta\mu$	p	p BH	n	rank.sig
Closeness	0.0000917	0.0000921	-0.00000678	2.01×10^{-9}	1.30×10^{-7}	7	2
Degree	10.6875000	6.0000000	-6.9567002	1.428×10^{-2}	0.79	7	NS
kcore	6.625	5.0	-2.526	5.319×10^{-4}	0.03	10	10
eigenvector	0.0167635	0.0089846	-0.0311651	8.827×10^{-12}	5.600×10^{-10}	16	2
transitivity 0	0.117	0.0503507	0.008	4.206×10^{-4}	2.500×10^{-2}	9	7
Betweenness	2272.6	247.42	1955.57	7.582×10^{-1}	1	4	NS

Table 5.16: Ranking of community 5 amongst 65 communities for centrality measures. community 5 shows low eigenvector centrality and low closeness and belongs to an outer kcore - features that have been associated with less severe clinical phenotypes

Community 7 shows enrichment in the $\text{Intelligence}_{\text{Discovery}}$ sample (p=0.0012) and $\text{Intelligence}_{\text{Replication}}$ samples (p=0.0026). Community 7 is composed of 154 genes (table 5.17) but the number shown in the tables of results may be less than this if a gene was not assigned any SNPs using MAGMA (n=144).

If we had used the Louvain algorithm with the the same study design as the spectral method we would have taken forward communities 2,3,7 and 12 (considering only MAGMA) so the adjusted level of α for the $\text{Intelligence}_{\text{Replication}}$ sample would be 0.05/4= 0.0125. The significance level for community 7 $\text{Intelligence}_{\text{Replication}}$ is lower (p=0.0026) than the penalty for including all of the Louvain communities α=0.00357 (0.05/14) (table 5.17).

Community 7 is enriched for significant genes in the $\text{Educational attainment}_{\text{Discovery}}$ sample but not the replication cohort (p=0.0146 and replication p=0.4750 ; see table 5.18).

Using gene ontology enrichment in ToppGene community 7 is also enriched for glutamate receptor activity (Molecular function, GO:0008066, glutamate receptor activity; table 5.19). Disease term enrichment is shown in table 5.20.

	Discovery					*Replication*				
Community	*n*	β	β STD	SE	*p*	*n*	β	β STD	SE	*p*
1	196	0.10	0.01	0.07	0.0790	198	-0.05	-0.00	0.06	0.7749
2	334	0.09	0.01	0.05	**0.0407**	335	0.04	0.01	0.05	0.1916
3	510	0.10	0.02	0.04	**0.0075**	512	0.04	0.01	0.04	0.1006
4	245	0.06	0.01	0.06	0.1514	246	0.07	0.01	0.05	0.1053
5	112	-0.08	-0.01	0.09	0.8107	112	0.05	0.00	0.07	0.2525
6	159	0.12	0.01	0.07	0.0580	159	0.05	0.00	0.07	0.2284
7	144	0.26	0.02	0.08	**0.0012**	144	0.21	0.02	0.07	**0.0026**
8	519	0.05	0.01	0.04	0.1103	519	0.07	0.01	0.04	0.0311
9	427	0.01	0.00	0.05	0.4054	428	-0.02	-0.00	0.04	0.6428
10	57	0.11	0.01	0.12	0.1910	57	0.01	0.00	0.11	0.4497
11	331	-0.03	-0.00	0.05	0.6911	332	-0.00	-0.00	0.05	0.5412
12	21	0.63	0.02	0.21	**0.0014**	21	0.17	0.01	0.17	0.1628
13	80	-0.06	-0.00	0.10	0.7250	80	-0.03	-0.00	0.09	0.6106
14	166	0.05	0.00	0.08	0.2554	166	0.17	0.02	0.07	0.0062

Table 5.17: Louvain clustering Intelligence phenotype

RProjects/utils2/R/collate_results/generic_compile_gsa_lourvain.R

	Discovery					*Replication*				
Community	*n*	β	β STD	SE	*p*	*n*	β	β STD	SE	*p*
1	196	0.05	0.00	0.07	0.2610	194	-0.02	-0.00	0.06	0.5955
2	334	0.13	0.02	0.05	**0.0071**	331	0.07	0.01	0.05	0.0568
3	510	0.06	0.01	0.04	0.0878	508	0.04	0.01	0.04	0.1212
4	245	-0.04	-0.00	0.06	0.7183	246	0.06	0.01	0.06	0.1351
5	112	0.23	0.02	0.09	**0.0063**	111	0.03	0.00	0.08	0.3478
6	159	-0.04	-0.00	0.08	0.6864	159	0.13	0.01	0.07	0.0343
7	144	0.19	0.02	0.09	**0.0146**	144	0.00	0.00	0.08	0.4750
8	519	-0.01	-0.00	0.04	0.5746	517	0.08	0.01	0.04	0.0259
9	427	0.07	0.01	0.05	0.0849	425	-0.01	-0.00	0.05	0.6032
10	57	-0.19	-0.01	0.13	0.9291	57	0.01	0.00	0.12	0.4783
11	331	0.02	0.00	0.05	0.3721	328	0.00	0.00	0.05	0.4812
12	21	0.08	0.00	0.21	0.3486	21	-0.08	-0.00	0.22	0.6508
13	80	0.10	0.01	0.10	0.1623	79	0.00	0.00	0.10	0.4885
14	166	0.00	0.00	0.08	0.4804	165	0.02	0.00	0.07	0.3795

Table 5.18: Louvain clustering Educational attainment phenotype. Sets 2,5 and 7 would be of interest but none are significant in the replication cohorts.

RProjects/utils2/R/collate_results/generic_compile_gsa_lourvain.R

ID	Name	n	PSP	p	q FDR B.H
GO:0030165	PDZ domain binding	26	58	6.01×10^{-21}	2.75×10^{-18}
GO:0022803	passive transmembrane transporter activi	40	179	5.70×10^{-19}	8.70×10^{-17}
GO:0015267	channel activity	40	179	5.70×10^{-19}	8.70×10^{-17}
GO:0038023	signaling receptor activity	33	136	8.38×10^{-17}	7.67×10^{-15}
GO:0060089	molecular transducer activity	33	136	8.38×10^{-17}	7.67×10^{-15}
GO:0008066	glutamate receptor activity	15	21	1.20×10^{-16}	9.14×10^{-15}
GO:0022857	transmembrane transporter activity	47	292	3.49×10^{-16}	2.24×10^{-14}
GO:0046873	metal ion transmembrane transporter acti	34	152	3.97×10^{-16}	2.24×10^{-14}
GO:0022836	gated channel activity	28	100	4.40×10^{-16}	2.24×10^{-14}
GO:0005216	ion channel activity	35	163	5.33×10^{-16}	2.44×10^{-14}

Table 5.19: community 7 Louvain GO: Molecular Function 91 significant terms

RProjects/group_sizes/R/louvain_enrichment/group7_molecular_function.R

ID	Name	n	PSP	p	q FDR B.H
C0524528	Pervasive Development Disorder	22	113	2.87×10^{-9}	8.05×10^{-6}
C0004352	Autism Disorder	36	285	5.26×10^{-9}	8.05×10^{-6}
C0036341	Schizophrenia	36	291	9.36×10^{-9}	9.56×10^{-6}
C1510586	Autism Spectrum Disorders	35	300	7.44×10^{-8}	5.70×10^{-5}
C1263846	Attention deficit hyperactivity disorder	23	152	1.82×10^{-7}	1.11×10^{-4}

Table 5.20: Community 7 Louvain: Disease term enrichment using ToppGene. Seventy-four significant terms found showing five most enriched.

5.1.10 Summary

Community detection has proven to be useful combined with GSA in the analysis of GWA studies of intelligence and educational attainment. A cohesive community (community five) was found using spectral community detection that has an enriched association in discovery and replication samples for differences in intelligence and educational attainment. This community is one of the most cohesive (lowest conductance figure 4.8). Community five has enriched Gene Disease Associations for schizophrenia and depressive disorder and its components are enriched for murine models of abnormal long term potentiation. Community five has a number of glutamate receptors and g protein receptors but the enrichment signal arises from the neighbouring proteins.

5.2 Replication in larger samples: Intelligence 2 and EA3

After completing the initial analysis two large cohorts became available that Incorporated data from our samples but enlarged them allowing us to see if our findings remained valid as the number of individuals involved in the studies increased.

5.2.1 Intelligence

Savage et al. [367] reported a GWA study of intelligence involving 269,867 individuals and allowed the identification of 1016 genes of which 939 are newly reported. This is the second population sample produced by the Complex Trait Genomics lab that conducted the analysis that formed the $\text{Intelligence}_{\text{Replication}}$ sample and is referred to by them as Intelligence 2.

The largest contributor of subjects to the study was UKBiobank (n=195,653) with a significant contribution from the Cognitive Genomics consorTium (COGENT) (n=35,289) [242]. There is therefore a significant overlap with our discovery cohort and we do not have access to samples that do not overlap ours. It does however represent a significant increase in sample size and we hypothesise our findings if valid would increase in significance with the increase in sample size.

5.2.2 Education

In the same way that the most recent intelligence sample is an extension of the sample used for the $\text{Intelligence}_{\text{Replication}}$ sample the education replication sample was mostly made from the EA2 analysis (Educational Attainment 2)[309]. More recently summary statistics have been made available for the EA3 samples made up of over 1 million individuals (n=1,131,881)[237]. The analysis identified 1,271 genome wide significant SNP's and polygenic risk scores based on the data were able to explain 11-13% of the variance in educational attainment and 7-10% of intelligence.

5.2.2.1 Methods

Gene scores and gene set analysis were carried out as detailed in chapter 2 and the spectral community gene sets analyses using MAGMA GSA. The summary statistics

were downloaded from the same website of the Complex Trait Genomics group that the Intelligence Replication sample had been previously[3].

Summary statistics for the EA3 analysis were downloaded from the SSGAC [4]. The summary statistics again excluded data from individuals in 23andMe.

5.2.2.2 Results

MAGMA GSA results are shown in table 5.21. Group 5 again shows significant enrichment (p=0.00267) for educational attainment and is also enriched in the intelligence sample. It is interesting to note that group 20, which was found to be significantly enriched only in the PASCAL analysis of educational attainment (table 5.6) also shows significant enrichment in the Educational attainment (EA3) sample (p=0.00096). Subgroup analysis of group 5 is shown in table 5.22.

Group 5 is also enriched for intelligence (p=0.00004) in the Intelligence wave 2 Savage sample. Although there is enrichment of group 20 in the large educational attainment sample EA3, group 5 has been consistently enriched in all samples and shows increased enrichment with sample size.

Subgroup analysis again shows that the signal is not confined to the glutamate receptor although for the larger intelligence sample (Intelligence 2) these receptors were responsible for the majority of the signal. In the larger educational attainment sample (EA3) the remaining proteins again had a greater level of enrichment than the glutamate receptors (p=0.00014 vs 0.00267 table 5.22).

Analysis of the Louvain clustering algorithm which had shown promise in the previous section showed enrichment for group 7 (glutamate associated) in both the larger intelligence and educational attainment samples (p=0.0002 Intelligence (Savage), p=0.0017 EA3 table 5.23 and table 5.24). Group 7 was the most enriched of the Louvain method groups in the intelligence discovery and replication samples (p=0.0012 and p=0.0026 table 5.17).

Using the largest cohorts available for educational attainment we find enrichment of Louvain group 7 (p=0.0017) and group 8 (p=0.0003). Louvain group 8 did not appear of interest in the earlier analyses but is enriched in the larger intelligence sample (Savage p=0.0003) and educational attainment sample (see table 5.23). The overall picture is of increasing significance of communities of interest as the sample size increases.

[3] https://ctg.cncr.nl/documents/p1651/SavageJansen_IntMeta_sumstats.zip
[4] https://www.dropbox.com/s/ho58e9jmytmpaf8/GWAS_EA_excl23andMe.txt?dl=0

	EA3 Educational Attainment					*Savage 2018 Intelligence*				
Community	*n*	β	β STD	SE	*p*	*n*	β	β STD	SE	*p*
1	161	0.20	0.02	0.10	0.02255	169	0.03	0.00	0.08	0.37667
2	153	0.02	0.00	0.10	0.42371	161	-0.03	-0.00	0.08	0.63671
3	39	-0.00	-0.00	0.19	0.50666	41	0.10	0.00	0.15	0.24404
4	29	-0.05	-0.00	0.21	0.59064	29	-0.40	-0.02	0.19	0.98326
5	106	0.36	0.03	0.13	**0.00267**	111	0.43	0.03	0.11	**0.00004**
6	76	-0.06	-0.00	0.15	0.64806	77	0.00	0.00	0.13	0.49470
7	125	-0.05	-0.00	0.11	0.68003	129	0.16	0.01	0.09	0.04848
9	61	-0.09	-0.01	0.15	0.72539	62	0.09	0.01	0.13	0.23801
10	146	0.04	0.00	0.10	0.35081	152	0.19	0.02	0.08	0.00819
11	55	0.26	0.01	0.16	0.05692	56	-0.05	-0.00	0.13	0.65762
12	32	0.06	0.00	0.20	0.39111	34	0.19	0.01	0.17	0.13704
16	68	0.05	0.00	0.15	0.37751	72	-0.19	-0.01	0.13	0.92713
17	144	0.06	0.01	0.10	0.27056	146	0.16	0.01	0.08	0.03087
19	20	-0.08	-0.00	0.22	0.63438	20	0.04	0.00	0.19	0.41751
20	144	0.33	0.03	0.11	0.00096	149	0.17	0.01	0.09	0.03405
22	42	0.16	0.01	0.19	0.19590	42	0.11	0.01	0.16	0.24855
23	67	0.09	0.01	0.15	0.29214	71	-0.03	-0.00	0.13	0.58327
24	74	0.27	0.02	0.15	0.03782	76	0.12	0.01	0.13	0.17961
25	26	-0.09	-0.00	0.23	0.65285	27	0.17	0.01	0.20	0.20409
26	116	0.10	0.01	0.12	0.18681	120	0.16	0.01	0.10	0.05441
28	35	0.35	0.02	0.21	0.04473	38	0.33	0.02	0.18	0.03277
32	50	-0.03	-0.00	0.18	0.57302	52	0.15	0.01	0.16	0.16530
33	180	0.25	0.02	0.10	0.00473	189	0.19	0.02	0.08	0.01065
34	209	0.02	0.00	0.09	0.42670	214	0.04	0.00	0.08	0.31938
43	137	0.12	0.01	0.11	0.12573	144	0.11	0.01	0.09	0.10033
44	100	-0.01	-0.00	0.12	0.52754	103	-0.03	-0.00	0.10	0.62441
45	118	0.02	0.00	0.11	0.43314	122	0.12	0.01	0.09	0.09530
46	47	0.14	0.01	0.16	0.18425	49	0.02	0.00	0.13	0.42924
47	23	-0.04	-0.00	0.25	0.56587	26	-0.15	-0.01	0.22	0.75716
48	22	0.07	0.00	0.26	0.39749	23	-0.10	-0.00	0.24	0.65749
51	110	0.04	0.00	0.12	0.35718	111	0.04	0.00	0.09	0.34068
53	386	0.05	0.01	0.06	0.18562	397	0.11	0.02	0.05	0.01655
55	42	0.27	0.01	0.18	0.06464	43	0.32	0.02	0.15	0.01828
58	36	0.04	0.00	0.21	0.42733	37	-0.21	-0.01	0.18	0.87947
62	34	-0.13	-0.01	0.20	0.74314	34	-0.14	-0.01	0.17	0.79092

Table 5.21: MAGMA GSA Education and Intelligence gene set analysis for larger samples EA3 (Educational attainment) and Savage 2018 (Intelligence). β is the regression coefficient of the gene set. SE is the standard error of the regression coefficient. β STD the the semi-standardized regression coefficient, corresponding to the predicted change in Z-value given a change of one standard deviation in the predictor gene set / gene covariate

RProjects/chapter_community_detection/R/magma_gsa/rawread/larger_sample_gsa_new_raw_table.R

	EA3 Educational Attainment					*Savage 2018 Intelligence*				
Community	*n*	β	β STD	SE	*p*	*n*	β	β STD	SE	*p*
remaining	61	0.62	0.04	0.17	0.00014	65	0.42	0.02	0.14	0.00125
glutamate	16	0.34	0.01	0.35	0.16798	17	1.19	0.04	0.30	0.00004
gprotein	38	-0.06	-0.00	0.21	0.60453	38	0.29	0.01	0.19	0.06316
group 5	106	0.36	0.03	0.13	0.00267	111	0.43	0.03	0.11	0.00004
gprotein & glutamate	45	0.02	0.00	0.20	0.46941	46	0.44	0.02	0.17	0.00543

Table 5.22: MAGMA GSA Educational attainment (EA3) and Intelligence (Savage 2018) subgroup analysis. Glutamate is genes in group 5 that appear in GO:0007215 Glutamate signalling pathway, gprotein is genes in group 5 found in GO:0007215 Glutamate signalling pathway. Remaining refers to the proteins not in the gprotein or glutamate setsβ is the regression coefficient of the gene set. SE is the standard error of the regression coefficient. β STD the the semi-standardized regression coefficient, corresponding to the predicted change in Z-value given a change of one standard deviation in the predictor gene set / gene covariate

RProjects/chapter_community_detection/R/magma_gsa/rawread/largersample_sub_group_gsa_new_raw_table.R

Community	n	β	β STD	SE	p
1	196	0.11	0.01	0.0897	0.1185
2	334	0.06	0.01	0.0660	0.2021
3	510	0.08	0.01	0.0521	0.0740
4	245	0.10	0.01	0.0792	0.0943
5	111	0.09	0.01	0.1110	0.2132
6	159	0.01	0.00	0.0958	0.4421
7	144	0.32	0.03	0.1080	0.0017
8	519	0.19	0.03	0.0558	0.0003
9	425	0.04	0.01	0.0640	0.2426
10	57	0.03	0.00	0.1570	0.4303
11	331	0.08	0.01	0.0679	0.1087
12	21	-0.11	-0.00	0.2620	0.6587
13	80	-0.12	-0.01	0.1330	0.8143
14	166	0.09	0.01	0.0989	0.1692

Table 5.23: Enrichment of communities for educational attainment using MAGMA GSA. Louvain community detection PSP EA3 Education sample

Community	n	β	β STD	SE	p
2	345	0.11	0.02	0.0550	0.0208
3	527	0.11	0.02	0.0431	0.0055
7	153	0.32	0.03	0.0899	0.0002
8	537	0.14	0.02	0.0475	0.0018

Table 5.24: Enrichment of communities for intelligence using MAGMA GSA. Louvain community detection PSP Savage 2018 Intelligence phase2 cohorts

Chapter 6

Discussion

The two aims of this thesis were to investigate the feasibility of combining network-based analysis of the synaptic proteome with Gene Set Analysis and discover general principles applicable to such analyses. The second was to use this method to understand better a specific complex trait (intelligence and educational attainment), which seemed appropriate for such analysis (see section 1.3).

My first hypothesis was that particular network communities would be differentially enriched for genetic differences associated with intelligence or educational attainment. The second was that important, central genes in the network would be more associated with differences in intelligence and educational attainment.

This chapter will assess whether this has been achieved and summarise the findings of the thesis. In doing so, I will discuss the strengths and limitations of this analysis and outline future work. I will conclude with some more speculative observations on this work informed by my experience in carrying out this analysis.

6.1 Principal findings

6.1.1 Intelligence and Educational Ability

The principal finding is that a community (community five) detected using the spectral clustering method, containing metabotropic glutamate receptors and ionotropic AMPA receptors, was found to be enriched for differences in intelligence and educational ability (section 5.1.4). The enrichment is present in the discovery, and replication samples for both intelligence and educational attainment and the findings become stronger using samples from more recent studies that overlap with our discovery and replication

samples but provide a larger sample size.

The part of the community responsible for the enrichment signal is not the receptors themselves which are implicated in animal models of learning, but their neighbours, proteins that contribute to the macromolecular complexes the receptors are embedded in. The results mirror Lundy et al.[259] who used a tissue-specific network to examine long QT syndrome and found that the significant genes were amongst the neighbours of the receptors.

Analysis of the most recent studies (Savage et al. and EA3)[367],[237] (section 5.2.2.2) show enrichment both of glutamatergic receptors and their neighbours, but a substantial amount of the enrichment remains in the neighbouring proteins. These proteins would not be identified as part of a standard gene set analysis approach and support network-based community gene set analysis as a method of analysing complex traits.

The enriched proteins reveal few common biological themes using Gene Ontology enrichment analysis save for their proximity to receptor proteins that are known to participate in glutamatergic neurotransmission. The *combination* of community-based GSA and Gene Ontology analysis may be combined profitably to interrogate better ontology or topology defined areas of enrichment in complex traits.

The second-best performing algorithm (Louvain) also identified a similar group of proteins related to glutamatergic receptors that showed enrichment in the intelligence samples suggesting that assessing algorithms on both performance on the LFR benchmark and PSP was useful.

6.1.2 Vertex statistic and intelligence

There was no association between a genes' centrality and its association with intelligence or educational attainment on MAGMA GWGAS (section 3.13.3). Lee et al[238] had previously found that complex traits were more likely to occur in hubs and bottlenecks. This analysis was limited to genome-wide significant SNPs in a wide range of disorders. This thesis finds no evidence to support an effect of network centrality on intelligence or educational attainment using the most up to date data and a tissue specific network. This is consistent with intelligence being the product of many genes of small effect and is in contrasts with diseases including cancer and neuro-degenerative disorders that are highly enriched in genes with high eigenvector centrality, and degree centrality even compared to the rest of the PSP.

Cancer genes have previously been reported to be central[419],[317] in the proteome which my findings support. The finding that neurodegenerative disorders are associated with high centrality is limited to an over-representation of gene-disease associations terms but to my knowledge, an association between neurodegenerative disorders and synaptic proteome centrality has not been previously reported. Further studies would be required to confirm if these findings can be replicated in GWA studies of these conditions. Previous studies have suggested that disease genes are not central[26] in the proteome. This finding would suggest that at least some disorders may have more central genes.

There is a correlation between centrality measures and loss of function intolerance and an association between loss of function intolerance and intelligence and educational ability (although the association with educational attainment has been recently reported[206]). This association is more marked in PSP genes than the rest of the genome. These factors (centrality and loss of function intolerance) may interact to produce the overall phenotype.

Why was there no association between gene-level GWA study p values for educational attainment or intelligence and measures of vertex importance such as degree? It may be that genetic changes in nodes of high degree result in a severe phenotype (as suggested by disease term enrichment section 3.13) and are not represented in population studies of complex traits. We would hypothesise that these core nodes are necessary for basic functions, and the finding further supports that central nodes are associated with an increased measure of evolutionary constraint (both by DEG and pLI). In this thesis, I show that genes associated with intelligence and educational attainment are associated with increased levels of genetic constraint compared to the rest of the PSP and that high centrality nodes have high levels of selection pressure. The thesis extends the evidence for what might be termed centrality-severity to human populations using the DEG as a nominal measure and Gnomad as a quantitative measure. The PSP is under a remarkable degree of genetic constraint compared to the rest of the genome, and here provide evidence of this from large scale human genomic studies.

Central genes are enriched for gene ontology terms associated with basic cellular functions such as the cell cycle (table 3.28), and it seems reasonable that in turn, they are associated with diseases that typically involve such processes (viral infection, cancer, etc.). The neurological disorders affected by these are primarily neurodegenerative, and this enrichment is against the stringent criteria of correcting for multiple comparisons using all diseases rather than only neurological ones. This suggests that

these disorders may show common features in terms of network architecture. It may be that these disease are associated with genes under genetic constraint (high centrality) as they are later in onset.

In turn, the most peripheral nodes are those associated with the plasma membrane and extracellular matrix and these, in turn, are associated with epilepsy. This suggests that there may exist a common network architecture for such disorders and that network architecture may be another means to subdivide neuropsychiatric traits. Caution should be applied, however given my previous statements about gene-disease databases. Low centrality genes have not to my knowledge been studied. This suggests a different structure to the PSP network than that found in ecological or financial networks where an outer (expendable) shell protects an inner core[63]. The outer shell (kcore) are responsible for vital day to day functions; it is just that the inner core is required for *any* coherent function.

6.2 GSA and Communities

These results suggest that genetic variation in the genes encoding a group of proteins forming a tightly interconnected sub-network in the postsynaptic proteome have an enriched association with differences in educational attainment and intelligence. This group contains the majority of the genes encoding glutamate receptors and heterotrimeric G protein in the postsynaptic proteome, and so it appears to be a homogeneous unit with a common function. The community is indeed one of the most cohesive as measured by conductance. These genes are commonly associated with animal models of learning and memory and intuitively appear to have a role in human learning and intelligence. Glutamate mediated neurotransmission is required to initiate long-term potentiation, and blocking glutamate receptors in the hippocampus of mice causes failure of memory encoding. More recently, differential expression of glutamate receptors has been shown to be associated with differences in problem-solving in two species of wild finch[19] which appears even more analogous to human intelligence.

The glutamate receptors and heterotrimeric G proteins are not the principal sources of the enrichment signal; it is instead their neighbouring and supporting proteins. The potential connection between the GWA study findings and these receptors is made possible by network analysis of the community structure of the synaptic proteome.

This thesis is, to my knowledge, the first to use GSA in association with community detection and use the complete data present in GWAS summary results. Mooney and

Wilmot proposed using communities found in protein interaction networks as gene sets for GSA[282] and gave as examples a modularity maximisation method and an annotation mining tool that incorporates module detection. The first did not use genetic data [440] and the second implements an over-representation test and does not use the full amount of data from GWAS[87]. Ghiassian et al.[144] used the link community[3],Louvain[47] and Markov Clustering[113] algorithms to detect communities in the STRING protein database and concluded that communities were not associated with diseases. Their analysis did not, however, use a tissue-specific network or a competitive GSA test. The GWA study data was limited to adding genes with significant SNPs in GWA catalog[144] to a Gene Disease Association list for over-representation analysis. It did not use the full polygenic signal information in GWA study summary statistics.

In this thesis, I have used a protein-protein interaction network specific to a cellular component believed implicated in the trait (the synapse and intelligence[177]). Goh et al. found that genes specific to tissues were more likely to be associated with diseases than others in the connectome [151], it seems a reasonable approach to take, and similar approaches have been successful in other systems[259]. This technique could be extended to other disorders or complex traits that may be mediated by synaptic function. However, the synapse may behave differently from other tissues in these forms of analyses, as we know it forms modular functional units from its proteins analogous to communities[155] our findings might not generalise to proteomes in non-synaptic tissue or metabolic pathways.

This work extends that of Pocklington[338], McLean[266], Bayes[35] and coworkers who have found differential enrichment of disease terms in network components in the synapse. Disease term enrichment based on disease-gene association contains heterogeneous data, and the ability to combine the information in the summary data of the GWA study with an analysis of network structure represents an improvement in these methods. Future extensions may include rare variants found in exome studies or network fusion studies incorporating information from other sources such as gene expression and eQTL.

Lamperter et al. [228], in their paper describing the Pascal software, pointed to the desirability of methods that can generate gene sets from primary experimental data[228]. Community and network analysis is an example of such a technique.

The genes in the region of the affected groups could be subject to experiment; one could see if they and the others had a role in animal learning or caused changes in

synaptic electrophysiology.

6.2.0.0.1 Limitations There are limitations to the current study. The set of genes found in proteomic studies of the postsynaptic areas continues to increase, although the increase has reached a plateau[173]. If the network changes, then the community allocations will change, and a different network may lead to different community allocations. The network structure itself will have errors in databases and experiments and collation. There are several principled approaches to network errors that could be adopted in future work[289].

The community allocations, in addition, are not a globally optimal allocation concerning modularity, and alternative partitions with similar modularity exist.

The clustering also does not consider that nodes may belong to more than one component or cluster. Probabilistic methods exist to determine the most likely community membership[443] if vertices can belong to more than one community but the loss of non-overlapping gene sets may result in a great increase in the number of sets. We have also not considered a model where some genes are assigned to communities while others remain unassigned.

That communities may be topologically good and yet not represents ground truth an entity is an important point [325],[134] and several partitions may have similar modularities[325]. In addition different community detection algorithms discover different communities. I do not believe communities represent exact natural entities (they cannot overlap, they have only one component) but are a *useful* mapping between network structure and genetic architecture. One approach to the existence of multiple plausible partitions would be to use node metadata. I have not used any data other than the distribution of edges to perform community detection. However, there are principled approaches[297],[270] that use metadata and this may allow the best of a number of partitions to be selected. An overlooked advantage of modularity based methods however, compared to some of the network analysis shown described in section 1.6 is that given the edge list and community allocations one can confirm the modularity of a network and high modularity partitions are unlikely to occur by chance.

A further limitation of the study is that gene-based analysis discards approximately 50% of the information from SNPs that do not lie between the boundaries of protein-encoding genes. Future work may incorporate this information by, for example, using genetic variants known to affect genes such as eQTLs[424],[140].

The use of gene ontology terms as gene sets has been widespread as part of the

secondary analysis of GWA studies; however, these ignore gene ontology structure and overlap. This results in a significant multiple testing penalty which in the case of GWA studies of educational attainment and intelligence has been overcome by sample size, but for other phenotypes, it may be worth looking again at methods that use the tree structure of Gene Ontology terms[10]. Specialised ontologies such as Neural /Immune[141] are available but are non-trivial to implement, and to my knowledge, there are no readily available packages that have this as a 'drop in' solution.

Any enrichment software should provide the ability to provide one's own background set. The recent addition of this feature in ToppGene[73] led me to move all of my analysis to this; had it been available earlier would have saved me a considerable amount of time.

A small world model of genetic influence networks has been argued to support an Omni-genic model of traits[54]. Network structure may still affect the genetic architecture of complex traits; even in an undirected network with unweighted edges and a small world structure, the flow rate of information between nodes (particularly over short timescales) depends on network features such as community structure (this is the basis of methods including Infomap).

The most obvious extension of this work would be to apply the systematic method outline to other neuropsychiatric disorders for which large scale GWA studies exist such as schizophrenia. Carrying out the study required writing numerous scripts. FUMA[428] has gained widespread use in the secondary analysis of GWA studies by providing an easy to use interface to carry out gene level tests and enrichment. Encapsulating the methods and data used here into one application would allow wider adoption of these techniques.

Further directions in the analysis of the synaptic proteome would include looking at how its structure changes over time. Intelligence and educational attainment are relatively stable over the life course but other neuropsychiatric disorders that involve similar genes occur at different times in the life course (compare, for example, schizophrenia and Autism Spectrum Disorder). There may be network features that emerge at critical periods (due to changes in gene expression, for example). We also know that the proteome varies across the brain[159]

The study provides an outline for a network analysis of GWA studies of neuropsychiatric disorders with synaptic involvement. An analysis of vertex properties with gene-level results, an analysis of enrichment core or periphery, community enrichment analysis and topological subgroup analysis.

6.3 Study design

I have included several different methods of assessing enrichment (GSEA, PASCAL and MAGMA) for two reasons. One was so findings would not be dependent on a particular method. The other was to limit the number of 'interesting' sets taken forward for replication and to ensure these were more likely to replicate. One option would have been to reduce the significance level, but this seemed somewhat arbitrary.

With increasing sample sizes in the most recent studies (see chapter 5), the significance level increases relative to the penalty for multiple comparisons, but not all disorders will have studies that include 1.1 million participants. For example, a meticulous study of the genetic epidemiology of drug-induced psychosis and its progression to schizophrenia between 1997 and 2015 using Swedish national registries (which are of very high quality) achieved a sample size 7,606[210]. Therefore, a careful study design will be necessary to limit the number of potential comparisons that arise testing network communities. Perhaps it would be reasonable practice also to 'lookup' previously discovered gene sets in future studies as is the case for SNPs in GWA studies before testing all Gene Ontology terms.

6.3.1 Benchmarks

Benchmarks of clustering algorithm performance were helpful in choosing an algorithm but only in combination with performance on the PSP network. Yang et al. [446] provide a helpful review of the performance of the igraph algorithms on the LFR benchmark. The way in which algorithms behave as they approach resolution limits may be important.

The LFR benchmark remains the best available, but it is not easy to fit all network structures. Ultimately, a probabilistic benchmark that considers features such as overlapping communities, assortativity, and core-periphery structure would be desirable but challenging to implement.

Nevertheless, there were essential differences between the performance of algorithms on the LFR benchmark and the PSP graph. The infomap algorithm, in particular, was the best on the benchmark but gave rise to unrealistic community structures. Testing algorithms on one's network to ensure that they give good results seems a sensible precaution.

Appendix A

Abbreviations

AMPA - α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid

ARC - Activity-regulated cytoskeleton associated protein

ASD - Autistic Spectrum Disorder

B H - Benjamini-Hochberg Procedure

BS - Bootstrap

CAGE - Cognitive Ageing Genetics in England and Scotland

CCACE - Centre for Cognitive Ageing and Cognitive Epidemiology

CDM - Community detection software implementing spectral clustering

CaMKII - Calcium/calmodulin-dependent protein kinase II

CHARGE - Cohorts for Heart and Aging Research in Genomic Epidemiology

ClinVar - Database of genomic variation and effect on health (NCBI hosted)

CTD - Comparative Toxicogenomics Database

COGENT - Cognitive Genomics consorTium

dbGAP - The database of Genotypes and Phenotypes

dbSNP - Single Nucleotide Polymorphism Database

DEG - Database Essential Genes

DEG - Differentially expressed genes

df - degrees of freedom

DisGeNET - database of disease gene associations

eQTL - Expression quantitative trait loci

ExAC - The Exome Aggregation Consortium

FDR - False Discovery Rate

FG - Fast Greedy

kB - kilobase

LD - Linkage disequilibrium

LEC - Leading eigenvector clustering (default spectral method in igraph)

LFR - Lancichinetti–Fortunato–Radicchi benchmark

LTP - Long term potentiation

GABA - Gamma aminobutyric acid

GATES - a modification of the Simes test

GCTA - Genome wide complex trait analysis

gml - graph modelling language

gmx - Gene matrix file format

gnomAD - Genome Aggregation Database

GN - Girvan Newman benchmark

GRIA - Glutamate Ionotropic Receptor AMPA Type

GRIK - Glutamate receptor, ionotropic, kainate

GSA - Gene Set Analysis

GSEA - Gene Set Enrichment Analysis

GPCRs - G-protein-coupled-receptors

GTEx - Gene Tissue Expression project

GWAS - Genome Wide Association Study

GWGAS -genome-wide gene association analysis (GWGAS)

hPSD - human Post Synaptic Density

KGG - Knowledge-Based Mining System for Genome-wide Genetic Studies

iGluR - ionotropic glutamate receptor

Ma - Mega-Annum (million years)

MAGUK - Membrane Associated Guanylate Kinase

MAGMA - Multi-marker Analysis of GenoMic Annotation

MC - Monte Carlo

MCL - Markov Clustering

MDD - Major depressive disorder

mGLU - metabotropic glutamate

MTAG - Multi trait analysis of genome wide association studies

NCBI - National Centre for Biotechnology Information

NMDA - N-methyl-D-aspartate

NMI - Normalised Mutual Information

NRC - NMDA receptor complex

NRC/MASC - NRC/MAGUK associated signalling complex

OMIM - Online Mendelian Inheritance in Man database

PASCAL - Pathway scoring algorithm; a Gene set analysis and gene level scoring package.

PPI - Protein-protein interaction

PPIN - Protein-protein interaction network

PSD - Postsynaptic density

PSD-95 - postsynaptic density protein 95

PSI-MI - Proteomics Standard Initiative Molecular Interaction

PSP - Post synaptic proteome

SBM - Stochastic Block Model

SG - Spinglass (SgG1 - Spinglass with gamma of one)

SNP - Single nucleotide polymorphism

STP - Short term potentiation

STRING - A protein interaction database. Originally Search Tool for Recurring Instances of Neighbouring Genes

TPM - Transcripts per million

VEGAS - versatile gene-based association study; software to obtain gene level scores

WT - Walktrap clustering

UKBB - UK Biobank

VNR - Verbal-numerical reasoning

Appendix B

Centrality measures - Supplemental

B.1 Correlation of centrality measures

The global positive correlation between transitivity and centrality measures is largely caused by nodes of transitivity 0 or 1 (see section B.1.3.1) particularly when isolates are treated as 0 ($C_{WS}0$) as the lowest degree nodes ($k = 1$) have zero transitivity by definition.

This is at variance with the view of local transitivity as a local betweenness where a node is central if transitivity is *low* (see section 3.2.8.3) and can be seen clearly if we test correlation for nodes of degree five or greater (table B.1).

In addition it further obscures the relationship between transitivity and degree ($C(k)$ and k - section B.1.2) reported in biological networks, where high degree nodes have lower local clustering [4]. Using C_i^{NaN} has little effect on the correlation with other centrality measures or with gene statistics (table 3.9).

The most correlated centrality measures were degree and kcore (0.986). Excluding transitivity and kcore which are not amongst the canonical centrality measures [308][1] the most correlated were eigenvector and closeness (0.969). This is higher than that reported in a collection of social science networks (0.63 in Valente et al[409] - see table B.2 and B.3)

See Oldham[310] for recent work on correlation across networks. Table B.2 shows the results of the correlation between degree, eigenvector centrality, closeness and betweenness found by Valente et al.[409] for 58 sociometric network datasets. Centrality

[1] Four measures of centrality that are widely used in network analysis are degree centrality, betweenness centrality,closeness centrality,and eigenvector centrality

measures appear more correlated in the PSP than in those reported by Valente[2] particularly closeness and betweenness (0.44 vs 0.782) and degree and closeness (0.66 vs 0.85; see table B.3) and persists even if lower degree nodes are removed from the calculation (table B.1) .

Whether the higher correlation of closeness and other centrality measures may be a feature found in biological networks compared with social networks (as for example assortativity) is unclear.

B.1.0.1 Principal component analysis of centrality measures

Principal component analysis was carried out to create a low dimensional representation of the relationship between centrality measures and in constructing a linear model of their effect upon intelligence and educational attainment gene scores.

Principal component analysis of centrality variables was carried out using the R command `prcomps`[3] with the centralities scaled to unit variance and all other arguments default. All centrality measures (six) other than transitivity NA were included (this was removed to avoid NA handling). The majority of the variance is explained by the first three principal components (93.4%) (figure B.1 and table B.4). Transitivity and eigenvector centrality appear almost orthogonal in a plot of the first two dimensions (figure B.2).

[2] although Valente[409] uses Pearson's product moment correlation

[3] which uses SVD

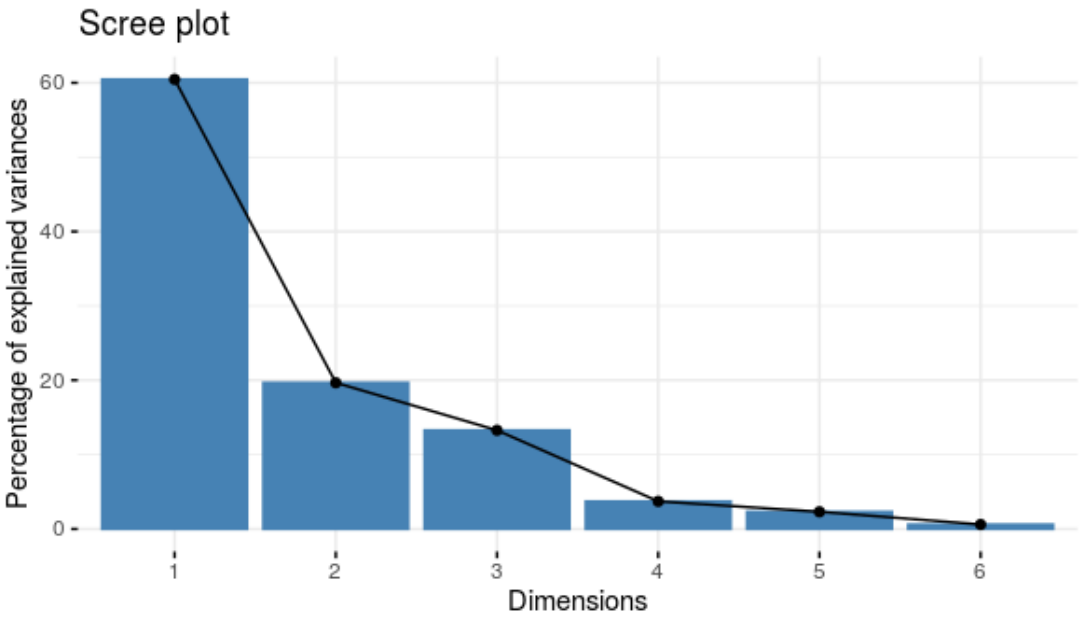


Figure B.1: Screeplot of the principal components of six centrality measures (table 3.3). TransitivityNA has been omitted to avoid dealing with NA.

RProjects/chapter3/R/centrality_ch3/prcomps/3_plotPCA.R

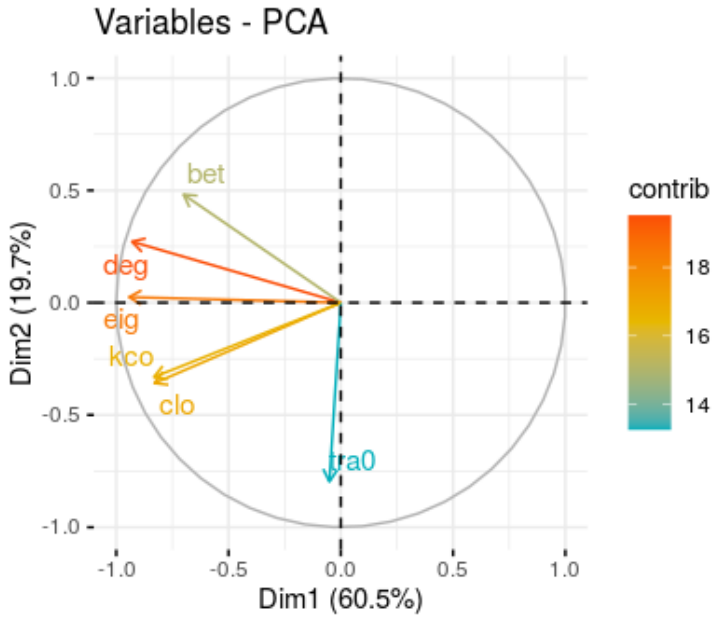


Figure B.2: Contribution of different variables. deg = degree, eig = eigenvector centrality, kco = kcore index, tra0 = transitivity with isolates set to zero, clo = closeness, bet = betweenness centrality.

RProjects/chapter3/R/centrality_ch3/prcomps/3_plotPCA.R

Rank	Centrality 1	Centrality 2	ρ	$\rho_{k>=5}$	Rank $k >= 5$
1	transitivity0	transitivityNA	1.00	1.00	1
2	kcoreness	degree	0.99	0.96	2
3	closeness	eigenvector	0.97	0.94	3
4	kcoreness	eigenvector	0.91	0.88	4
5	betweenness	degree	0.90	0.83	5
6	kcoreness	closeness	0.87	0.82	6
7	eigenvector	degree	0.88	0.81	7
8	closeness	degree	0.85	0.77	8
9	kcoreness	betweenness	0.85	0.71	9
10	closeness	betweenness	0.78	0.65	10
11	eigenvector	betweenness	0.76	0.60	11
12	transitivity0	eigenvector	0.46	0.17	13
13	transitivityNA	eigenvector	0.33	0.17	12
14	transitivity0	closeness	0.43	0.12	15
15	transitivityNA	closeness	0.29	0.12	14
16	kcoreness	transitivity0	0.43	0.06	17
17	kcoreness	transitivityNA	0.26	0.06	16
18	transitivity0	degree	0.38	-0.09	19
19	transitivityNA	degree	0.20	-0.09	18
20	transitivity0	betweenness	0.21	-0.31	21
21	transitivityNA	betweenness	-0.00	-0.31	20

Table B.1: Correlation of centrality measures. Spearman. Degree five or more. Ordered by correlation. Method for missing data include only pairwise complete. TransitivityNA treats isolates as NA and removes from calculation. Transitivity0 uses the option to set isolates to 0. There is minimal perturbation of the rank order but the relationship between transitivity and betweenness and degree become negative.

RProjects/db_essential_genes/R/get_corr/cor_centrality_measures_ordered_cortable_single_long_table.R

	Degree	Eigenvector centrality	Closeness	Betweenness
Degree	1	0.92	0.66	0.85
Eigenvector centrality	0.92	1	0.63	0.72
Closeness	0.66	0.63	1	0.44
Betweenness	0.85	0.72	0.44	1

Table B.2: Correlation of centrality measures for 58 sociometric network datasets from Valente et al [409]. Pearson correlation coefficient. Values are for symmetrised closeness and betweenness centrality. Eigenvector centrality is necessarily symmetrised. Range of average network size 45-83. Some studies had maximal nominations for degree (i.e there was an artificial upper bound using a survey).

	Degree	Eigenvector centrality	Closeness	Betweenness
Degree	1	0.92 (0.88)	0.66 (0.851)*	0.85 (0.902)*
Eigenvector centrality	0.92 (0.88)	1	0.63 (0.969)*	0.72 (0.762)
Closeness	0.66 (0.851)	0.63 (0.969)	1	0.44 (0.782)*
Betweenness	0.8 (0.902)	0.72 (0.762)	0.44 (0.782)	1

Table B.3: Correlation of centrality measures for 58 sociometric network datasets from Valente et al.[409]. Pearson correlation coefficient. Values are for symmetrised closeness and betweenness centrality. Eigenvector centrality is necessarily symmetrised. Range of average network size 45-83. Some studies had maximal nominations for degree (i.e there was an artificial upper bound using a survey). Comparison added with PSP * shows where PSP notably higher correlation.

Dimension	Eigenvalue	Variance (%)	Cumulative variance (%)
Dim 1	3.63	60.46	60.46
Dim 2	1.18	19.66	80.12
Dim 3	0.80	13.26	93.38
Dim 4	0.22	3.71	97.09
Dim 5	0.14	2.31	99.40
Dim 6	0.04	0.60	100.00

Table B.4: Eigenvalue of dimensions of PCA for centrality measures with variance.

RProjects/chapter3/R/centrality_ch3/prcomps/5_PCA_eigval.R

B.1.1 Supplemental degree distribution

The plot of log-degree against the log of the cumulative distribution function (CDF) of the degree distribution is shown in figure B.5b and B.4. The graph forms a straight line for most of the distribution with degree greater than x_{min} [4]. In figure B.4 the fit of the power law distribution is shown in red.

The lines of fit of a power-law, log-normal, exponential and Poisson distribution to the data are shown figure B.4 over the whole distribution (i.e. x_{min} is calculated for each distribution as being the value for the fit that minimises the KS distance rather than x_{min} equal to the power-law x_{min} when comparing the log-likelihood of the distribution). This shows that the log-normal is a good statistical model of the whole distribution.

An estimate of the plausibility of a power law distribution was made by calculating the bootstrap p value (100 sims) for the power law distribution is 0.583 with goodness of fit KS 0.19 [5]. A bootstrap p value greater than 0.1 indicates that a power law distribution cannot be ruled out (see [145] and [77]).

To determine the best fitting distribution, I calculated the log-likelihood of the power law, lognormal, exponential and Poisson distribution fitted to the data. The log-likelihoods were calculated for each distribution from x_{min} determined for the power-law distribution $x_{min} = 24$ (see[77]) to allow direct comparison of the models. The values of the parameters are in table B.5 along with the bootstrap p value for each distribution. The plot of the distributions using $x_{min} = 24$ is shown in figure B.3.

B.1.1.0.1 Fit over xmin Over the portion for which a power law distribution holds ($x_{min} > 24$) the log-likelihood of the Poisson distribution is substantially lower than the exponential, log-normal[6] or power law.

The exponential distribution is clearly inferior to log-normal and power law using goodness of fit bootstrap p or the two sided Vuong test (p=3.455×10^{-7}). The one sided Vuong test p = 1.72×10^{-7} shows that the power law distribution is a better fit than the exponential as is clear in figure B.3.

There is no significant difference between the power law and log-normal using Vuong's two sided test (p=0.187). The one sided test favours the log-normal over power law (p=0.093) but was not significantly different[7].

[4] the gradient of the CDF is γ-1

[5] `RProjects/group_size_distribution/R/calulate_cdf.R`

[6] `https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/lognormal-distribution`

[7] power law v log-normal One sided p=0.906

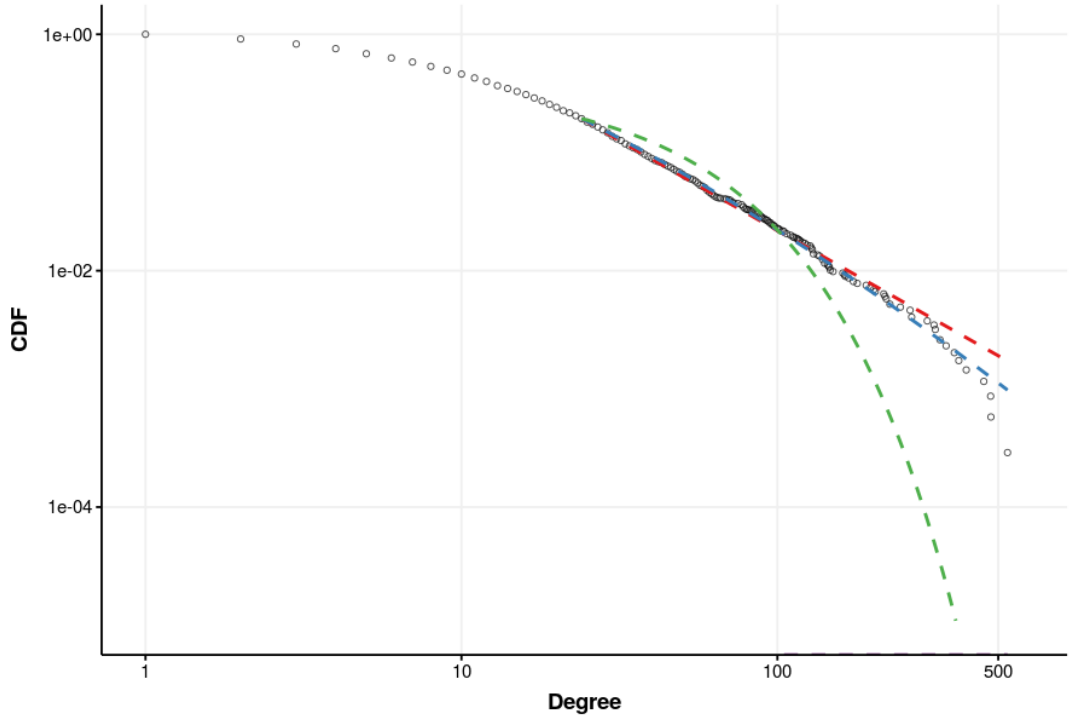


Figure B.3: CDF of degree distribution and lines of best fit for power law (red), log-normal(blue) exponential(green) and poisson (purple). Poisson distribution is not visible as the probability is too low, the exponential is a poorer fit in the tail of the distribution than either the powerlaw or log-normal . x axis degree log 10 scaled, y axis CDF log 10 scaled. Exponential appears a good fit but is fitted to a much smaller part of the distribution than the power law and log-normal. Poisson distribution is clearly not a good fit.

`RProjects/chapter3/R/fit_power_law/compare_distributions.R`

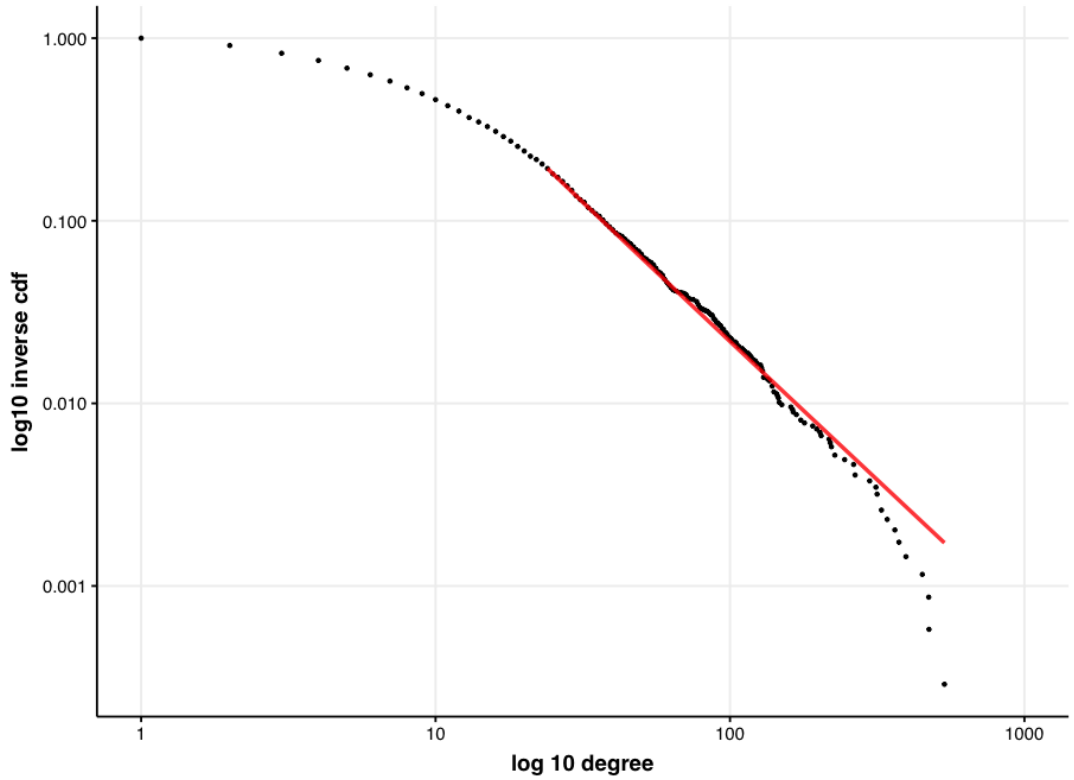
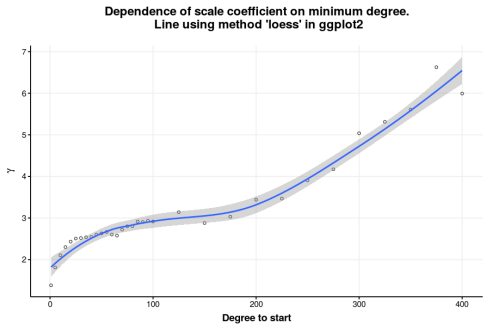


Figure B.4: Plot of fit of power law to degree distribution for the PSP. X axis is log 10 degree, Y axis is log 10 inverse cdf. Red line fit of power law distribution $\gamma = 2.51$ and $x_{min} = 24$ `RProjects/group_size_distribution/R/calulate_cdf.R`

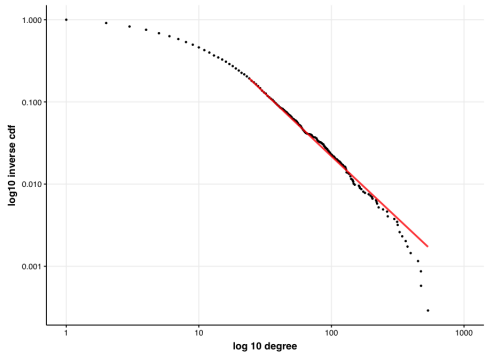
The level of confidence in the power law distribution according to the criteria of Broido and Clauset[58] for the PSP network is strong. The network is plausible by bootstrap p (therefore stronger than weakest), n_{tail} is greater than 50 (668 - at least weak) and meets the strong criteria by satisfying the previous criteria and having an exponent between two and three (2.51). 10% of the networks examined by Broido and Clauset had this characteristic. It does not meet strongest as the likelihood ratio test is inconclusive and does not favour either the log-normal or power law ($< 4\%$ of studied networks)

The distribution could be characterised either as a power law distribution or as a log-normal distribution, given recent results it could be argued it is likely to possess the features of robustness of a power-law distribution[376].



(a) γ as a function of x_{min}.

RProjects/PhD_graphs/calculate_gamma_theme.R



(b) The plot of log cumulative cdf of degree distribution log 10 scale x and y.

Figure B.5: Plots of degree distribution

model	LL	p_{bs}	g.o.f	pars$_1$	pars$_2$	Vuong$_{\text{statistic}}$	Vuong p_1	Vuong p_2
Power-law	-2943.61	0.57	0.02	2.511		-1.32	0.906	0.0187
log-normal	-2941.56	0.38	0.01	-2.345	2.11	1.32	0.094*	0.0187
exponential	-3044.66	< 0.001	0.09	0.029		5.10	1.73×10^{-7}	3.46×10^{-7}
poisson	-14376.28	< 0.001	0.09	58.590		7.31	1.29×10^{-13}	2.58×10^{-13}

Table B.5: Fit of distributions for PSP. Vuong test power law versus other test other than log-normal where one sided is log-normal versus power law marked with*. One sided test favours log-normal but not significantly so.

RProjects/chapter3/R/fit_power_law/fit_pl_setxmin.R

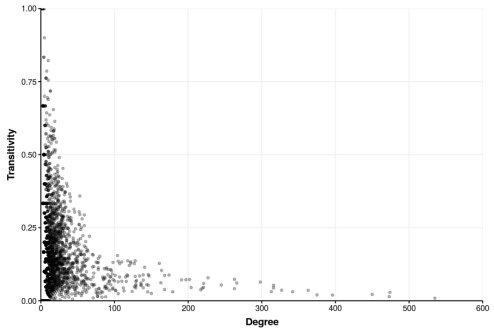
B.1.2 Global measures: Relationship of $C(k)$ to k

In section 3.8 all centrality measures are positively correlated yet mean local transitivity for nodes of degree k: $C(k)$, constant in random networks, is reported to decrease with increasing degree[447],[4] as,

$$C(k) = \frac{B}{k^{\beta}}, \tag{B.1}$$

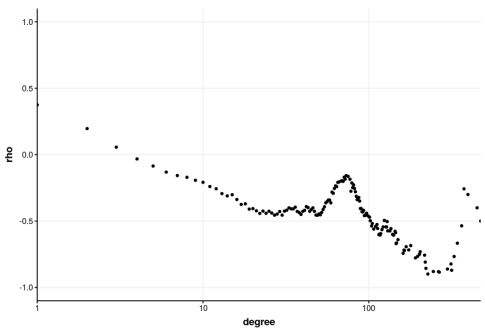
where B is a constant and β is typically between one and two[4],[447] in networks with hierarchical organisation. High degree nodes should have a lower mean transitivity and the relationship be roughly linear relationship when both axes are logarithmic. The relationship between $C(k)$ and degree k in the PSP network is shown in figure B.7b[8]. Initially $C(k)$ does not change with increasing degree but eventually $C(k)$ clearly decreases as k increases. An almost exactly identical pattern is seen in[447]

[8] RProjects/graph2community/R/transitivity_degree/Cluster_nodes.R



(a) Scatter plot of the relationship between degree and local transitivity for the PSP nodes. Nodes with degree $k = 1$ have transitivity zero. Plot appears to show a negative correlation between degree and transitivity. The opacity of points has been decreased to reduce overplotting and show the number of points of low degree.

RProjects/chapter3/R/transitivity/transitivity_and_degree/mean_deg_seq.R plot p



(b) Plot of minumum degree included in the calculation of Spearmans rank correlation between degree and ρ. Although the overall coefficient is positive the correlation for nodes above degree 5 are clearly negative and this becomes more apparent at high degree.

- RProjects/chapter3/R/transitivity/transitivity_and_degree/plot/plot_start_degree.R

Figure B.6: $C(k)$ and k_{min}

(figure 2b) reporting the relationship in Eq(B.1).

The overall correlation coefficient between degree and transitivity$_{\mathrm{NaN}}$ is positive ($\rho = 0.196$; $p < 2.2 \times 10^{-16}$) and is higher when isolates are zero (ρ=0.375 for transitivity$_0$). The reason for the difference between the positive correlation across the entire distribution and the relationship in figure B.7b) is the effect of nodes of degree one ($k = 1$) and hence transitivity 0 - having very low degree and minimal transitivity and so being highly correlated(figure B.6a). The transitivity is also low for nodes of degree two.

For nodes with degree $k >= 2$, ρ=0.196; $k >= 3$, ρ=0.056 and by $k >= 5$ the correlation is negative (ρ=-0.085). The median degree is eight and five is the 0.369 quantile. The positive correlation between degree and transitivity found across all nodes is the result of the effect of the third of nodes with smallest degree (histogram of transitivity see fig B.7a). Plotting ρ against the minimum degree included in calculating the correlation coefficient(k_{min}) clearly shows the switch from positive to negative correlation between low degree and high degree regions (figure B.6b).

The scaling described as being typical of hierarchical organisation is therefore seen clearly only when degree is above 5 and the relationship changes from a positive correlation in the low degree regime to the previously described scaling. In networks

with large numbers of low degree node the correlation coefficient will not show this relationship and setting isolates to zero makes this positive correlation stronger[9].

B.1.3 Results Small world

The small world effect is a combination of a high clustering coefficient and a short average path length (section 3.6.5)

B.1.3.1 Global clustering coefficient

The global clustering coefficient defined by Watts and Strogatz (C^0_{WS}), is 0.156 when isolates are set to zero[429] and 0.171 if the isolates are undefined and excluded (C^{NaN}_{WS} - the default in igraph) is 0.171 (see above section 3.2.8.2). A histogram of local transitivity values is seen in figure B.7a, although bounded in the region 0-1 the probability mass is concentrated on the left of the graph due to the number of nodes with low transitivity.

The value of the global transitivity described by Newman (C section 3.2.8.1) is 0.070. The different global transitivity measures are reviewed in table B.6.

Name	Symbol	type	isolates	Global Transitivity
Watts-Strogatz	C^0_{WS}	'localaverageundirected'	isolates='zero'	0.156
Watts-Strogatz	C^{NaN}_{WS}	'localaverageundirected'*	isolates='NaN'	0.171
Newman	C	'globalundirected'	isolates='zero'	0.070
Newman	C	'globalundirected'	isolates='NaN'	0.070

Table B.6: Different global transitivities in igraph. * or mean of 'localundirected' with isolates 'NaN' RProjects/chapter3/R/transitivity/global_transitivity/global_transitivity.R same project R/transitivity/global_transitivity/mkdown/global_transitivity_markdown.html

The global transitivity values compared to Monte Carlo estimates using Erdos Renyi random graphs with the same number of edges and nodes as the PSP (table B.7). The Watts-Strogatz transitivity is larger than the Newman transitivity (0.156 vs 0.70). It is not clear what the significance of the divergence in values is [289] but it is absent in the random graphs and appears to be a feature of the network. The absolute size of the difference between the PSP and ER transitivity is small but is of the order of

[9] RProjects/centrality/R/transitivity/calculate_correlation_min_transitivity_and_plot.R',

several standard deviations (377.5 for Newman Transitivity) and the samples of the ER transitivity appear normally distributed (see KS in table B.7).

	Median ER	Mean ER	PSP	difference	sd	n sd	ks d	ks p	p
C	5.096×10^{-3}	5.106×10^{-3}	0.070	0.065	1.712×10^{-4}	377.5	0.026	0.51	0
C_{WS}^{NaN}	5.095×10^{-3}	5.104×10^{-3}	0.171	0.166	1.816×10^{-4}	915.4	0.024	0.63	0
C_{WS}^{0}	5.095×10^{-3}	5.104×10^{-3}	0.156	0.151	1.816×10^{-4}	832.4	0.024	0.63	0

Table B.7: The global transitivity measures of the PSP are compared with Monte Carlo estimates of the transitivity measures in Erdos Renyi random graphs with matching numbers of vertices and edges. 1000 iterations. difference = PSP - mean Monte Carlo Estimate n sd number of times difference greater than standard deviation of MC estimate mean. ks d = Kolmogorov-Smirnov statistic distance,ks p Kolmogorov-Smirnov test statistic comparing Monte Carlo estimate to normal distribution with identical mean and standard deviation, p probability of result as extreme as PSP using mean and standard deviation of MC estimates.

RProjects/chapter3/R/transitivity/global_transitivity/monte_carlo_er_global_transitivity.R

Monte Carlo estimates of the transitivity in the degree constant configuration model are shown in table B.8. There is less difference between the PSP transitivity and the Configuration model than between the PSP and the ER graphs in both absolute terms or in the number of standard distributions it represents. The Watts-Strogatz global transitivity is again differs most from the random model. There is a difference in the value of the mean transitivity between the different types of transitivity with their rank order being the same as the PSP (Watts-Strogatz is greatest). I would hypothesise some but not all of the divergence in these values (WS and Newman) are due to degree structure but the divergence appears to be a feature of real world graphs.

The global undirected transitivity of the PSP network is 0.0697. This is similar to the value reported by Newman[10] [289] in metabolic networks (0.090) and PPI (0.072).

B.1.3.2 Results: Average path length and transitivity

Having dealt with the results for the global clustering coefficient we also need to determine the average path length in order to investigate the small world phenomenon.

[10] p305

	Median Config	Mean	PSP	difference	sd	n sd	ks d	ks p	p
C	5.835×10^{-2}	5.836×10^{-2}	0.070	0.011	8.032×10^{-4}	14.2	0.016	0.95	0
C_{WS}^{NaN}	7.475×10^{-2}	7.480×10^{-2}	0.171	0.097	1.981×10^{-3}	48.8	0.024	0.63	0
C_{WS}^{0}	6.818×10^{-2}	6.822×10^{-2}	0.156	0.088	1.807×10^{-3}	48.8	0.024	0.63	0

Table B.8: PSP vs 1000 Configuration model Monte Carlo. 1000 iterations. difference = PSP - mean Monte Carlo Estimate n sd number of times difference greater than standard deviation of MC estimate mean. ks d = Kolmogorov-Smirnov statistic distance,ks p Kolmogorov-Smirnov test statistic comparing Monte Carlo estimate to normal distribution with identical mean and standard deviation, p probability of result as extreme as PSP using mean and standard deviation of MC estimates.
RProjects/chapter3/R/transitivity/global_transitivity/monte_carlo_config2_global_transitivity.R

Graph	APL	sd	C_{WS}	$C_{WS}0$	C
PSP	2.98	NA	0.171	0.156	0.0697
Config	3.0069	0.0053	0.0683	0.0748	0.0584
ER	3.118	0.00047	0.0051	0.0051	0.0051

Table B.9: Alternative way of presenting the data take transitivity from PSP vs 1000 Configuration model Monte Carlo RProjects/chapter3/R/transitivity/global_transitivity/monte_carlo_config2_global_transitivity.R

The average path length of the PSP was again compared Monte Carlo estimates of the average path length for the Erdos Renyi models and configuration models.

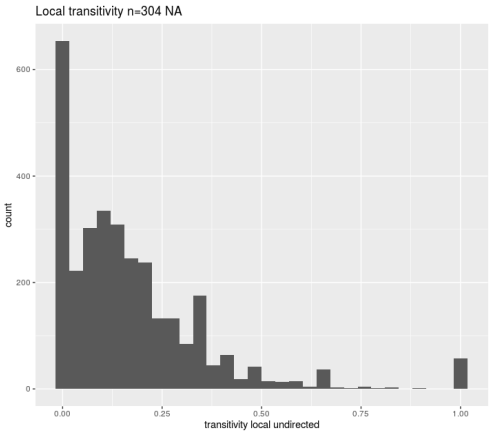
The average path length, calculated using the `mean_distance` function in igraph, was shortest for the PSP (2.98) and longest for the Erdos Renyi graph with the Configuration Model again in the middle. The value for one sample t tests of the difference between the Monte Carlo estimates and the PSP[11] indicate that the PSP has a significantly shorter APL B.13.

For the degree corrected configuration model see table B.9. The average path length is shorter than in the Erdos Renyi model likely due to the presence of hub nodes[4][12].

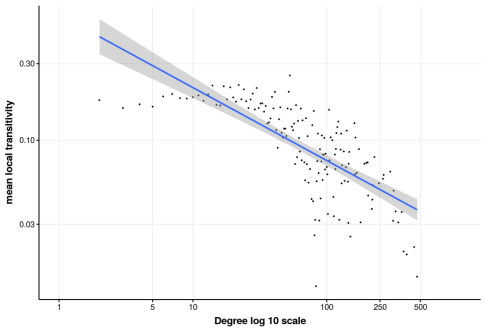
This is similar to the L of 2.65 and C of 0.28 of C elegans, where L is the "shortest path between two vertices, averaged over all pairs of vertices and C is the clustering coefficient" as defined by Watts and Strogatz [429]. The random rewiring with a uniform

[11] `RProjects/chapter3/R/transitivity/global_transitivity/t_test/compare_er_config.R`

[12] `RProjects/graph_sw/R/small_world_random/degree_corrected_config_random_ws.R`



(a) Histogram of transitivity



(b) Plot of degree with mean local transitivity. Log10-log10 scale. Omitting degree $k = 1$ as transitivity C will be undefined.

Figure B.7: Plots of transitivity.

probability leads to a L of 3.11. For comparison the path length and global clustering from [429] is shown in table B.10

Network	L_{actual}	L_{random}	C_{actual}	C_{random}
Film actors	3.65	2.99	0.79	0.00027
Power grid	18.7	12.4	0.080	0.005
C elegans	2.65	2.25	0.28	0.05

Table B.10: Clustering coefficient and average path length of three networks compared to random Erdos Renyi graph with same number of nodes and edges - from Watts and Strogatz[429]. L is average path length, C is global transitivity of type Watts-Strogatz

B.1.3.3 Distance distribution

An alternative test presence of the small world effect ([23][13]) is to compare the distance distribution (pairwise distances between all nodes) with the Erdos-Renyi graph and the configuration model. This is essentially using all of the path lengths that make up the average path length to have a distribution of values for the PSP to compare to a random model.

[13] p143

	Median	Mean	PSP	difference	sd	n_sd	ks_d	ks_p	p
APL	3.01	3.01	2.980	-0.027	5.304×10^{-3}	-5.2	0.017	0.93	1

Table B.11: PSP vs 1000 Configuration model Monte Carlo

	Median	Mean	PSP	CI	difference	sd	n_sd	ks_d	ks_p	p
APL	3.12	3.12	2.980	3.1179-3.1180	-0.138	4.875×10^{-4}	-283.5	0.028	0.43	$< 2.2 \times 10^{-16}$

Table B.12: PSP vs 1000 ER model Monte Carlo

The results show that the PSP has a significantly shorter mean path length compared to the other two. The configuration model has a shorter path length than the ER model but longer than the PSP. This is in keeping with our other results - tables B.14 and table B.15. Infinite values have been removed from the random graphs (distance values are infinite when the graph is disconnected).

The results support the PSP having a small world configuration.

B.1.3.3.1 Small worldness measure The small world measure S where a graph is small world if $S > 1$ [190] is 14.3[14]. Using the definition of small worldness from Humphries and Gurney.

S is $S > 1$ for small world networks:

$$\gamma_g = \frac{C_g}{C_{random}} \tag{B.2}$$

and

$$\lambda_g = \frac{C_g}{C_{random}} \tag{B.3}$$

then

$$S = \frac{\gamma}{\lambda} \tag{B.4}$$

S using the global measure of transitivity (C) was calculated to be 14.3[15] and S using the Watts-Strogatz global clustering coefficient C_{ws}= 35.08[16]

[14] `RProjects/chapter3/R/small_world/calculate_humphries_small_world_metric.R`
[15] `RProjects/chapter3/R/small_world/calculate_humphries_small_world_metric.R`
[16] `RProjects/chapter3/R/small_world/calculate_humphries_small_world_metricWS.R`

	estimate	t-statistic	p.value	d.f	conf.low	conf.high	method	alternative
t-test APL ER	3.118	8964.21	0	999	3.1180	3.1180	One Sample t-test	two.sided

Table B.13: T test Average path length difference with ER random graph

RProjects/chapter3/R/small_world/average_path_length/monte_carlo_er_apl.R

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Graph types	2	64211.02	32105.51	81460.66	0.0000
Residuals	17921085	7063109.61	0.39		

Table B.14: ANOVA (aov in R) of distance distribution of PSP, Erdos Renyi Graph and Configuration model. Erdos-Renyi with n,m matching PSP. Configuration uses degree distribution of PSP

RProjects/chapter3/R/small_world/distance_distribution/barabasi_distance_distribution.R

A value greater than one is defined as denoting the small world property[190]. Thirty three real world networks showed a range of values from 1.34 (student relationships) to approximately 11,660 for math coauthorship ($S\ C$) and for WS from S=0.27 student relationship (not small world) to 61,967 for biology co-authorship n=11,803,064. They found that S scaled near linearly with n the number of vertices in the network. .

The predicted value for the PSP is is 57.39 ($0.023n^{0.96}$=57.39), Δ ie not WS) and for S using WS S_{ws}=$0.012n^{1.11}$ = 101.65

B.1.3.3.2 Conclusion The network shows the small world property. It is less than theoretical maximum small world and behaves similarly to other networks in its scaling with n. There is a wide variation to the small worldness

	diff	lwr	upr	p adj
ER-PSP	0.139	0.138	0.140	< 0.001
Config-PSP	0.024	0.023	0.025	< 0.001
Config-ER	-0.115	-0.116	-0.114	< 0.001

Table B.15: Post hoc Tukey multiple comparisons of means 95% family-wise confidence level Distance Distribution. p is 0. Diff= difference in means. Mean pairwise distance n=5,973,696 for PSP, Erdos Renyi gnm graph with identical n vertices and edges to PSP and configuration random graph using PSP degree distribution.

RProjects/chapter3/R/small_world/distance_distribution/barabasi_distance_distribution.R

B.2 Chapter 2

Gene set name	n genes	β	p
Exocytosis	83	0.19	0.039
Neurotransmitter metabolism	27	-0.10	0.686
Excitability	57	0.38	0.005
Protein cluster	43	0.20	0.091
Ligand-gated ion channel signaling	33	0.03	0.438
Tyrosine kinase signaling	7	0.22	0.308
Unassigned	59	0.07	0.301
Peptide/neurotrophin signals	28	0.04	0.430
Intracellular signal transduction	148	-0.04	0.697
Ion balance/transport	41	-0.01	0.522
G-protein relay	27	-0.10	0.688
RNA and protein synthesis, folding and breakdown	67	0.01	0.448
Endocytosis	25	-0.17	0.804
Structural plasticity	95	-0.06	0.731
Cell metabolism	51	-0.13	0.827
Cell adhesion and trans-synaptic signaling	76	0.29	0.009
G-protein-coupled receptor signaling	41	0.09	0.300
Intracellular trafficking	75	0.04	0.344

Table B.16: Synaptic enrichment using components. Intelligence Discovery

RProjects/redo_chapter2/R/redo_synapse_set_int_curated.R

Gene set name	n genes	β	p
Exocytosis	83	0.04	0.348
Neurotransmitter metabolism	27	-0.08	0.648
Excitability	57	0.43	0.003
Protein cluster	43	-0.14	0.813
Ligand-gated ion channel signaling	33	0.18	0.178
Tyrosine kinase signaling	7	-0.18	0.658
Unassigned	59	0.08	0.276
Peptide/neurotrophin signals	28	-0.20	0.828
Intracellular signal transduction	148	0.07	0.226
Ion balance/transport	41	0.11	0.242
G-protein relay	27	-0.09	0.671
RNA and protein synthesis, folding and breakdown	67	0.10	0.198
Endocytosis	25	0.12	0.281
Structural plasticity	95	0.04	0.355
Cell metabolism	51	0.05	0.374
Cell adhesion and trans-synaptic signaling	76	0.39	0.001
G-protein-coupled receptor signaling	41	-0.01	0.517
Intracellular trafficking	75	-0.06	0.713

Table B.17: Synaptic enrichment using components. Education Discovery

Gene set name	n genes	β	p
Cell adhesion and trans-synaptic signaling	76	0.11	0.155
Cell metabolism	51	-0.07	0.726
Endocytosis	25	-0.03	0.572
Excitability	57	0.13	0.159
Exocytosis	83	0.18	0.016
G-protein-coupled receptor signaling	41	0.15	0.162
G-protein relay	27	0.11	0.251
Intracellular signal transduction	148	0.07	0.150
Intracellular trafficking	75	-0.01	0.522
Ion balance/transport	41	-0.17	0.898
Ligand-gated ion channel signaling	33	-0.27	0.944
Neurotransmitter metabolism	27	0.25	0.099
Peptide/neurotrophin signals	28	-0.10	0.720
Protein cluster	43	0.11	0.207
RNA and protein synthesis, folding and breakdown	67	0.11	0.120
Structural plasticity	95	-0.10	0.874
Tyrosine kinase signaling	7	-0.45	0.880
Unassigned	58	-0.03	0.620

Table B.18: Synaptic enrichment using component Intelligence Discovery sample

Gene set name	n genes	β	p
Cell adhesion and trans-synaptic signaling	76	0.17	0.067
Cell metabolism	53	-0.02	0.549
Endocytosis	26	0.11	0.251
Excitability	57	0.16	0.122
Exocytosis	83	-0.11	0.877
G-protein-coupled receptor signaling	41	0.08	0.315
G-protein relay	27	-0.01	0.521
Intracellular signal transduction	147	0.02	0.418
Intracellular trafficking	75	-0.02	0.590
Ion balance/transport	40	0.05	0.370
Ligand-gated ion channel signaling	33	0.18	0.158
Neurotransmitter metabolism	27	-0.15	0.774
Peptide/neurotrophin signals	28	0.11	0.267
Protein cluster	42	-0.20	0.910
RNA and protein synthesis, folding and breakdown	66	0.22	0.016
Structural plasticity	95	0.10	0.140
Tyrosine kinase signaling	7	-0.27	0.742
Unassigned	58	0.07	0.290

Table B.19: Synaptic enrichment using components. Educational attainment replication

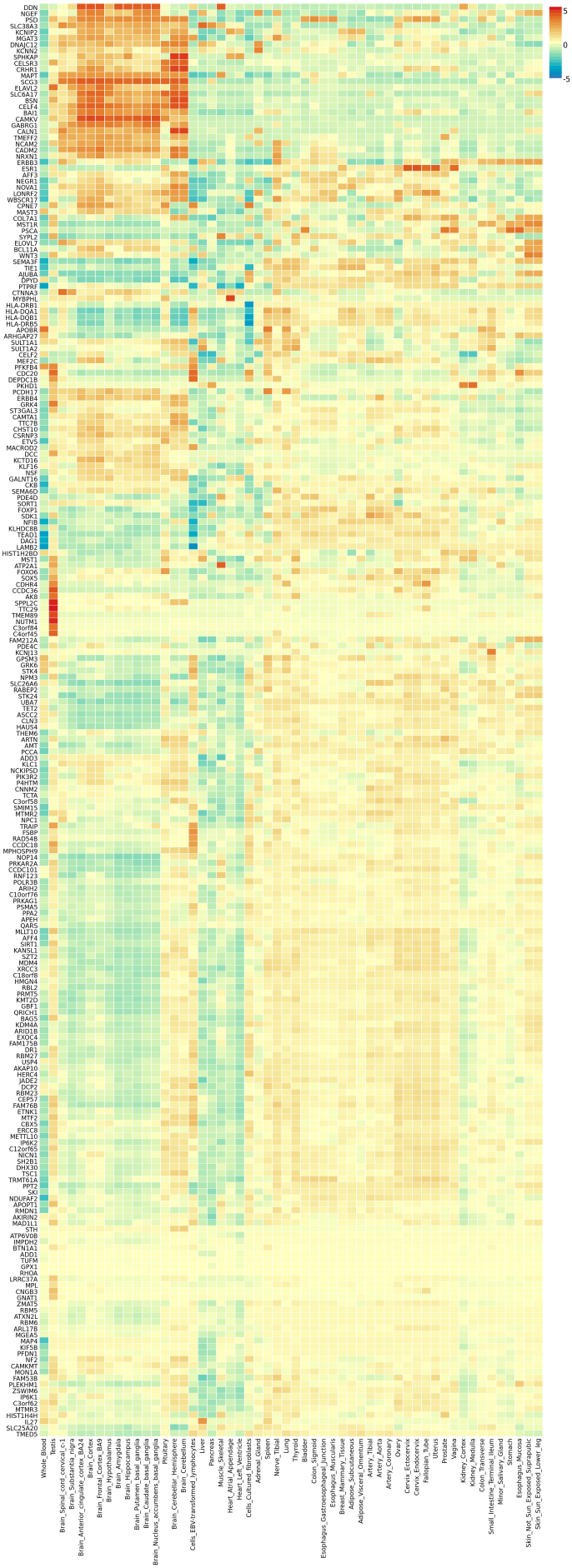


Figure B.8: Heatmap gene expression FUMA UKBBEd Average value of the relavitve expression level zero mean centred and log2 transformed. Shows that a small number of genes amongst those enriched are very enriched

B.3 Chapter Community Detection

B.3.1 Community statistics following Fortunato[134]

g	n_c	k_{int}	μk_{int}	δ_{kint}	k_{ext}	μ_{kext}	δ_{kext}	$\sum_k$	μ_k	conductance
1	1228	24454	19.914	0.016	8388	6.831	0.003	32842	26.744	0.255
2	130	2082	16.015	0.124	1849	14.223	0.004	3931	30.238	0.470
3	118	894	7.576	0.065	1407	11.924	0.004	2301	19.500	0.611
4	83	594	7.157	0.087	995	11.988	0.004	1589	19.145	0.626
5	52	508	9.769	0.192	771	14.827	0.004	1279	24.596	0.603
6	41	548	13.366	0.334	608	14.829	0.004	1156	28.195	0.526
7	64	326	5.094	0.081	583	9.109	0.003	909	14.203	0.641
8	66	132	2.000	0.031	550	8.333	0.002	682	10.333	0.806
9	50	286	5.720	0.117	463	9.260	0.003	749	14.980	0.618
10	50	248	4.960	0.101	463	9.260	0.003	711	14.220	0.651
11	51	272	5.333	0.107	373	7.314	0.002	645	12.647	0.578
12	43	424	9.860	0.235	246	5.721	0.002	670	15.581	0.367
13	41	80	1.951	0.049	478	11.659	0.003	558	13.610	0.857
14	40	310	7.750	0.199	256	6.400	0.002	566	14.150	0.452
15	34	206	6.059	0.184	389	11.441	0.003	595	17.500	0.654
16	34	172	5.059	0.153	237	6.971	0.002	409	12.029	0.579
17	23	122	5.304	0.241	313	13.609	0.004	435	18.913	0.720
18	32	86	2.688	0.087	266	8.312	0.002	352	11.000	0.756
19	27	80	2.963	0.114	265	9.815	0.003	345	12.778	0.768
20	28	172	6.143	0.228	182	6.500	0.002	354	12.643	0.514
21	23	84	3.652	0.166	205	8.913	0.003	289	12.565	0.709
22	26	86	3.308	0.132	168	6.462	0.002	254	9.769	0.661
23	21	86	4.095	0.205	207	9.857	0.003	293	13.952	0.706
24	22	92	4.182	0.199	126	5.727	0.002	218	9.909	0.578
25	17	68	4.000	0.250	148	8.706	0.003	216	12.706	0.685
27	18	52	2.889	0.170	135	7.500	0.002	187	10.389	0.722
28	18	48	2.667	0.157	137	7.611	0.002	185	10.278	0.741
29	18	64	3.556	0.209	114	6.333	0.002	178	9.889	0.640
30	17	58	3.412	0.213	115	6.765	0.002	173	10.176	0.665
31	15	34	2.267	0.162	161	10.733	0.003	195	13.000	0.826
32	15	104	6.933	0.495	82	5.467	0.002	186	12.400	0.441
34	18	52	2.889	0.170	94	5.222	0.002	146	8.111	0.644
35	15	32	2.133	0.152	130	8.667	0.003	162	10.800	0.802
36	18	38	2.111	0.124	100	5.556	0.002	138	7.667	0.725
40	15	42	2.800	0.200	78	5.200	0.002	120	8.000	0.650
46	15	30	2.000	0.143	47	3.133	0.001	77	5.133	0.610
52	15	30	2.000	0.143	35	2.333	0.001	65	4.333	0.538

Table B.20: Community statistics after fortunato infomap clustering size > 15

g	n_c	k_{int}	μk_{int}	δ_{kint}	k_{ext}	μ_{kext}	δ_{kext}	$\sum_k$	μ_k	conductance
1	171	916	5.36	3.15×10^{-2}	1536	8.98	1.63×10^{-3}	2452	14.34	0.626
2	165	1250	7.58	4.62×10^{-2}	2248	13.62	2.30×10^{-3}	3498	21.20	0.643
3	41	110	2.68	6.71×10^{-2}	253	6.17	7.85×10^{-4}	363	8.85	0.697
4	31	94	3.03	1.01×10^{-1}	198	6.39	8.85×10^{-4}	292	9.42	0.678
5	112	460	4.11	3.70×10^{-2}	737	6.58	1.23×10^{-3}	1197	10.69	0.616
6	79	172	2.18	2.79×10^{-2}	593	7.51	6.45×10^{-4}	765	9.68	0.775
7	129	364	2.82	2.20×10^{-2}	863	6.69	8.48×10^{-4}	1227	9.51	0.703
9	65	166	2.55	3.99×10^{-2}	451	6.94	7.53×10^{-4}	617	9.49	0.731
10	154	1042	6.77	4.42×10^{-2}	1069	6.94	2.05×10^{-3}	2111	13.71	0.506
11	56	154	2.75	5.00×10^{-2}	687	12.27	8.09×10^{-4}	841	15.02	0.817
12	34	124	3.65	1.11×10^{-1}	317	9.32	1.07×10^{-3}	441	12.97	0.719
16	72	294	4.08	5.75×10^{-2}	589	8.18	1.21×10^{-3}	883	12.26	0.667
17	148	852	5.76	3.92×10^{-2}	1136	7.68	1.74×10^{-3}	1988	13.43	0.571
19	20	46	2.30	1.21×10^{-1}	191	9.55	6.69×10^{-4}	237	11.85	0.806
20	151	800	5.30	3.53×10^{-2}	2012	13.32	1.60×10^{-3}	2812	18.62	0.716
22	42	94	2.24	5.46×10^{-2}	409	9.74	6.55×10^{-4}	503	11.98	0.813
23	72	162	2.25	3.17×10^{-2}	671	9.32	6.65×10^{-4}	833	11.57	0.806
24	76	348	4.58	6.11×10^{-2}	686	9.03	1.35×10^{-3}	1034	13.61	0.663
25	27	54	2.00	7.69×10^{-2}	234	8.67	5.83×10^{-4}	288	10.67	0.812
26	120	480	4.00	3.36×10^{-2}	1238	10.32	1.20×10^{-3}	1718	14.32	0.721
28	38	132	3.47	9.39×10^{-2}	272	7.16	1.02×10^{-3}	404	10.63	0.673
32	52	152	2.92	5.73×10^{-2}	514	9.88	8.58×10^{-4}	666	12.81	0.772
33	191	1104	5.78	3.04×10^{-2}	1787	9.36	1.77×10^{-3}	2891	15.14	0.618
34	215	2680	12.47	5.82×10^{-2}	2520	11.72	3.84×10^{-3}	5200	24.19	0.485
43	150	972	6.48	4.35×10^{-2}	2317	15.45	1.96×10^{-3}	3289	21.93	0.704
44	103	616	5.98	5.86×10^{-2}	1807	17.54	1.78×10^{-3}	2423	23.52	0.746
45	124	624	5.03	4.09×10^{-2}	1998	16.11	1.51×10^{-3}	2622	21.15	0.762
46	49	302	6.16	1.28×10^{-1}	810	16.53	1.81×10^{-3}	1112	22.69	0.728
47	27	58	2.15	8.26×10^{-2}	290	10.74	6.26×10^{-4}	348	12.89	0.833
48	23	74	3.22	1.46×10^{-1}	244	10.61	9.37×10^{-4}	318	13.83	0.767
51	112	494	4.41	3.97×10^{-2}	1423	12.71	1.32×10^{-3}	1917	17.12	0.742
53	405	7072	17.46	4.32×10^{-2}	5479	13.53	5.72×10^{-3}	12551	30.99	0.437
55	44	164	3.73	8.67×10^{-2}	678	15.41	1.09×10^{-3}	842	19.14	0.805
58	37	128	3.46	9.61×10^{-2}	950	25.68	1.01×10^{-3}	1078	29.14	0.881
62	34	172	5.06	1.53×10^{-1}	305	8.97	1.48×10^{-3}	477	4.03	0.639

Table B.21: Community statistics for Spectral Clustering of the PSP following Fortunato. Com = community number, n = number of genes in community, k int = internal degree, sum of the internal degree of the vertices in the community, av. int. deg = average internal degree, delta int = internal degree density:ratio number of internal edges and all possible edges, k ext = external degree, av. ext. degree = average external degree, delta ext: external degree density: ration of external edges and all possible external edges (a.k.a cut ratio), average external degree (expansion) conductance, k c = total degree or volume. av. com. degree = average combined degree, Showing only communities greater than or equal to 15 in size. Total communities = 65. Code see section B.4.2 for details on statistics

`RProjects/fortunato/R/calculate_fortunato_measures_name_input.R`

	Discovery					*Replication*				
Community	*n*	β	β STD	SE	*p*	*n*	β	β STD	SE	*p*
1	161	0.10	0.01	0.08	0.104	160	0.07	0.01	0.07	0.169
2	153	0.09	0.01	0.08	0.120	153	0.17	0.02	0.07	0.012
3	39	-0.21	-0.01	0.16	0.905	39	-0.02	-0.00	0.14	0.556
4	29	-0.30	-0.01	0.19	0.944	29	-0.12	-0.00	0.15	0.789
5	106	0.29	0.02	0.10	**0.002**	106	0.20	0.02	0.09	**0.015**
6	76	0.01	0.00	0.12	0.454	76	0.10	0.01	0.10	0.162
7	125	0.01	0.00	0.09	0.470	123	-0.09	-0.01	0.08	0.867
9	61	0.10	0.01	0.12	0.191	61	0.03	0.00	0.11	0.402
10	146	0.09	0.01	0.08	0.133	146	0.06	0.01	0.07	0.206
11	55	-0.02	-0.00	0.14	0.570	55	-0.05	-0.00	0.11	0.665
12	32	0.11	0.00	0.16	0.244	32	0.19	0.01	0.16	0.112
16	68	-0.14	-0.01	0.12	0.868	68	0.06	0.00	0.11	0.282
17	144	-0.02	-0.00	0.08	0.605	144	-0.05	-0.00	0.07	0.754
19	20	0.01	0.00	0.19	0.475	20	0.02	0.00	0.19	0.464
20	144	0.00	0.00	0.09	0.489	142	0.22	0.02	0.08	0.002
22	42	0.27	0.01	0.15	**0.035**	41	0.09	0.00	0.13	0.250
23	67	-0.00	-0.00	0.12	0.509	67	0.12	0.01	0.11	0.125
24	74	0.10	0.01	0.12	0.188	74	0.12	0.01	0.11	0.122
25	26	-0.51	-0.02	0.20	0.995	26	0.12	0.00	0.17	0.238
26	117	0.08	0.01	0.10	0.209	116	-0.09	-0.01	0.09	0.841
28	35	-0.13	-0.01	0.17	0.783	35	0.09	0.00	0.15	0.289
32	50	0.17	0.01	0.15	0.130	49	-0.24	-0.01	0.14	0.961
33	180	0.10	0.01	0.08	0.097	180	0.12	0.01	0.07	0.049
34	210	0.02	0.00	0.07	0.400	209	-0.02	-0.00	0.06	0.620
43	137	-0.04	-0.00	0.08	0.669	137	0.01	0.00	0.08	0.448
44	101	-0.02	-0.00	0.10	0.563	100	-0.01	-0.00	0.08	0.562
45	118	0.03	0.00	0.09	0.374	118	0.04	0.00	0.08	0.294
46	47	0.12	0.01	0.13	0.169	47	0.11	0.01	0.12	0.167
47	23	0.37	0.01	0.20	**0.035**	23	-0.05	-0.00	0.21	0.599
48	22	0.07	0.00	0.21	0.374	22	0.07	0.00	0.22	0.367
51	110	0.01	0.00	0.09	0.466	110	-0.04	-0.00	0.08	0.710
53	386	0.11	0.02	0.05	*0.015*	381	0.06	0.01	0.04	0.085
55	42	0.05	0.00	0.15	0.378	42	-0.16	-0.01	0.15	0.855
58	36	-0.12	-0.01	0.16	0.767	35	-0.05	-0.00	0.16	0.629
62	34	-0.14	-0.01	0.16	0.808	34	-0.14	-0.01	0.14	0.826

Table B.22: MAGMA GSA Education BETA is the regression coefficient of the gene set. SE is the standard error of the regression coefficient. BETA SD the the semi-standardized regression coefficient, corresponding to the predicted change in Z-value given a change of one standard deviation in the predictor gene set / gene covariate. The significant discovery genes that achieve nominal significance in GSA are in bold. Those nominally significant only in MAGMA are in italics. Four have nominally significant p values but only three are significant in both MAGMA and GSEA

RProjects/chapter_community_detection/R/magma_gsa/education_gsa_table.R

RProjects/chapter_community_detection/R/magma_gsa/rawread/education_gsa_new_raw_table.R

g	n_c	k_{int}	μk_{int}	δ_{kint}	k_{ext}	μ_{kext}	δ_{kext}	$\sum_k$	μ_k	conductance
1	39	154	3.95	1.04×10^{-1}	367	9.41	1.16×10^{-3}	521	13.36	0.704
2	60	618	10.30	1.75×10^{-1}	631	10.52	3.03×10^{-3}	1249	20.82	0.505
3	82	984	12.00	1.48×10^{-1}	930	11.34	3.56×10^{-3}	1914	23.34	0.486
4	172	2202	12.80	7.49×10^{-2}	2038	11.85	3.90×10^{-3}	4240	24.65	0.481
5	26	108	4.15	1.66×10^{-1}	231	8.88	1.21×10^{-3}	339	13.04	0.681
6	44	332	7.55	1.75×10^{-1}	500	11.36	2.21×10^{-3}	832	18.91	0.601
7	34	208	6.12	1.85×10^{-1}	331	9.74	1.79×10^{-3}	539	15.85	0.614
8	18	114	6.33	3.73×10^{-1}	164	9.11	1.84×10^{-3}	278	15.44	0.590
9	18	144	8.00	4.71×10^{-1}	202	11.22	2.33×10^{-3}	346	19.22	0.584
10	40	372	9.30	2.38×10^{-1}	445	11.12	2.72×10^{-3}	817	20.43	0.545
11	196	2328	11.88	6.09×10^{-2}	2286	11.66	3.64×10^{-3}	4614	23.54	0.495
12	40	292	7.30	1.87×10^{-1}	393	9.82	2.14×10^{-3}	685	17.12	0.574
13	16	84	5.25	3.50×10^{-1}	171	10.69	1.53×10^{-3}	255	15.94	0.671
14	26	150	5.77	2.31×10^{-1}	257	9.88	1.68×10^{-3}	407	15.65	0.631
15	125	1748	13.98	1.13×10^{-1}	1618	12.94	4.20×10^{-3}	3366	26.93	0.481
16	65	298	4.58	7.16×10^{-2}	617	9.49	1.35×10^{-3}	915	14.08	0.674
17	50	408	8.16	1.67×10^{-1}	548	10.96	2.40×10^{-3}	956	19.12	0.573
18	336	5370	15.98	4.77×10^{-2}	4254	12.66	5.12×10^{-3}	9624	28.64	0.442
19	89	736	8.27	9.40×10^{-2}	989	11.11	2.46×10^{-3}	1725	19.38	0.573
20	39	142	3.64	9.58×10^{-2}	410	10.51	1.07×10^{-3}	552	14.15	0.743
21	231	3302	14.29	6.21×10^{-2}	2919	12.64	4.43×10^{-3}	6221	26.93	0.469
22	31	234	7.55	2.52×10^{-1}	333	10.74	2.20×10^{-3}	567	18.29	0.587
23	48	368	7.67	1.63×10^{-1}	513	10.69	2.25×10^{-3}	881	18.35	0.582
24	58	520	8.97	1.57×10^{-1}	599	10.33	2.64×10^{-3}	1119	19.29	0.535
25	140	1656	11.83	8.51×10^{-2}	1605	11.46	3.57×10^{-3}	3261	23.29	0.492
26	558	8776	15.73	2.82×10^{-2}	6512	11.67	5.43×10^{-3}	15288	27.40	0.426
27	61	620	10.16	1.69×10^{-1}	683	11.20	2.99×10^{-3}	1303	21.36	0.524
28	61	290	4.75	7.92×10^{-2}	568	9.31	1.40×10^{-3}	858	14.07	0.662
29	222	2706	12.19	5.52×10^{-2}	2619	11.80	3.77×10^{-3}	5325	23.99	0.492
30	48	446	9.29	1.98×10^{-1}	477	9.94	2.73×10^{-3}	923	19.23	0.517
31	203	3002	14.79	7.32×10^{-2}	2507	12.35	4.54×10^{-3}	5509	27.14	0.455
32	69	694	10.06	1.48×10^{-1}	775	11.23	2.97×10^{-3}	1469	21.29	0.528
33	27	154	5.70	2.19×10^{-1}	309	11.44	1.66×10^{-3}	463	17.15	0.667
34	102	1120	10.98	1.09×10^{-1}	1194	11.71	3.27×10^{-3}	2314	22.69	0.516
35	56	546	9.75	1.77×10^{-1}	624	11.14	2.87×10^{-3}	1170	20.89	0.533
36	27	126	4.67	1.79×10^{-1}	263	9.74	1.36×10^{-3}	389	14.41	0.676

Table B.23: Community statistics for gtcommunity which is the ground truth allocation of nodes for N 3457 E 40617 LFR benchmark following Fortunato[134]. Com = community number, n = number of genes in community, k int = internal degree, sum of the internal degree of the vertices in the community, av. int. deg = average internal degree, delta int = internal degree density:ratio number of internal edges and all possible edges, k ext = external degree, av. ext. degree = average external degree, delta ext: external degree density: ration of external edges and all possible external edges (a.k.a cut ratio), average external degree (expansion) conductance, k c = total degree or volume. av. com. degree = average combined degree, Showing only communities greater than or equal to 1 in size. Total communities = 36

g	n_c	k_{int}	μk_{int}	δ_{kint}	k_{ext}	μ_{kext}	δ_{kext}	$\sum_k$	μ_k	conductance
1	209	658	3.148	0.015	1062	5.081	0.002	1720	8.230	0.617
2	350	1904	5.440	0.016	2117	6.049	0.002	4021	11.489	0.526
3	536	9554	17.825	0.033	6628	12.366	0.004	16182	30.190	0.410
4	265	1516	5.721	0.022	2848	10.747	0.003	4364	16.468	0.653
5	114	820	7.193	0.064	1349	11.833	0.004	2169	19.026	0.622
6	166	1198	7.217	0.044	1861	11.211	0.003	3059	18.428	0.608
7	154	564	3.662	0.024	646	4.195	0.001	1210	7.857	0.534
8	539	4738	8.790	0.016	5734	10.638	0.004	10472	19.429	0.548
9	448	4458	9.951	0.022	4036	9.009	0.003	8494	18.960	0.475
10	58	320	5.517	0.097	563	9.707	0.003	883	15.224	0.638
11	344	2288	6.651	0.019	2888	8.395	0.003	5176	15.047	0.558
12	21	58	2.762	0.138	174	8.286	0.002	232	11.048	0.750
13	81	424	5.235	0.065	828	10.222	0.003	1252	15.457	0.661
14	172	834	4.849	0.028	928	5.395	0.002	1762	10.244	0.527

Table B.24: Community statistics for discovered communities following Fortunato[134] in the PSP using Louvain community detection algorithm. Com = community number, n_c = number of genes in community, k_{int} = internal degree, sum of the internal degree of the vertices in the community, μk_{int} = average internal degree, delta int = internal degree density:ratio number of internal edges and all possible edges, k ext = external degree, av. ext. degree = average external degree, delta ext: external degree density: ration of external edges and all possible external edges (a.k.a cut ratio), average external degree (expansion) conductance, k c = total degree or volume. av. com. degree = average combined degree, Showing only communities greater than or equal to 1 in size.

B.3.2 Supplemental tables

B.3.3 Group 5 centrality measures

	group	n	mean	median	δ mean	p	p.adj
1	53	405	0.0001057	0.0001072	0.0000073	2.114×10^{-35}	1.400×10^{-33}
2	5	112	0.0000917	0.0000921	-0.0000068	2.009×10^{-9}	1.300×10^{-7}
3	9	65	0.0000914	0.0000917	-0.0000070	1.521×10^{-5}	9.600×10^{-4}
4	34	215	0.0001014	0.0001015	0.0000030	5.907×10^{-5}	3.700×10^{-3}
5	28	38	0.0000920	0.0000910	-0.0000065	1.847×10^{-4}	1.100×10^{-2}
6	6	79	0.0000945	0.0000947	-0.0000039	1.878×10^{-4}	1.100×10^{-2}
7	17	148	0.0000950	0.0000975	-0.0000034	6.080×10^{-4}	3.600×10^{-2}

Table B.25: Significantly close groups compared to rest of PSP Wilcoxon Group 5

RProjects/chapter_community_detection/R/group5/centrality_measures/plot_histogram_closenessbygroup.R

	group	n	mean	median	δ mean	p	p.adj
1	53	405	0.1213551	0.0940556	0.0734266	7.006×10^{-60}	4.600×10^{-58}
2	5	112	0.0167635	0.0089846	-0.0311651	8.827×10^{-12}	5.600×10^{-10}
3	7	129	0.0190148	0.0098960	-0.0289138	6.548×10^{-10}	4.100×10^{-8}
4	6	79	0.0193213	0.0092125	-0.0286073	4.070×10^{-9}	2.500×10^{-7}
5	34	215	0.0641540	0.0443283	0.0162254	3.944×10^{-8}	2.400×10^{-6}
6	9	65	0.0167916	0.0099634	-0.0311370	5.466×10^{-7}	3.300×10^{-5}
7	17	148	0.0204287	0.0161364	-0.0274999	2.869×10^{-6}	1.700×10^{-4}
8	43	150	0.0651438	0.0326865	0.0172152	8.470×10^{-6}	4.900×10^{-4}
9	28	38	0.0146490	0.0058165	-0.0332796	1.435×10^{-5}	8.200×10^{-4}
10	45	124	0.0625388	0.0399817	0.0146102	4.154×10^{-5}	2.300×10^{-3}
11	58	37	0.0954283	0.0466383	0.0474998	8.308×10^{-5}	4.600×10^{-3}
12	16	72	0.0193063	0.0107053	-0.0286223	1.025×10^{-4}	5.500×10^{-3}
13	44	103	0.0745549	0.0364416	0.0266264	1.164×10^{-4}	6.200×10^{-3}
14	10	154	0.0233720	0.0153370	-0.0245566	1.547×10^{-4}	8.000×10^{-3}
15	3	41	0.0168392	0.0133506	-0.0310893	6.063×10^{-4}	3.100×10^{-2}
16	2	165	0.0589695	0.0350183	0.0110409	9.861×10^{-4}	4.900×10^{-2}

Table B.26: Eigenvector centrality. Mean grouped by spectral clustering group. Wilcoxon test for difference in means (with PSP) with Benjamini and Hochberg correction for multiple comparisons

	group	n	mean	median	δ mean	p	p.adj
1	53	405	0.2225620	0.2132353	0.1143369	1.156×10^{-22}	7.500×10^{-21}
2	34	215	0.2365085	0.2000000	0.1282833	1.458×10^{-10}	9.300×10^{-9}
3	6	79	0.0837273	0.0000000	-0.0244978	5.177×10^{-9}	3.300×10^{-7}
4	9	65	0.0746643	0.0317460	-0.0335608	2.959×10^{-6}	1.800×10^{-4}
5	44	103	0.2333367	0.1961538	0.1251116	9.597×10^{-6}	5.900×10^{-4}
6	58	37	0.3054424	0.1823362	0.1972173	1.648×10^{-4}	9.900×10^{-3}
7	5	112	0.1165576	0.0503507	0.0083325	4.206×10^{-4}	2.500×10^{-2}
8	26	120	0.0987124	0.0663918	-0.0095127	5.520×10^{-4}	3.200×10^{-2}
9	7	129	0.1675753	0.0000000	0.0593502	6.092×10^{-4}	3.500×10^{-2}

Table B.27: transitivity0 Wilcoxon

RProjects/chapter_community_detection/R/group5/centrality_measures/plot_histogram_transitivity0bygroup.R

Rank	group	n	mean	median	δ mean	p	p.adj
1	53	405	14.3382716	15	5.1869844	8.319×10^{-34}	5.400×10^{-32}
2	7	129	4.2015504	3	-4.9497369	1.538×10^{-19}	9.800×10^{-18}
3	6	79	3.5443038	3	-5.6069834	1.042×10^{-16}	6.600×10^{-15}
4	34	215	12.1674419	13	3.0161546	4.168×10^{-8}	2.600×10^{-6}
5	23	72	5.0694444	4	-4.0818428	8.999×10^{-8}	5.500×10^{-6}
6	25	27	4.4814815	3	-4.6698058	5.968×10^{-5}	3.600×10^{-3}
7	43	150	11.3266667	10	2.1753794	9.209×10^{-5}	5.400×10^{-3}
8	11	56	5.7500000	6	-3.4012872	3.116×10^{-4}	1.800×10^{-2}
9	9	65	6.0615385	6	-3.0897488	3.483×10^{-4}	2.000×10^{-2}
10	5	112	6.6250000	5	-2.5262872	5.319×10^{-4}	3.000×10^{-2}

Table B.28: kcoreness deviation from mean by community NOT logarithmic Wilcoxon

RProjects/chapter_community_detection/R/group5/centrality_measures/plot_histogram_closenessbygroup.R

B.4 Supplementary Community Detection

B.4.1 Spectral methods using the modularity matrix

Following the notation in [289]. For a community structure or any vertex property the modularity is the difference between the number of edges between nodes sharing the property found in the network and those that might be expected to occur by chance.

The number of edges occurring between nodes that share the same property is

$$\sum_{\text{edges}_{i,j}} \delta_{g_i,g_j} = \frac{1}{2}\sum_{i,j} A_{i,j}\delta_{g_i,g_j} \tag{B.5}$$

where δ is the Kronecker delta and the factor of half accounts for the double counting of node pairs ie $A_{i,j}$ is identical to $A_{j,i}$ and $\mathbf{g}$ is a vector of group allocations.

The null model is obtained by rewiring the network but preserving the degree. In the rewiring model each edge is divided into two stubs and each stub can connected to $2m-1$ other stubs with a probability $\frac{1}{2m-1}$. The expected number of edges between two nodes is thus:

$$\frac{k_i k_j}{2m-1} \tag{B.6}$$

and the expected number of edges between groups is

$$\frac{1}{2}\sum_{i,j}\frac{k_i k_j}{2m}\delta_{g_i,g_j} \tag{B.7}$$

where $2m$ is used for $2m-1$ as the value is likely to be similar in large networks.

The difference in the number of edges within the groups from the null model is the difference between equation B.5 and equation B.7. The modularity is the number of edges expressed as the fraction of all edges obtained by dividing the difference in expected and actual edges between groups by the total number of edges in the network $2m$.

$$Q = \frac{1}{2m}\sum_{i,j} A_{i,j} - \frac{k_i k_j}{2m}\delta_{g_i,g_j} \tag{B.8}$$

defining the modularity as the difference between the actual number of edges and expected number, the modularity matrix $\mathbf{B}$ is:

$$\mathbf{B}_{i,j} = A_{i,j} - P_{i,j} = A_{i,j} - \frac{k_i k_j}{2m} \tag{B.9}$$

as before we define an index matrix s of community membership and write the modularity from Eq.(B.8) as

$$Q = \frac{1}{2m}\sum_{i,j}[A_{i,j} - P_{i,j}]\delta(g_i, g_j) \tag{B.10}$$

Rewriting the group allocation for two groups with s either +1 if in group 1 or -1 if in group 2.

$$\delta(g_i, g_j) = \frac{1}{2}(s_i s_j + 1) \tag{B.11}$$

$$Q = \frac{1}{4m}\sum_{i,j}[A_{i,j} - P_{i,j}](s_i s_j + 1) \tag{B.12}$$

$$Q = \frac{1}{4m}\sum_{i,j}[B_{i,j}](s_i s_j + 1) \tag{B.13}$$

as

$$\sum_{i,j} B_{i,j} = 0 \tag{B.14}$$

$$Q = \frac{1}{4m}\sum_{i,j}[B_{i,j}](s_i, s_j) \tag{B.15}$$

or

$$Q = \frac{1}{4m}\sum_{i,j}[A_{i,j} - P_{i,j}](s_i, s_j) \tag{B.16}$$

as a quadratic form:

$$Q = \frac{1}{4m}\mathbf{s}^T\mathbf{B}\mathbf{s} \tag{B.17}$$

we now seek to maximise Q by finding the eigenvector corresponding to the largest eigenvalue of the modularity matrix $\mathbf{B}$. The restriction of s to integer values is circumvented by ascribing s_i to the sign of the eigenvector v_i of $\mathbf{B}$ with maximum eigenvalue.

We can extend this to multiple groups by recursively partitioning the network based on the leading eigenvector of the partition and stopping partitioning when the partition does not result in a further increase in the modularity for the subdivision of the community (see Newman (2006) for further details).

The change in modularity of bisecting group g is[289][17].

[17] p510

$$\Delta Q = \frac{1}{4,} \sum_{i,j \in g} B_{i,j}^{(g)} s_i s_j \tag{B.18}$$

where the modularity matrix for the bisection is

$$B_{ij}^{P(g)} = B_{i,j} - \delta_{i,j} \sum_{k \in g} B_{i,k} \tag{B.19}$$

This is the method implemented by McLean [266]. This implementation include a fine tuning step where the node which being moved from one group to another results in the maximum gain in modularity is fixed in its community and the process is repeated until all nodes are moved. This process is repeated until the change in modularity is less than Δ ε[266]. The fine tuning step is the algorithm from [293] and [289][18].

B.4.2 Community statistics

Conductance has been used as a measure of the optimum size of communities[243], it is a community level statistic. In the same way that we were interested in importance measures for individual nodes in the previous chapters we are interested in measures of the properties of the discovered communities in addition to the global modularity property (Q)

A number of statistics can be calculated to characterise the properties of the communities detected such as how much they form a cohesive group isolated from the rest of the graph (conductance) and the number of edges outwith the group and within the group[134]. Statistics can also be calculated for individual vertices showing their relation to the community (for example are they well embedded in the community or peripheral). The notation and description here follows Fortunato[134] *exactly* and is for illustrative purposes. The following are defined on communities

Internal connectedness

- Internal degree k_c^{int}

$$k_c^{int} = \sum_{i,j \in C} A_{i,j} \tag{B.20}$$

- Average internal degree $k_c^{avg-int}$

$$k_c^{avg-int} = \frac{k_C^{int}}{n_c} \tag{B.21}$$

[18] p503

- Internal edge density δ_C^{int} The ratio of the number of internal edges to all possible internal edges $n_c(n_c-1)$

$$\delta_C^{int} = \frac{k_c^{int}}{n_C(n_C-1)} \tag{B.22}$$

External connectedness measures

- External degree or *cut* k_c^{ext}

$$k_c^{ext} = \sum_{i\in C, j\notin C} A_{i,j} \tag{B.23}$$

- Average external degree $k_c^{avg-ext}$ or *expansion*

$$k_c^{avg-ext} = \frac{k_c^{ext}}{n_C} \tag{B.24}$$

- External degree density or *cut ratio* δ_C^{ext}. The ratio of external degree to all possible external edges $n_c(n-n_c)$

$$\delta_C^{ext} = \frac{k_c^{ext}}{n_C(n_C-1)} \tag{B.25}$$

Other

- Total degree or *volume* $k_c = k_c^{int} + k_c^{ext}$
- Average degree k_c^{avg}

$$k_c^{avg} = \frac{k_C}{n_C} \tag{B.26}$$

- Conductance

$$C_c = \frac{k_c^{ext}}{k_c} \tag{B.27}$$

where c represents community and k degree, n is the number of nodes in the graph and n_c in community c

Vertex measures can also be calculated for individual vertices. Vertex internal and external degree are number of edges of a vertex linked to a member of the community or outwith the community respectively. Embeddedness ζ is the ratio of the internal degree to degree for a vertex and the mixing parameter μ is the ratio of external degree to the degree of a vertex[134].

What these measures provide are measures to assess the nature and quality of communities across a partition rather to augment the modularity which allows us to look at

the global quality of the partition. This will help select an algorithm or algorithms for use in community based GSA by choosing algorithms that have desirable properties distributed throughout groups rather than concentrated in a single group (for example a low conductance - equation B.27)

B.5 Generation of the synaptic network graph

Amstrong group https://homepages.inf.ed.ac.uk/jda/

Dr Oksana Sorokina

Dr Colin McLean

Prof J. Douglas Armstrong

Dr Ian Simpson

Dr Katharina Heil

B.6 Community detection of PSP

Dr Colin McClean

B.7 Samples

Generation of the summary statistics for the intelligence discovery and education discovery samples was carried out by Dr W. David Hill. Dr Hill performed the LD score regression and calculated genetic correlations between studies. Section 2.2.2. The information in this section is taken the part of a paper being prepared in collaboration with Dr Hill written by him.

Dr Okbay - provided samples for EA2 and performed the subset analysis.

Appendix C

Supplementary tables chapter 3

C.0.1 Degree

High degree nodes are enriched for terms associated with RNA binding, cell cycle and ubiquitin binding. This would be consistent with these high degree nodes being involved in central cellular processes[1].

C.0.1.0.1 Molecular function Molecular function table C.1 shows enrichment for ubiquitin protein ligase, RNA binding and transcription factor binding. The q values are not as significant as those for kcoreness and eigenvector centrality.

C.0.1.0.2 Biological process Biological process table C.2 shows enrichment for terms related to the cell cycle with low FDR values.

C.0.1.0.3 Cellular component Cellular component table C.3 shows enrichment for focal adhesion and cell junction terms, ribonucleoprotein and cytoskeletal fibre.

C.0.1.1 Low degree

GO enrichment for Cellular component has two significant terms (table C.4).

Table C.5 shows pathway enrichment using ToppGene.

[1] `source('~/RProjects/chapter3/R/sig_magma_genes/central_genes_ToppGene/degree/5_0_clip_centrality_degree_high_mf_n_10.R')`

ID	Molecular Function	n	n PSP	p	q FDR B H
GO:0031625	ubiquitin protein ligase binding	51	117	8.81×10^{-22}	9.12×10^{-19}
GO:0044389	ubiquitin-like protein ligase binding	52	123	1.88×10^{-21}	9.74×10^{-19}
GO:0008134	transcription factor binding	58	156	1.53×10^{-20}	5.26×10^{-18}
GO:0019904	protein domain specific binding	87	329	3.13×10^{-19}	8.11×10^{-17}
GO:0003723	RNA binding	120	554	5.07×10^{-19}	1.05×10^{-16}
GO:0044877	protein-containing complex binding	111	513	2.27×10^{-17}	3.91×10^{-15}
GO:0003720	telomerase activity	21	30	5.04×10^{-15}	7.00×10^{-13}
GO:0003964	RNA-directed DNA polymerase activity	22	33	5.41×10^{-15}	7.00×10^{-13}
GO:0034061	DNA polymerase activity	23	37	1.23×10^{-14}	1.41×10^{-12}
GO:0019900	kinase binding	78	325	1.50×10^{-14}	1.55×10^{-12}

Table C.1: **High degree genes. Molecular function - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Molecular Function tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category (e.g. MF,CC,BP): 186

C.0.2 Betweenness ToppGene

C.0.2.1 High betweenness

C.0.2.1.1 Molecular function (table C.6)

C.0.2.1.2 Biological process (table C.7)

C.0.2.1.3 Cellular component (table C.8)

ID	Biological Process	n	n PSP	p	q FDR B H
GO:0007049	cell cycle	139	489	1.98×10^{-36}	1.25×10^{-32}
GO:0022402	cell cycle process	119	414	7.10×10^{-31}	2.25×10^{-27}
GO:0000280	nuclear division	101	319	1.43×10^{-29}	3.02×10^{-26}
GO:0048285	organelle fission	103	337	9.65×10^{-29}	1.53×10^{-25}
GO:0000278	mitotic cell cycle	95	296	3.44×10^{-28}	3.10×10^{-25}
GO:0140014	mitotic nuclear division	95	296	3.44×10^{-28}	3.10×10^{-25}
GO:1903047	mitotic cell cycle process	95	296	3.44×10^{-28}	3.10×10^{-25}
GO:0044403	symbiotic process	106	364	1.23×10^{-27}	9.33×10^{-25}
GO:0051726	regulation of cell cycle	100	329	1.33×10^{-27}	9.33×10^{-25}
GO:0044419	interspecies interaction between organisms	106	368	3.48×10^{-27}	2.20×10^{-24}

Table C.2: **High degree genes. Biological Process - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Biological process tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category (e.g. MF,CC,BP): 1584

ID	Cellular Component	n	n PSP	p	q FDR B H
GO:1990904	ribonucleoprotein complex	81	303	2.75×10^{-18}	2.36×10^{-15}
GO:0015630	microtubule cytoskeleton	102	447	1.01×10^{-17}	4.33×10^{-15}
GO:0036464	cytoplasmic ribonucleoprotein granule	33	74	8.92×10^{-15}	2.55×10^{-12}
GO:0035770	ribonucleoprotein granule	33	78	5.84×10^{-14}	1.25×10^{-11}
GO:1902494	catalytic complex	86	392	1.07×10^{-13}	1.84×10^{-11}
GO:0030055	cell-substrate junction	62	244	7.12×10^{-13}	1.02×10^{-10}
GO:0005925	focal adhesion	61	239	9.13×10^{-13}	1.12×10^{-10}
GO:0000974	Prp19 complex	19	31	3.66×10^{-12}	3.92×10^{-10}
GO:0099513	polymeric cytoskeletal fiber	95	482	4.16×10^{-12}	3.96×10^{-10}
GO:0005815	microtubule organizing center	60	246	1.21×10^{-11}	1.04×10^{-9}

Table C.3: **High degree genes. Cellular Component - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Biological process tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category (e.g. MF,CC,BP): 159

ID	Cellular Component	p	q FDR B H	n	n PSP
GO:0031226	intrinsic component of plasma membrane	1.564×10^{-5}	7.462×10^{-3}	51	331
GO:0005887	integral component of plasma membrane	4.228×10^{-5}	1.008×10^{-2}	47	307

Table C.4: ToppGene GO: Cellular Component. 10 centile cwpsp.txtp Degree centrality = p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP.

ID	Pathway	p	q FDR B H	n	n PSP
M5889	Ensemble of genes encoding extracellular matrix and extracellular matrix-associated proteins	5.689×10^{-6}	5.831×10^{-3}	19	91
1269590	Adenylate cyclase activating pathway	9.603×10^{-5}	4.798×10^{-2}	4	5
1269923	Transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds	2.551×10^{-4}	4.798×10^{-2}	5	10
1272485	Aldosterone synthesis and secretion	2.896×10^{-4}	4.798×10^{-2}	9	34
1269930	Metal ion SLC transporters	3.277×10^{-4}	4.798×10^{-2}	3	3

Table C.5: ToppGene Pathway. 10 centile cwpsp.txtp = p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP

```
source('~/RProjects/chapter3/R/sig_magma_genes/central_genes_toppgene/betweenness/5_0_clip_centrality_bet_hig_mf.R')
```

ID	Molecular Function	n	n PSP	p	q FDR B H
GO:0044877	protein-containing complex binding	120	513	1.74×10^{-22}	1.83×10^{-19}
GO:0031625	ubiquitin protein ligase binding	51	117	5.84×10^{-22}	3.07×10^{-19}
GO:0044389	ubiquitin-like protein ligase binding	52	123	1.24×10^{-21}	4.34×10^{-19}
GO:0019904	protein domain specific binding	89	329	1.02×10^{-20}	2.69×10^{-18}
GO:0019900	kinase binding	76	325	9.81×10^{-14}	2.06×10^{-11}
GO:0008234	cysteine-type peptidase activity	54	189	1.31×10^{-13}	2.29×10^{-11}
GO:0008134	transcription factor binding	48	156	1.64×10^{-13}	2.45×10^{-11}
GO:0005102	signaling receptor binding	87	405	2.09×10^{-13}	2.74×10^{-11}
GO:0050839	cell adhesion molecule binding	69	288	5.05×10^{-13}	5.88×10^{-11}
GO:0019901	protein kinase binding	69	295	1.77×10^{-12}	1.86×10^{-10}

Table C.6: **High betweenness genes. Cellular Component - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Biological process tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category

ID	Biological Process	n	n PSP	p	q FDR B H
GO:0007049	cell cycle	124	489	5.10×10^{-27}	3.33×10^{-23}
GO:0051247	positive regulation of protein metabolic process	117	463	4.11×10^{-25}	1.34×10^{-21}
GO:0042981	regulation of apoptotic process	108	416	5.38×10^{-24}	1.17×10^{-20}
GO:0043067	regulation of programmed cell death	108	423	2.44×10^{-23}	3.99×10^{-20}
GO:0009894	regulation of catabolic process	95	350	2.16×10^{-22}	2.82×10^{-19}
GO:0006508	proteolysis	105	417	4.31×10^{-22}	4.29×10^{-19}
GO:0010941	regulation of cell death	112	464	4.60×10^{-22}	4.29×10^{-19}
GO:0032270	positive regulation of cellular protein metabolic process	106	433	2.89×10^{-21}	2.36×10^{-18}
GO:0009057	macromolecule catabolic process	106	443	2.02×10^{-20}	1.47×10^{-17}
GO:0043066	negative regulation of apoptotic process	73	240	3.94×10^{-20}	2.36×10^{-17}

Table C.7: **High betweenness genes. Biological Process - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Biological process tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category

source('~/RProjects/chapter3/R/sig_magma_genes/central_genes_toppgene/betweenness/5_0_clip_centrality_bet_hig_bp.R')

ID	Cellular Component	n	n PSP	p	q FDR B H
GO:0015630	microtubule cytoskeleton	98	447	4.56×10^{-16}	4.02×10^{-13}
GO:0005814	centriole	55	194	7.79×10^{-14}	3.43×10^{-11}
GO:0005813	centrosome	52	182	2.92×10^{-13}	8.59×10^{-11}
GO:0005815	microtubule organizing center	62	246	6.14×10^{-13}	1.35×10^{-10}
GO:0030990	intraciliary transport particle	51	212	5.95×10^{-10}	8.73×10^{-8}
GO:0005929	cilium	51	212	5.95×10^{-10}	8.73×10^{-8}
GO:0098589	membrane region	42	160	1.37×10^{-9}	1.72×10^{-7}
GO:0005819	spindle	35	120	1.74×10^{-9}	1.92×10^{-7}
GO:0099513	polymeric cytoskeletal fiber	87	482	3.13×10^{-9}	2.46×10^{-7}
GO:0045121	membrane raft	41	158	3.20×10^{-9}	2.46×10^{-7}

Table C.8: **High betweenness genes. Cellular Component - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Biological process tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category

C.0.2.2 Low betweenness

C.0.2.2.1 Cellular component

Cellular component table C.9

ID	Cellular Component	n	n PSP	p	q FDR B H
GO:0031226	intrinsic component of plasma membrane	61	331	1.85×10^{-6}	9.76×10^{-4}
GO:0005887	integral component of plasma membrane	56	307	7.36×10^{-6}	1.94×10^{-3}

Table C.9: **Low betweenness genes. Cellular Component - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Biological process tested gene set is below 10th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category

C.0.2.2.2 Pathway

Extra cellular matrix table C.10. Glycosylphosphatidylinositol proteins (GPI proteins) are directed to the endoplasmic reticulum.

ID	Pathway	n	n PSP	p	q FDR B H
M5889	Ensemble of genes encoding extracellular matrix and extracellular matrix-associated proteins	21	91	7.23×10^{-6}	8.28×10^{-3}
M1268710	Post-translational modification: synthesis of GPI-anchored proteins	6	11	9.13×10^{-5}	4.00×10^{-2}
M5884	Ensemble of genes encoding core extracellular matrix including ECM glycoproteins, collagens and proteoglycans	10	31	1.05×10^{-4}	4.00×10^{-2}

Table C.10: ToppGene Pathway. bet 10 centile cwpsp.txtp = p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP

C.0.3 Closeness Centrality ToppGene

C.0.3.0.1 Molecular function Molecular function Gene Ontology Enrichment High Closeness centrality in table C.11

ID	Molecular Function	n	n PSP	p	q FDR B H
GO:0003723	RNA binding	167	554	2.11×10^{-50}	2.14×10^{-47}
GO:0031625	ubiquitin protein ligase binding	57	117	1.38×10^{-27}	7.00×10^{-25}
GO:0044389	ubiquitin-like protein ligase binding	58	123	4.10×10^{-27}	1.39×10^{-24}
GO:0051082	unfolded protein binding	31	56	2.20×10^{-17}	5.57×10^{-15}
GO:0044877	protein-containing complex binding	110	513	4.36×10^{-17}	8.84×10^{-15}
GO:0019904	protein domain specific binding	81	329	5.59×10^{-16}	9.46×10^{-14}
GO:0003729	mRNA binding	39	100	6.55×10^{-15}	9.50×10^{-13}
GO:0008134	transcription factor binding	50	156	8.56×10^{-15}	1.09×10^{-12}
GO:0005102	signaling receptor binding	86	405	7.31×10^{-13}	8.25×10^{-11}
GO:0140097	catalytic activity, acting on DNA	31	76	1.13×10^{-12}	1.15×10^{-10}

Table C.11: **High Closeness genes. Molecular Function - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Molecular function tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category

C.0.3.0.2 Biological Process Biological process table C.12

C.0.3.0.3 Cellular component Cellular component table C.13

C.0.3.1 Low closeness

Note human phenotype epilepsy and disease enriched[2]

[2] '/RProjects/chapter3/R/sig_magma_genes/central_genes_toppgene/closeness/low_closeness/5_0_clip_centrality_clo_low_cc_n_10.R'

ID	Biological Process	n	n PSP	p	q FDR B H
GO:0044265	cellular macromolecule catabolic process	131	372	4.77×10^{-46}	2.98×10^{-42}
GO:0009057	macromolecule catabolic process	143	443	1.34×10^{-45}	4.18×10^{-42}
GO:0016071	mRNA metabolic process	106	260	2.87×10^{-43}	5.96×10^{-40}
GO:0044403	symbiotic process	125	364	3.12×10^{-42}	4.86×10^{-39}
GO:0044419	interspecies interaction between organisms	125	368	1.23×10^{-41}	1.53×10^{-38}
GO:0016032	viral process	118	348	6.85×10^{-39}	7.13×10^{-36}
GO:0006402	mRNA catabolic process	75	180	1.13×10^{-30}	1.00×10^{-27}
GO:0006401	RNA catabolic process	76	186	2.10×10^{-30}	1.63×10^{-27}
GO:0010608	posttranscriptional regulation of gene expression	82	225	1.11×10^{-28}	7.69×10^{-26}
GO:0034655	nucleobase-containing compound catabolic process	80	217	2.37×10^{-28}	1.48×10^{-25}

Table C.12: **High Closeness genes. Biological Process - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Biological process tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category

C.0.3.1.1 Molecular function Molecular function table C.14. Closeness is more normally distributed it appears the low part of its distribution is also more discriminating than those with power law distribution that have a number of genes separated by small absolute values at the lower end of the distribution. Genes encoding proteins with low closeness centrality are enriched for molecular functions involving ion channels and trans-membrane transporters.

C.0.3.1.2 Cellular component Trans-membrane transporters are also integral part of synaptic and post synaptic membrane(table C.15). The proteins encoded by these genes, the ion channels and synaptic membrane components, are also on the periphery of the network.

ID	Cellular Component	n	n PSP	p	q FDR B H
GO:1990904	ribonucleoprotein complex	102	303	5.76×10^{-33}	4.89×10^{-30}
GO:0005925	focal adhesion	77	239	4.94×10^{-23}	2.10×10^{-20}
GO:0030055	cell-substrate junction	77	244	2.18×10^{-22}	6.15×10^{-20}
GO:0005912	adherens junction	85	314	8.51×10^{-20}	1.80×10^{-17}
GO:0070161	anchoring junction	85	319	2.64×10^{-19}	4.47×10^{-17}
GO:0035770	ribonucleoprotein granule	36	78	1.03×10^{-16}	1.30×10^{-14}
GO:0036464	cytoplasmic ribonucleoprotein granule	35	74	1.07×10^{-16}	1.30×10^{-14}
GO:0000974	Prp19 complex	20	31	2.08×10^{-13}	2.21×10^{-11}
GO:0101031	chaperone complex	16	21	1.04×10^{-12}	9.79×10^{-11}
GO:0005681	spliceosomal complex	22	40	1.30×10^{-12}	1.11×10^{-10}

Table C.13: **High Closeness genes. Cellular Component - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Biological process tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category

ID	Molecular Function	n	n PSP	p	q FDR B H
GO:0005215	transporter activity	66	419	2.40×10^{-5}	1.02×10^{-2}
GO:0022857	transmembrane transporter activity	61	378	2.44×10^{-5}	1.02×10^{-2}
GO:0004930	G protein-coupled receptor activity	10	24	3.91×10^{-5}	1.09×10^{-2}
GO:0043176	amine binding	4	4	1.01×10^{-4}	2.10×10^{-2}
GO:0046873	metal ion transmembrane transporter activity	36	203	2.12×10^{-4}	3.19×10^{-2}
GO:0015075	ion transmembrane transporter activity	53	340	2.34×10^{-4}	3.19×10^{-2}
GO:0022803	passive transmembrane transporter activity	39	230	3.05×10^{-4}	3.19×10^{-2}
GO:0015267	channel activity	39	230	3.05×10^{-4}	3.19×10^{-2}
GO:0008324	cation transmembrane transporter activity	46	288	3.61×10^{-4}	3.36×10^{-2}
GO:0015318	inorganic molecular entity transmembrane transporter activity	50	324	4.60×10^{-4}	3.86×10^{-2}

Table C.14: **Low Closeness genes. Molecular Function - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Biological process tested gene set is below 10th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category

ID	Cellular Component	n	n PSP	p	q FDR B H
GO:0031226	intrinsic component of plasma membrane	72	331	8.90×10^{-12}	4.95×10^{-9}
GO:0005887	integral component of plasma membrane	66	307	1.41×10^{-10}	3.91×10^{-8}
GO:1990351	transporter complex	30	132	7.50×10^{-6}	1.35×10^{-3}
GO:1902495	transmembrane transporter complex	29	127	9.72×10^{-6}	1.35×10^{-3}
GO:0034702	ion channel complex	27	117	1.62×10^{-5}	1.57×10^{-3}
GO:0034703	cation channel complex	24	98	1.69×10^{-5}	1.57×10^{-3}
GO:0099055	integral component of postsynaptic membrane	19	72	4.33×10^{-5}	3.22×10^{-3}
GO:0072534	perineuronal net	5	6	5.12×10^{-5}	3.22×10^{-3}
GO:0098936	intrinsic component of postsynaptic membrane	20	79	5.22×10^{-5}	3.22×10^{-3}
GO:0099699	integral component of synaptic membrane	22	96	1.13×10^{-4}	6.28×10^{-3}

Table C.15: **Low Closeness genes. Cellular Component - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Biological process tested gene set is below 10th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category

C.0.4 Eigenvector centrality

C.0.4.1 High eigenvector centrality

C.0.4.1.1 Molecular function Molecular function table C.16 shows over representation of ubiquitin, RNA binding and ribosomal components. Similar to other centrality measures but very strong over representation[3].

ID	Molecular Function	n	n PSP	p	q FDR B H
GO:0003723	RNA binding	198	554	7.00×10^{-79}	7.01×10^{-76}
GO:0003735	structural constituent of ribosome	54	103	2.82×10^{-28}	1.41×10^{-25}
GO:0031625	ubiquitin protein ligase binding	56	117	1.15×10^{-26}	3.83×10^{-24}
GO:0044389	ubiquitin-like protein ligase binding	57	123	3.21×10^{-26}	8.03×10^{-24}
GO:0003729	mRNA binding	51	100	5.76×10^{-26}	1.15×10^{-23}
GO:0005198	structural molecule activity	82	303	5.13×10^{-19}	8.55×10^{-17}
GO:0019843	rRNA binding	25	34	1.26×10^{-18}	1.80×10^{-16}
GO:0003727	single-stranded RNA binding	21	31	1.19×10^{-14}	1.49×10^{-12}
GO:0045296	cadherin binding	59	224	4.31×10^{-13}	4.80×10^{-11}
GO:0003730	mRNA 3'-UTR binding	20	32	5.22×10^{-13}	5.22×10^{-11}

Table C.16: **High Eigenvector centrality genes. Molecular Function - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Molecular function tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category

C.0.4.1.2 Biological process Biological process table C.17

C.0.4.1.3 Cellular component Ribosomal components are significantly enriched (table C.18). PRP is pre-ribosomal rna processing complx 19.snRNP are small ribonuclear protein. Enrichment for ribosomes and for focal adhesion. Twenty eight of 34 U5 snRNP genes are identified [434].

[3] '/RProjects/chapter3/R/sig_magma_genes/central_genes_toppgene/eigenvector/5_0_clip_centrality_eig_mfunc_n_10.R')

ID	Biological Process	n	n PSP	p	q FDR B H
GO:0016071	mRNA metabolic process	143	260	1.30×10^{-83}	7.64×10^{-80}
GO:0009057	macromolecule catabolic process	159	443	1.09×10^{-59}	3.20×10^{-56}
GO:0044265	cellular macromolecule catabolic process	144	372	5.80×10^{-58}	9.34×10^{-55}
GO:0006402	mRNA catabolic process	101	180	6.34×10^{-58}	9.34×10^{-55}
GO:0006401	RNA catabolic process	102	186	3.28×10^{-57}	3.87×10^{-54}
GO:0044403	symbiotic process	139	364	1.05×10^{-54}	1.03×10^{-51}
GO:0044419	interspecies interaction between organisms	139	368	5.35×10^{-54}	4.50×10^{-51}
GO:0016032	viral process	134	348	6.30×10^{-53}	4.64×10^{-50}
GO:0034655	nucleobase-containing compound catabolic process	104	217	7.76×10^{-51}	5.08×10^{-48}
GO:0019439	aromatic compound catabolic process	104	232	2.67×10^{-47}	1.58×10^{-44}

Table C.17: **High Eigenvector centrality genes. Biological Process - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Molecular function tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category

ID	Cellular Component	n	n PSP	p	q FDR B H
GO:1990904	ribonucleoprotein complex	144	303	4.68×10^{-73}	3.68×10^{-70}
GO:0022626	cytosolic ribosome	60	97	3.21×10^{-37}	1.26×10^{-34}
GO:0044391	ribosomal subunit	61	124	4.99×10^{-30}	1.31×10^{-27}
GO:0005840	ribosome	65	141	6.93×10^{-30}	1.36×10^{-27}
GO:0000974	Prp19 complex	29	31	1.48×10^{-27}	2.33×10^{-25}
GO:0005681	spliceosomal complex	32	40	1.11×10^{-25}	1.46×10^{-23}
GO:0071013	catalytic step 2 spliceosome	27	29	1.49×10^{-25}	1.67×10^{-23}
GO:0097525	spliceosomal snRNP complex	29	35	3.56×10^{-24}	3.49×10^{-22}
GO:0005925	focal adhesion	78	239	7.67×10^{-24}	6.70×10^{-22}
GO:0005682	U5 snRNP	28	34	3.16×10^{-23}	2.48×10^{-21}

Table C.18: **High Eigenvector centrality genes. Cellular Component - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Molecular function tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category

C.0.4.2 Low eigenvector centrality

C.0.4.2.1 Molecular function Low eigenvector centrality PSP genes are enriched for molecular function terms related to trans-membrane transporter activity and voltage gated ion channels (table C.19).

There is no significant enrichment for biological process terms .

ID	Molecular Function	n	n PSP	p	q FDR B H
GO:0022857	transmembrane transporter activity	69	378	5.33×10^{-8}	2.75×10^{-5}
GO:0005215	transporter activity	74	419	6.69×10^{-8}	2.75×10^{-5}
GO:0022803	passive transmembrane transporter activity	47	230	3.54×10^{-7}	7.28×10^{-5}
GO:0015267	channel activity	47	230	3.54×10^{-7}	7.28×10^{-5}
GO:0015075	ion transmembrane transporter activity	61	340	7.16×10^{-7}	1.18×10^{-4}
GO:0015318	inorganic molecular entity transmembrane transporter activity	58	324	1.54×10^{-6}	1.67×10^{-4}
GO:0005244	voltage-gated ion channel activity	29	118	1.62×10^{-6}	1.67×10^{-4}
GO:0022832	voltage-gated channel activity	29	118	1.62×10^{-6}	1.67×10^{-4}
GO:0008324	cation transmembrane transporter activity	53	288	1.96×10^{-6}	1.79×10^{-4}
GO:0046873	metal ion transmembrane transporter activity	41	203	2.97×10^{-6}	2.44×10^{-4}

Table C.19: **Low Eigenvector centrality genes. Molecular Function - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Molecular function tested gene set is 10th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category

C.0.4.2.2 Cellular component Low eigenvector centrality genes in the PSP network (lowest 10%) are enriched for transmembrane transporter complexes, ion channels and integral and intrinsic part of synaptic and post synaptic membrane (table C.20).

ID	Cellular Component	n	n PSP	p	q FDR B H
GO:0031226	intrinsic component of plasma membrane	75	331	3.06×10^{-13}	1.69×10^{-10}
GO:0005887	integral component of plasma membrane	69	307	5.24×10^{-12}	1.45×10^{-9}
GO:1990351	transporter complex	36	132	5.47×10^{-9}	9.30×10^{-7}
GO:1902495	transmembrane transporter complex	35	127	6.71×10^{-9}	9.30×10^{-7}
GO:0034702	ion channel complex	33	117	9.91×10^{-9}	1.10×10^{-6}
GO:0034703	cation channel complex	29	98	2.53×10^{-8}	2.33×10^{-6}
GO:0099055	integral component of postsynaptic membrane	21	72	3.01×10^{-6}	2.37×10^{-4}
GO:0099699	integral component of synaptic membrane	25	96	3.42×10^{-6}	2.37×10^{-4}
GO:0098936	intrinsic component of postsynaptic membrane	22	79	4.08×10^{-6}	2.51×10^{-4}
GO:0099240	intrinsic component of synaptic membrane	26	108	1.04×10^{-5}	5.78×10^{-4}

Table C.20: **Low Eigenvector centrality genes. Molecular Function - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Molecular function tested gene set is 10th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category

C.0.5 Transitivity

C.0.5.1 High transitivity

In comparison with other centrality measures only two significant molecular function terms are enriched in high transitivity nodes (table C.21).

ID	Molecular Function	n	n PSP	p	q FDR B H
GO:0003735	structural constituent of ribosome	29	103	1.45×10^{-8}	1.06×10^{-5}
GO:0005198	structural molecule activity	48	303	6.32×10^{-5}	2.31×10^{-2}

Table C.21: **High Transitivity NaN genes. Molecular Function - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Molecular function tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category (e.g. MF,CC,BP): 2

C.0.5.1.1 Biological process High transitivity genes are enriched for Biological Process terms related to mitochondrial translation (table C.22).

C.0.5.1.2 Cellular component High transitivity network genes are associated with mitochondrial protein complex and ribosomal subunits (table C.23). No disease enrichment was found using ToppGene to lookup DisGeNet. No enrichment was found of any gene ontology clade for low (bottom 10%) of transitivity genes.

ID	Biological Process	n	n PSP	p	q FDR B H
GO:0006415	translational termination	14	35	8.46×10^{-7}	2.13×10^{-3}
GO:0070126	mitochondrial translational termination	13	32	1.71×10^{-6}	2.13×10^{-3}
GO:0070125	mitochondrial translational elongation	13	32	1.71×10^{-6}	2.13×10^{-3}
GO:0043624	cellular protein complex disassembly	25	108	7.84×10^{-6}	6.98×10^{-3}
GO:0032543	mitochondrial translation	15	47	9.37×10^{-6}	6.98×10^{-3}
GO:0006414	translational elongation	46	279	3.13×10^{-5}	1.66×10^{-2}
GO:0006412	translation	46	279	3.13×10^{-5}	1.66×10^{-2}
GO:0140053	mitochondrial gene expression	15	52	3.65×10^{-5}	1.70×10^{-2}
GO:0043043	peptide biosynthetic process	46	283	4.54×10^{-5}	1.72×10^{-2}
GO:0006521	regulation of cellular amino acid metabolic process	12	36	4.62×10^{-5}	1.72×10^{-2}

Table C.22: **High Transitivity NaN genes. Biological Process - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Molecular function tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category (e.g. MF,CC,BP): 22

ID	Cellular Component	n	n PSP	p	q FDR B H
GO:0044391	ribosomal subunit	31	124	8.04×10^{-8}	3.86×10^{-5}
GO:0098798	mitochondrial protein complex	33	139	1.15×10^{-7}	3.86×10^{-5}
GO:0005840	ribosome	32	141	5.48×10^{-7}	1.23×10^{-4}
GO:0005761	mitochondrial ribosome	13	31	1.03×10^{-6}	1.39×10^{-4}
GO:0000313	organellar ribosome	13	31	1.03×10^{-6}	1.39×10^{-4}
GO:0005762	mitochondrial large ribosomal subunit	9	16	2.44×10^{-6}	2.34×10^{-4}
GO:0000315	organellar large ribosomal subunit	9	16	2.44×10^{-6}	2.34×10^{-4}
GO:0015934	large ribosomal subunit	18	64	8.18×10^{-6}	6.88×10^{-4}
GO:0000502	proteasome complex	13	38	1.50×10^{-5}	1.01×10^{-3}
GO:1905369	endopeptidase complex	13	38	1.50×10^{-5}	1.01×10^{-3}

Table C.23: **High Transitivity NaN genes. Cellular Component - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Molecular function tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category (e.g. MF,CC,BP): 26

C.0.6 kcoreness

C.0.6.1 High kcoreness

C.0.6.1.1 Molecular function Similar to eigenvector centrality, terms related to RNA binding and ubiquitin ligase binding are highly significantly enriched (table C.24)[4].

ID	Molecular Function	n	n PSP	p	q FDR B H
GO:0003723	RNA binding	188	554	2.64×10^{-58}	2.68×10^{-55}
GO:0044389	ubiquitin-like protein ligase binding	60	123	2.82×10^{-26}	1.43×10^{-23}
GO:0031625	ubiquitin protein ligase binding	58	117	7.13×10^{-26}	2.42×10^{-23}
GO:0003729	mRNA binding	48	100	1.05×10^{-20}	2.66×10^{-18}
GO:0003735	structural constituent of ribosome	43	103	8.30×10^{-16}	1.69×10^{-13}
GO:0101005	ubiquitinyl hydrolase activity	36	77	2.84×10^{-15}	4.12×10^{-13}
GO:0004843	thiol-dependent ubiquitin-specific protease activity	36	77	2.84×10^{-15}	4.12×10^{-13}
GO:0008134	transcription factor binding	53	156	1.02×10^{-14}	1.30×10^{-12}
GO:0044877	protein-containing complex binding	112	513	4.46×10^{-14}	5.04×10^{-12}
GO:0019843	rRNA binding	22	34	1.33×10^{-13}	1.35×10^{-11}

Table C.24: **High kcoreness genes. Molecular Function - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Molecular function tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category (e.g. MF,CC,BP): 180

C.0.6.1.2 Biological process In addition to mRNA catabolism, viral processes and interspecies interaction are enriched perhaps reflecting basic preserved functions across different organisms (table C.25).

C.0.6.1.3 Cellular component Ribonucleoprotein terms and splicasome terms similar to high eigenvector centrality enrichment are prominent (table C.26). The FDR q values are low and a substantial proportion of PSP components (29 of 31 for Prp19 complex) are identified. Adherens junction contains proteins,including cadherins, at cell to cell junctions.

[4] 'RProjects/chapter3/R/sig_magma_genes/central_genes_toppgene/kcoreness/5_0_clip_centrality_kco_high_mf_n_10.R'

ID	Biological Process	n	n PSP	p	q FDR B H
GO:0016071	mRNA metabolic process	142	260	3.73×10^{-74}	2.26×10^{-70}
GO:0009057	macromolecule catabolic process	166	443	4.75×10^{-57}	1.44×10^{-53}
GO:0044265	cellular macromolecule catabolic process	150	372	1.43×10^{-55}	2.88×10^{-52}
GO:0006402	mRNA catabolic process	100	180	2.06×10^{-51}	3.12×10^{-48}
GO:0006401	RNA catabolic process	100	186	1.34×10^{-49}	1.62×10^{-46}
GO:0044403	symbiotic process	140	364	2.98×10^{-48}	3.01×10^{-45}
GO:0044419	interspecies interaction between organisms	140	368	1.46×10^{-47}	1.27×10^{-44}
GO:0016032	viral process	135	348	8.48×10^{-47}	6.43×10^{-44}
GO:0034655	nucleobase-containing compound catabolic process	102	217	1.67×10^{-43}	1.13×10^{-40}
GO:0019439	aromatic compound catabolic process	102	232	3.61×10^{-40}	2.19×10^{-37}

Table C.25: **High kcoreness genes. Biological Process - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Biological Process tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category (e.g. MF,CC,BP): 1141

C.0.6.1.4 Others Murine phenotypes include embryonic lethality and human phenotypes are associated with neoplasm and limb abnormalities.

Disease enrichment reflects the involvement of core biological processess with prominent neoplasms and viral infections. The most enriched neurological disoder is Frontotemporal dementia (C0338451, 7th most enriched disorder, n in annotation 44, n in PSP 119, p 3.75×10^{-13} FDR q 3.87×10^{-10}).

ID	Cellular Component	n	n PSP	p	q FDR B H
GO:1990904	ribonucleoprotein complex	137	303	6.81×10^{-58}	5.48×10^{-55}
GO:0005925	focal adhesion	87	239	9.15×10^{-27}	3.68×10^{-24}
GO:0022626	cytosolic ribosome	53	97	1.86×10^{-26}	4.98×10^{-24}
GO:0000974	Prp19 complex	29	31	4.58×10^{-26}	8.40×10^{-24}
GO:0030055	cell-substrate junction	87	244	5.22×10^{-26}	8.40×10^{-24}
GO:0071013	catalytic step 2 spliceosome	27	29	3.59×10^{-24}	4.81×10^{-22}
GO:0005681	spliceosomal complex	32	40	4.61×10^{-24}	5.29×10^{-22}
GO:0097525	spliceosomal snRNP complex	29	35	1.04×10^{-22}	1.05×10^{-20}
GO:0005682	U5 snRNP	28	34	8.18×10^{-22}	7.31×10^{-20}
GO:0005912	adherens junction	94	314	1.07×10^{-21}	8.59×10^{-20}

Table C.26: **High kcoreness genes. Cellular Component - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Cellular Component tested gene set is 90th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category (e.g. MF,CC,BP): 181

C.0.6.2 Low kcoreness

C.0.6.2.1 Cellular component Over representation of integral part of plasma membrane similar to low closeness the low kcore is also on the periphery of the network (table C.27).

ID	Cellular Component	n	n PSP	p	q FDR B H
GO:0031226	intrinsic component of plasma membrane	98	331	8.26×10^{-9}	5.35×10^{-6}
GO:0005887	integral component of plasma membrane	91	307	2.94×10^{-8}	9.51×10^{-6}

Table C.27: **Low kcoreness genes. Cellular Component - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Cellular Component tested gene set is 10th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category (e.g. MF,CC,BP): 2

Pathway (table 3.20) are enriched for extra cellular matrix and matrix associated proteins and post translational modification of GPI. A high proportion of the members of the pathway in the PSP are identified amongst low kcore genes.

Gene family ORA in ToppGene (table C.28) is enriched for calcium voltage gate channel subunits, fibronectin and solute carriers.

ID	Gene Family	n	n PSP	p	q FDR B H
752	Solute carriers	21	45	7.61×10^{-7}	1.58×10^{-4}
253	Calcium voltage-gated channel subunits	8	14	4.36×10^{-4}	4.53×10^{-2}
1210	ATPase phospholipid transporting	4	4	7.39×10^{-4}	4.77×10^{-2}
593	Fibronectin type III domain containing|I-set domain containing| Immunoglobulin like domain containing	15	42	1.11×10^{-3}	4.77×10^{-2}
594	Immunoglobulin like domain containing|Interleukin receptors| TIR domain containing	13	34	1.15×10^{-3}	4.77×10^{-2}

Table C.28: **Low kcoreness genes. Gene Family - Gene ontology enrichment analysis** using the ToppGene web client and all PSP genes as the background set. Gene Family tested gene set is 10th centile of degree distribution. p value; q FDR B H = q adjusted significance level False Discovery Rate using Benjamini and Hochberg adjustment; n= n genes annotated in test group; n PSP= n genes annotated in PSP. n ontology terms significantly (q FDR$>= 0.05$) enriched for this centrality measure and ontology category (e.g. MF,CC,BP): 5

Appendix D

Data sources and code

Supplementary data referred to in the text can be found at

`https://datasync.ed.ac.uk/index.php/s/ZjPL0bC6PFecQah`. Code paths in the text refer to the relative path from the folder `RProjects` and include this as the first directory name. Contact `gr.infthesis@gmail.com`.

Additional code is at `https://github.com/14760/RProjects`

Bibliography

[1] Alexandre Abraham et al. “Machine learning for neuroimaging with scikit-learn”. In: *Frontiers in neuroinformatics* 8 (2014), p. 14.

[2] Lada A Adamic and Natalie Glance. “The political blogosphere and the 2004 US election: divided they blog”. In: *Proceedings of the 3rd international workshop on Link discovery*. 2005, pp. 36–43.

[3] Yong-Yeol Ahn, James P Bagrow, and Sune Lehmann. “Link communities reveal multiscale complexity in networks”. In: *nature* 466.7307 (2010), pp. 761–764.

[4] Reka Albert. “Scale-free networks in cell biology”. In: *Journal of cell science* 118.21 (2005), pp. 4947–4957.

[5] Réka Albert, Hawoong Jeong, and Albert-László Barabási. “Diameter of the world-wide web”. In: *nature* 401.6749 (1999), pp. 130–131.

[6] Réka Albert, Hawoong Jeong, and Albert-László Barabási. “Error and attack tolerance of complex networks”. In: *nature* 406.6794 (2000), pp. 378–382.

[7] Rodrigo Aldecoa and Ignacio Marín. “Exploring the limits of community detection strategies in complex networks”. In: *Scientific reports* 3.1 (2013), pp. 1–11.

[8] A Alexa and J Rahnenführer. *Gene set enrichment analysis with topGO. R package version 2.40. 0*. 2020.

[9] Adrian Alexa and Jörg Rahnenführer. “Gene set enrichment analysis with topGO”. In: *Bioconductor Improv* 27 (2009).

[10] Adrian Alexa, Jörg Rahnenführer, and Thomas Lengauer. “Improved scoring of functional groups from gene expression data by decorrelating GO graph structure”. In: *Bioinformatics* 22.13 (2006), pp. 1600–1607.

[11] Diego Alonso-López et al. “APID database: redefining protein–protein interaction experimental evidences and binary interactomes”. In: *Database* 2019 (2019).

[12] J Ignacio Alvarez-Hamelin et al. “Large scale networks fingerprinting and visualization using the k-core decomposition”. In: *Advances in neural information processing systems*. 2006, pp. 41–50.

[13] Alessia Amelio and Clara Pizzuti. “Is normalized mutual information a fair measure for comparing community detection methods?” In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. 2015, pp. 1584–1585.

[14] Johansen B Amin, Gabrielle R Moody, and Lonnie P Wollmuth. “From bedside-to-bench: What disease-associated variants are teaching us about the NMDA receptor”. In: *The Journal of physiology* 599.2 (2021), pp. 397–416.

[15] Rosalind Arden et al. “The association between intelligence and lifespan is mostly genetic”. In: *International Journal of Epidemiology* 45.1 (2016), pp. 178–185.

[16] J Douglas Armstrong and Oksana Sorokina. “Evolution of the cognitive proteome: from static to dynamic network models”. In: *Advances in Systems Biology*. Springer, 2012, pp. 119–134.

[17] Michael Ashburner et al. “Gene ontology: tool for the unification of biology”. In: *Nature genetics* 25.1 (2000), pp. 25–29.

[18] Uri Ashery et al. “The Molecular Mechanisms Underlying Synaptic Transmission: A View of the Presynaptic Terminal”. In: *The Synapse*. Elsevier, 2014, pp. 21–109.

[19] Jean-Nicolas Audet et al. “Divergence in problem-solving skills is associated with differential expression of glutamate receptors in wild finches”. In: *Science advances* 4.3 (2018), eaao6369.

[20] Erich J Baker et al. “GeneWeaver: a web-based system for integrative functional genomics”. In: *Nucleic acids research* 40.D1 (2012), pp. D1067–D1076.

[21] Trygve E Bakken et al. “A comprehensive transcriptional map of primate brain development”. In: *Nature* 535.7612 (2016), pp. 367–375.

[22] David M Bannerman et al. “Hippocampal synaptic plasticity, spatial memory and anxiety”. In: *Nature reviews neuroscience* 15.3 (2014), pp. 181–192.

[23] Albert-László Barabási et al. *Network science*. Cambridge university press, 2016.

[24] Albert-László Barabási and Réka Albert. "Emergence of scaling in random networks". In: *science* 286.5439 (1999), pp. 509–512.

[25] Albert-László Barabási, Réka Albert, and Hawoong Jeong. "Mean-field theory for scale-free random networks". In: *Physica A: Statistical Mechanics and its Applications* 272.1-2 (1999), pp. 173–187.

[26] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. "Network medicine: a network-based approach to human disease". In: *Nature reviews genetics* 12.1 (2011), pp. 56–68.

[27] Sergio E Baranzini et al. "Pathway and network-based analysis of genome-wide association studies in multiple sclerosis". In: *Human molecular genetics* 18.11 (2009), pp. 2078–2090.

[28] Sergio E Baranzini et al. "Network-based multiple sclerosis pathway analysis with GWAS data from 15,000 cases and 30,000 controls". In: *The American Journal of Human Genetics* 92.6 (2013), pp. 854–865.

[29] Fredrik Barrenas et al. "Network properties of complex human disease genes identified through genome-wide association studies". In: *PloS one* 4.11 (2009), e8090.

[30] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. "Gephi: an open source software for exploring and manipulating networks". In: *Third international AAAI conference on weblogs and social media*. 2009.

[31] Vladimir Batagelj and Matjaz Zaversnik. "An O (m) algorithm for cores decomposition of networks". In: *arXiv preprint cs/0310049* (2003).

[32] G David Batty, Ian J Deary, and Linda S Gottfredson. "Premorbid (early life) IQ and later mortality risk: systematic review". In: *Annals of epidemiology* 17.4 (2007), pp. 278–288.

[33] Alex Bavelas. "A mathematical model for group structures". In: *Human organization* 7.3 (1948), pp. 16–30.

[34] Alex Bayes et al. "Comparative study of human and mouse postsynaptic proteomes finds high compositional conservation and abundance differences for key synaptic proteins". In: *PloS one* 7.10 (2012).

[35] Àlex Bayés et al. “Characterization of the proteome, diseases and evolution of the human postsynaptic density”. In: *Nature neuroscience* 14.1 (2011), pp. 19–21.

[36] Àlex Bayés et al. “Human post-mortem synapse proteome integrity screening for proteomic studies of postsynaptic complexes”. In: *Molecular brain* 7.1 (2014), pp. 1–11.

[37] Edward Beltrami. *Mathematical models for society and biology*. Academic Press, 2013. Chap. 7.

[38] Yoav Benjamini and Yosef Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[39] Beben Benyamin et al. “Childhood intelligence is heritable, highly polygenic and associated with FNBP1L”. In: *Molecular psychiatry* 19.2 (2014), pp. 253–258.

[40] Tapan K Bera et al. “POTE, a highly homologous gene family located on numerous chromosomes and expressed in prostate, ovary, testis, placenta, and prostate cancer”. In: *Proceedings of the National Academy of Sciences* 99.26 (2002), pp. 16975–16980.

[41] Lukas Bernauer et al. “Komagataella phaffii as Emerging Model Organism in Fundamental Research”. In: *Frontiers in Microbiology* 11 (2021), p. 3462.

[42] B Bettler and C Mulle. “AMPA and kainate receptors”. In: *Neuropharmacology* 34.2 (1995), pp. 123–139.

[43] Ginestra Bianconi and A-L Barabási. “Competition and multiscaling in evolving networks”. In: *EPL (Europhysics Letters)* 54.4 (2001), p. 436.

[44] Miriam Bibel and Yves-Alain Barde. “Neurotrophins: key regulators of cell fate and cell shape in the vertebrate nervous system”. In: *Genes & development* 14.23 (2000), pp. 2919–2937.

[45] Thóra K Bjarnadóttir et al. “Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse”. In: *Genomics* 88.3 (2006), pp. 263–273.

[46] Tim VP Bliss and Terje Lømo. “Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path”. In: *The Journal of physiology* 232.2 (1973), pp. 331–356.

[47] Vincent D Blondel et al. “Fast unfolding of communities in large networks”. In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[48] Miriam E Bocarsly et al. “Obesity diminishes synaptic markers, alters microglial morphology, and impairs cognitive function”. In: *Proceedings of the National Academy of Sciences* 112.51 (2015), pp. 15731–15736.

[49] Stefano Boccaletti et al. “Complex networks: Structure and dynamics”. In: *Physics reports* 424.4-5 (2006), pp. 175–308.

[50] Phillip Bonacich. “Factoring and weighting approaches to status scores and clique identification”. In: *Journal of mathematical sociology* 2.1 (1972), pp. 113–120.

[51] Phillip Bonacich. “Power and centrality: A family of measures”. In: *American journal of sociology* 92.5 (1987), pp. 1170–1182.

[52] Severin Borenstein. “Hubs and high fares: dominance and market power in the US airline industry”. In: *The RAND Journal of Economics* (1989), pp. 344–365.

[53] Stephen P Borgatti. “Centrality and network flow”. In: *Social networks* 27.1 (2005), pp. 55–71.

[54] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. “An expanded view of complex traits: from polygenic to omnigenic”. In: *Cell* 169.7 (2017), pp. 1177–1186.

[55] Ulrik Brandes, Stephen P Borgatti, and Linton C Freeman. “Maintaining the duality of closeness and betweenness centrality”. In: *Social Networks* 44 (2016), pp. 153–159.

[56] Ulrik Brandes et al. “On modularity clustering”. In: *IEEE transactions on knowledge and data engineering* 20.2 (2007), pp. 172–188.

[57] Alvis Brazma, Maria Krestyaninova, and Ugis Sarkans. “Standards for systems biology”. In: *Nature Reviews Genetics* 7.8 (2006), pp. 593–605.

[58] Anna D Broido and Aaron Clauset. “Scale-free networks are rare”. In: *Nature communications* 10.1 (2019), pp. 1–10.

[59] Morton B Brown. “400: A method for combining non-independent, one-sided tests of significance”. In: *Biometrics* (1975), pp. 987–992.

[60] Brendan K Bulik-Sullivan et al. “LD Score regression distinguishes confounding from polygenicity in genome-wide association studies”. In: *Nature genetics* 47.3 (2015), p. 291.

[61] Simon C Bull and Andrew J Doig. “Properties of protein drug target classes”. In: *PloS one* 10.3 (2015), e0117955.

[62] Carol J Bult et al. “Mouse genome database (MGD) 2019”. In: *Nucleic acids research* 47.D1 (2019), pp. D801–D806.

[63] Kate Burleson-Lesser et al. “K-core robustness in ecological and financial networks”. In: *Scientific reports* 10.1 (2020), pp. 1–14.

[64] Ronald S Burt. *Structural holes: The social structure of competition*. Harvard university press, 2009.

[65] Clare Bycroft et al. “The UK Biobank resource with deep phenotyping and genomic data”. In: *Nature* 562.7726 (2018), pp. 203–209.

[66] Francesco Cadini, Enrico Zio, and Cristina-Andreea Petrescu. “Using centrality measures to rank the importance of the components of a complex network infrastructure”. In: *International Workshop on Critical Information Infrastructures Security*. Springer. 2008, pp. 155–167.

[67] Na Cai, Karmel W Choi, and Eiko I Fried. “Reviewing the genetics of heterogeneity in depression: operationalizations, manifestations and etiologies”. In: *Human molecular genetics* 29.R1 (2020), R10–R18.

[68] Keith E Campbell et al. “Representing thoughts, words, and things in the UMLS”. In: *Journal of the American Medical Informatics Association* 5.5 (1998), pp. 421–431.

[69] Paul Charlesworth et al. “Canalization of genetic and pharmacological perturbations in developing primary neuronal activity patterns”. In: *Neuropharmacology* 100 (2016), pp. 47–55.

[70] Andrew Chatr-Aryamontri et al. “The BioGRID interaction database: 2017 update”. In: *Nucleic acids research* 45.D1 (2017), pp. D369–D379.

[71] Sreenivas Chavali et al. “Network properties of human disease genes with pleiotropic effects”. In: *BMC systems biology* 4.1 (2010), pp. 1–11.

[72] Jing Chen, Bruce J Aronow, and Anil G Jegga. “Disease candidate gene identification and prioritization using protein interaction networks”. In: *BMC bioinformatics* 10.1 (2009), p. 73.

[73] Jing Chen et al. “ToppGene Suite for gene list enrichment analysis and candidate gene prioritization”. In: *Nucleic acids research* 37.suppl_2 (2009), W305–W311.

[74] Xiumin Chen et al. “MAGUKs are essential, but redundant, in long-term potentiation”. In: *Proceedings of the National Academy of Sciences* 118.28 (2021).

[75] Neil R Clark and Avi Ma’ayan. “Introduction to statistical methods for analyzing large data sets: gene-set enrichment analysis”. In: *Science signaling* 4.190 (2011), tr4–tr4.

[76] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. “Finding community structure in very large networks”. In: *Physical review E* 70.6 (2004), p. 066111.

[77] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. “Power-law distributions in empirical data”. In: *SIAM review* 51.4 (2009), pp. 661–703.

[78] Reuven Cohen and Shlomo Havlin. “Scale-free networks are ultrasmall”. In: *Physical review letters* 90.5 (2003), p. 058701.

[79] Jonathan RI Coleman et al. “Biological annotation of genetic loci associated with intelligence in a meta-analysis of 87,740 individuals”. In: *Molecular psychiatry* 24.2 (2019), pp. 182–197.

[80] Graham L Collingridge, SJ Kehl, and H t McLennan. “Excitatory amino acids in synaptic transmission in the Schaffer collateral-commissural pathway of the rat hippocampus.” In: *The Journal of physiology* 334.1 (1983), pp. 33–46.

[81] Mark O Collins et al. “Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome”. In: *Journal of neurochemistry* 97 (2006), pp. 16–23.

[82] Ana Conesa et al. “A survey of best practices for RNA-seq data analysis”. In: *Genome biology* 17.1 (2016), p. 13.

[83] 1000 Genomes Project Consortium et al. “An integrated map of genetic variation from 1,092 human genomes”. In: *Nature* 491.7422 (2012), pp. 56–65.

[84] GTEx Consortium et al. “The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans”. In: *Science* 348.6235 (2015), pp. 648–660.

[85] International HapMap 3 Consortium et al. “Integrating common and rare genetic variation in diverse human populations”. In: *Nature* 467.7311 (2010), p. 52.

[86] Anis Contractor, Christophe Mulle, and Geoffrey T Swanson. “Kainate receptors coming of age: milestones of two decades of research”. In: *Trends in neurosciences* 34.3 (2011), pp. 154–163.

[87] Mark J Cowley et al. “PINA v2. 0: mining interactome modules”. In: *Nucleic acids research* 40.D1 (2012), pp. D862–D865.

[88] Mike DR Croning et al. “G2Cdb: the Genes to Cognition database”. In: *Nucleic acids research* 37.suppl_1 (2009), pp. D846–D851.

[89] Gabor Csardi, Tamas Nepusz, et al. “The igraph software package for complex network research”. In: *InterJournal, complex systems* 1695.5 (2006), pp. 1–9.

[90] David Curtis, Anna E Vine, and Jo Knight. “A simple method for assessing the strength of evidence for association at the level of the whole gene”. In: *Advances and applications in bioinformatics and chemistry: AABC* 1 (2008), p. 115.

[91] Petr Danecek et al. “The variant call format and VCFtools”. In: *Bioinformatics* 27.15 (2011), pp. 2156–2158.

[92] Leon Danon et al. “Comparing community structure identification”. In: *Journal of statistical mechanics: Theory and experiment* 2005.09 (2005), P09008.

[93] Hassan S Dashti et al. “Genome-wide association study identifies genetic loci for self-reported habitual sleep duration supported by accelerometer-derived estimates”. In: *Nature communications* 10.1 (2019), pp. 1–12.

[94] Gail Davies et al. “Genome-wide association studies establish that human intelligence is highly heritable and polygenic”. In: *Molecular psychiatry* 16.10 (2011), pp. 996–1005.

[95] Gail Davies et al. “Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N= 53 949)”. In: *Molecular psychiatry* 20.2 (2015), pp. 183–192.

[96] Gail Davies et al. “Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N= 112 151)”. In: *Molecular psychiatry* 21.6 (2016), pp. 758–767.

[97] Robert B Davies. “The distribution of a linear combination of $\chi 2$ random variables”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 29.3 (1980), pp. 323–333.

[98] Allan Peter Davis et al. “The comparative toxicogenomics database: update 2019”. In: *Nucleic acids research* 47.D1 (2019), pp. D948–D954.

[99] MO Dayhoff. “The origin and evolution of protein superfamilies.” In: *Federation Proceedings*. Vol. 35. 10. 1976, pp. 2132–2138.

[100] Christiaan A De Leeuw et al. “The statistical properties of gene-set analysis”. In: *Nature Reviews Genetics* 17.6 (2016), p. 353.

[101] Thomas Dean, Greg S Corrado, and Jonathon Shlens. “Three controversial hypotheses concerning computation in the primate cortex”. In: *Twenty-Sixth AAAI Conference on Artificial Intelligence*. 2012.

[102] Ian Deary. “Why do intelligent people live longer?” In: *Nature* 456.7219 (2008), pp. 175–176.

[103] Ian J Deary. “Human intelligence differences: a recent history”. In: *Trends in cognitive sciences* 5.3 (2001), pp. 127–130.

[104] Ian J Deary. *Intelligence: A very short introduction*. Oxford University Press, USA, 2001.

[105] Ian J Deary. “Cognitive epidemiology: Its rise, its current issues, and its challenges”. In: *Personality and individual differences* 49.4 (2010), pp. 337–343.

[106] Ian J Deary. “The stability of intelligence from childhood to old age”. In: *Current Directions in Psychological Science* 23.4 (2014), pp. 239–245.

[107] Ian J Deary and G David Batty. “Cognitive epidemiology”. In: *Journal of Epidemiology & Community Health* 61.5 (2007), pp. 378–384.

[108] Ian J Deary, Wendy Johnson, and Lorna M Houlihan. “Genetic foundations of human intelligence”. In: *Human genetics* 126.1 (2009), pp. 215–232.

[109] Ian J Deary et al. “Cognitive change and the APOE E4 allele”. In: *Nature* 418.6901 (2002), pp. 932–932.

[110] Ute Distler et al. “In-depth protein profiling of the postsynaptic density from mouse hippocampus using data-independent acquisition proteomics”. In: *Proteomics* 14.21-22 (2014), pp. 2607–2613.

[111] Luca Donetti and Miguel A Munoz. “Detecting network communities: a new systematic and efficient algorithm”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2004.10 (2004), P10012.

[112] Shaowei Dong and Nicholas J Provart. “Analyses of protein interaction networks using computational tools”. In: *Two-Hybrid Systems*. Springer, 2018, pp. 97–117.

[113] S van Dongen. “Graph clustering by flow simulation”. In: *PhD thesis, University of Utrecht* (2000).

[114] Sergey N Dorogovtsev, Alexander V Goltsev, and Jose Ferreira F Mendes. “K-core organization of complex networks”. In: *Physical review letters* 96.4 (2006), p. 040601.

[115] Ayse Dosemeci et al. “Composition of the synaptic PSD-95 complex”. In: *Molecular & Cellular Proteomics* 6.10 (2007), pp. 1749–1760.

[116] Eric Eaton and Rachael Mansbach. “A spin-glass model for semi-supervised community detection”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 26. 1. 2012.

[117] Clara C Elbers et al. “Using genome-wide pathway analysis to unravel the etiology of complex diseases”. In: *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 33.5 (2009), pp. 419–431.

[118] Paul Elliott and Tim C Peakman. “The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine”. In: *International journal of epidemiology* 37.2 (2008), pp. 234–244.

[119] Richard D Emes and Seth GN Grant. “Evolution of synapse complexity and diversity”. In: *Annual review of neuroscience* 35 (2012), pp. 111–131.

[120] Sabine Endele et al. “Mutations in GRIN2A and GRIN2B encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes”. In: *Nature genetics* 42.11 (2010), pp. 1021–1026.

[121] Anton J Enright, Stijn Van Dongen, and Christos A Ouzounis. “An efficient algorithm for large-scale detection of protein families”. In: *Nucleic acids research* 30.7 (2002), pp. 1575–1584.

[122] Paul Erdős and Alfréd Rényi. “On the evolution of random graphs”. In: *Publ. Math. Inst. Hung. Acad. Sci* 5.1 (1960), pp. 17–60.

[123] Ernesto Estrada. “Virtual identification of essential proteins within the protein interaction network of yeast”. In: *Proteomics* 6.1 (2006), pp. 35–40.

[124] Ernesto Estrada. “Combinatorial study of degree assortativity in networks”. In: *Physical review E* 84.4 (2011), p. 047101.

[125] Ernesto Estrada. “When local and global clustering of networks diverge”. In: *Linear Algebra and its Applications* 488 (2016), pp. 249–263.

[126] Iakes Ezkurdia et al. “Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes”. In: *Human molecular genetics* 23.22 (2014), pp. 5866–5878.

[127] Carol D Farr et al. “Proteomic analysis of native metabotropic glutamate receptor 5 protein complexes reveals novel molecular constituents”. In: *Journal of neurochemistry* 91.2 (2004), pp. 438–450.

[128] Francesco Fazio et al. “Targeting metabotropic glutamate receptors for the treatment of neuroinflammation”. In: *Current Opinion in Pharmacology* 38 (2018), pp. 16–23.

[129] Humberto A Filho, Jeaneth Machicao, and Odemir M Bruno. “A hierarchical model of metabolic machinery based on the k core decomposition of plant metabolic networks”. In: *PloS one* 13.5 (2018), e0195843.

[130] Hilary K Finucane et al. “Partitioning heritability by functional annotation using genome-wide association summary statistics”. In: *Nature genetics* 47.11 (2015), p. 1228.

[131] Walter M Fitch. “Aspects of molecular evolution”. In: *Annual review of genetics* 7.1 (1973), pp. 343–380.

[132] Melanie Föcking et al. “Proteomic and genomic evidence implicates the postsynaptic density in schizophrenia”. In: *Molecular psychiatry* 20.4 (2015), pp. 424–432.

[133] Melanie Föcking et al. “Proteomic analysis of the postsynaptic density implicates synaptic function and energy pathways in bipolar disorder”. In: *Translational psychiatry* 6.11 (2016), e959–e959.

[134] Santo Fortunato and Darko Hric. “Community detection in networks: A user guide”. In: *Physics reports* 659 (2016), pp. 1–44.

[135] René AW Frank et al. “NMDA receptors are selectively partitioned into complexes and supercomplexes during synapse maturation”. In: *Nature communications* 7.1 (2016), pp. 1–13.

[136] Linton C Freeman. “A set of measures of centrality based on betweenness”. In: *Sociometry* (1977), pp. 35–41.

[137] Linton C Freeman. “Centrality in social networks conceptual clarification”. In: *Social networks* 1.3 (1978), pp. 215–239.

[138] Charles Randy Gallistel and Louis D Matzel. “The neuroscience of learning: beyond the Hebbian synapse”. In: *Annual review of psychology* 64 (2013), pp. 169–200.

[139] Francis Galton. *Hereditary genius*. Macmillan, 1869.

[140] Eric R Gamazon et al. “Using an atlas of gene regulation across 44 human tissues to inform complex disease-and trait-associated variation”. In: *Nature genetics* 50.7 (2018), pp. 956–967.

[141] Nophar Geifman, Alon Monsonego, and Eitan Rubin. “The Neural/Immune Gene Ontology: clipping the Gene Ontology for neurological and immunological systems”. In: *BMC bioinformatics* 11.1 (2010), p. 458.

[142] Alan Genz. “Numerical computation of multivariate normal probabilities”. In: *Journal of computational and graphical statistics* 1.2 (1992), pp. 141–149.

[143] Mahdieh Ghasemi et al. “Centrality measures in biological networks”. In: *Current Bioinformatics* 9.4 (2014), pp. 426–441.

[144] Susan Dina Ghiassian, Jörg Menche, and Albert-László Barabási. “A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome”. In: *PLoS computational biology* 11.4 (2015).

[145] Colin S Gillespie et al. “Fitting Heavy Tailed Distributions: The poweRlaw Package”. In: *Journal of Statistical Software* 64.i02 (2015).

[146] Michelle Girvan and Mark EJ Newman. "Community structure in social and biological networks". In: *Proceedings of the national academy of sciences* 99.12 (2002), pp. 7821–7826.

[147] Enrico Glaab et al. "TopoGSA: network topological gene set analysis". In: *Bioinformatics* 26.9 (2010), pp. 1271–1272.

[148] Enrico Glaab et al. "EnrichNet: network-based gene set enrichment analysis". In: *Bioinformatics* 28.18 (2012), pp. i451–i457.

[149] Oleg O Glebov et al. "Nanoscale structural plasticity of the active zone matrix modulates presynaptic function". In: *Cell reports* 18.11 (2017), pp. 2715–2728.

[150] David F Gleich. "PageRank beyond the Web". In: *Siam Review* 57.3 (2015), pp. 321–363.

[151] Kwang-Il Goh et al. "The human disease network". In: *Proceedings of the National Academy of Sciences* 104.21 (2007), pp. 8685–8690.

[152] Alan J Gow et al. "Stability and change in intelligence from age 11 to ages 70, 79, and 87: the Lothian Birth Cohorts of 1921 and 1936." In: *Psychology and aging* 26.1 (2011), p. 232.

[153] Laura C Graham et al. "Regional molecular mapping of primate synapses during normal healthy aging". In: *Cell reports* 27.4 (2019), pp. 1018–1026.

[154] Seth G Grant et al. "Impaired long-term potentiation, spatial learning, and hippocampal development in fyn mutant mice". In: *Science* 258.5090 (1992), pp. 1903–1910.

[155] Seth GN Grant. "Synaptopathies: diseases of the synaptome". In: *Current opinion in neurobiology* 22.3 (2012), pp. 522–529.

[156] Seth GN Grant. "The molecular evolution of the vertebrate behavioural repertoire". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 371.1685 (2016), p. 20150051.

[157] Seth GN Grant. "Synapse molecular complexity and the plasticity behaviour problem". In: *Brain and Neuroscience Advances* 2 (2018), pp. 1–7.

[158] Seth GN Grant. "The synaptomic theory of behavior and brain disease". In: *Cold Spring Harbor symposia on quantitative biology*. Vol. 83. Cold Spring Harbor Laboratory Press. 2018, pp. 45–56.

[159] Seth GN Grant. “Synapse diversity and synaptome architecture in human genetic disorders”. In: *Human molecular genetics* 28.R2 (2019), R219–R225.

[160] Edward G Gray. “Electron microscopy of synaptic contacts on dendrite spines of the cerebral cortex”. In: *Nature* 183.4675 (1959), pp. 1592–1593.

[161] Kristian A Gray et al. “Genenames. org: the HGNC resources in 2013”. In: *Nucleic acids research* 41.D1 (2012), pp. D545–D552.

[162] Sabrice Guerrier et al. “The F-BAR domain of srGAP2 induces membrane protrusions required for neuronal migration and morphogenesis”. In: *Cell* 138.5 (2009), pp. 990–1004.

[163] Roger Guimera, Marta Sales-Pardo, and Luis A Nunes Amaral. “Modularity from fluctuations in random graphs and complex networks”. In: *Physical Review E* 70.2 (2004), p. 025101.

[164] Eckart D Gundelfinger, Carsten Reissner, and Craig C Garner. “Role of bassoon and piccolo in assembly and molecular organization of the active zone”. In: *Frontiers in synaptic neuroscience* 7 (2016), p. 19.

[165] Aric Hagberg, Pieter Swart, and Daniel S Chult. *Exploring network structure, dynamics, and function using NetworkX*. Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[166] Matthew W Hahn and Andrew D Kern. “Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks”. In: *Molecular biology and evolution* 22.4 (2005), pp. 803–806.

[167] Jeremy Hall et al. “Genetic risk for schizophrenia: convergence on synaptic pathways involved in plasticity”. In: *Biological psychiatry* 77.1 (2015), pp. 52–58.

[168] Ada Hamosh et al. “Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders”. In: *Nucleic acids research* 33.suppl_1 (2005), pp. D514–D517.

[169] Kristen M Harris and Richard J Weinberg. “Ultrastructure of synapses in the mammalian brain”. In: *Cold Spring Harbor perspectives in biology* 4.5 (2012), a005587.

[170] Xionglei He and Jianzhi Zhang. “Why do hubs tend to be essential in protein networks?” In: *PLoS Genet* 2.6 (2006), e88.

[171] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Wiley, New York, 1949.

[172] Fritz Heider. "Attitudes and cognitive organization". In: *The Journal of psychology* 21.1 (1946), pp. 107–112.

[173] Katharina Friedlinde Heil. "Systems biological approach to Parkinson's disease". PhD thesis. 2018.

[174] Jeremy M Henley and Kevin A Wilkinson. "AMPA receptor trafficking and the mechanisms underlying synaptic plasticity and cognitive aging". In: *Dialogues in clinical neuroscience* 15.1 (2013), p. 11.

[175] Henning Hermjakob et al. "IntAct: an open source molecular interaction database". In: *Nucleic acids research* 32.suppl_1 (2004), pp. D452–D455.

[176] Henning Hermjakob et al. "The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data". In: *Nature biotechnology* 22.2 (2004), pp. 177–183.

[177] W David Hill et al. "Functional gene group analysis indicates no role for heterotrimeric G proteins in cognitive ability". In: *PloS one* 9.3 (2014), e91690.

[178] W David Hill et al. "Human cognitive ability is influenced by genetic variation in components of postsynaptic signalling complexes assembled by NMDA receptors and MAGUK proteins". In: *Translational psychiatry* 4.1 (2014), e341–e341.

[179] W David Hill et al. "Age-dependent pleiotropy between general cognitive function and major psychiatric disorders". In: *Biological psychiatry* 80.4 (2016), pp. 266–273.

[180] WD Hill et al. "A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence". In: *Molecular psychiatry* 24.2 (2019), pp. 169–181.

[181] William David Hill et al. "Molecular genetic aetiology of general cognitive function is enriched in evolutionarily conserved regions". In: *Translational psychiatry* 6.12 (2016), e980–e980.

[182] Michael Himsolt. *GML: A portable graph file format*. Tech. rep. Technical report, Universitat Passau, 1997.

[183] Lucia A Hindorff et al. “Potential etiologic and functional implications of genome-wide association loci for human diseases and traits”. In: *Proceedings of the National Academy of Sciences* 106.23 (2009), pp. 9362–9367.

[184] Geoffrey E Hinton. “Learning multiple layers of representation”. In: *Trends in cognitive sciences* 11.10 (2007), pp. 428–434.

[185] Peter Holmans et al. “Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder”. In: *The American Journal of Human Genetics* 85.1 (2009), pp. 13–24.

[186] Petter Holme. “Rare and everywhere: Perspectives on scale-free networks”. In: *Nature communications* 10.1 (2019), pp. 1–3.

[187] Michael Hsing, Kendall Grant Byler, and Artem Cherkasov. “The use of Gene Ontology terms for predicting highly-connected’hub’nodes in protein-protein interaction networks”. In: *BMC systems biology* 2.1 (2008), p. 80.

[188] Wei Hu et al. “The glutamate hypothesis of schizophrenia: evidence from human brain tissue studies”. In: *Annals of the New York Academy of Sciences* 1338.1 (2015), p. 38.

[189] Hongzhan Huang et al. “A comprehensive protein-centric ID mapping service for molecular data integration”. In: *Bioinformatics* 27.8 (2011), pp. 1190–1191.

[190] Mark D Humphries and Kevin Gurney. “Network ‘small-world-ness’: a quantitative method for determining canonical network equivalence”. In: *PloS one* 3.4 (2008), e0002051.

[191] Holger Husi et al. “Proteomic analysis of NMDA receptor–adhesion protein signaling complexes”. In: *Nature neuroscience* 3.7 (2000), pp. 661–669.

[192] Trey Ideker and Roded Sharan. “Protein networks in disease”. In: *Genome research* 18.4 (2008), pp. 644–652.

[193] Ross Ihaka and Robert Gentleman. “R: a language for data analysis and graphics”. In: *Journal of computational and graphical statistics* 5.3 (1996), pp. 299–314.

[194] Ruth Isserlin, Rashad A El-Badrawi, and Gary D Bader. “The biomolecular interaction network database in PSI-MI 2.5”. In: *Database* 2011 (2011).

[195] Kei Ito et al. “A systematic nomenclature for the insect brain”. In: *Neuron* 81.4 (2014), pp. 755–765.

[196] Mahdi Jalili et al. “CentiServer: a comprehensive resource, web-based application and R package for centrality analysis”. In: *PloS one* 10.11 (2015), e0143111.

[197] Hawoong Jeong et al. “Lethality and centrality in protein networks”. In: *Nature* 411.6833 (2001), pp. 41–42.

[198] Peilin Jia et al. “Common variants conferring risk of schizophrenia: a pathway analysis of GWAS data”. In: *Schizophrenia research* 122.1-3 (2010), pp. 38–42.

[199] Peilin Jia et al. “Pathway-based analysis of GWAS datasets: effective but caution required”. In: *International Journal of Neuropsychopharmacology* 14.4 (2011), pp. 567–572.

[200] Wendy Johnson et al. “Just one g: Consistent results from three test batteries”. In: *Intelligence* 32.1 (2004), pp. 95–107.

[201] Samuel E Jones et al. “Genome-wide association analyses of chronotype in 697,828 individuals provides insights into circadian rhythms”. In: *Nature communications* 10.1 (2019), pp. 1–11.

[202] Maliackal Poulo Joy et al. “High-betweenness proteins in the yeast protein interaction network”. In: *BioMed Research International* 2005.2 (2005), pp. 96–103.

[203] Ran Kafri, Michael Springer, and Yitzhak Pilpel. “Genetic redundancy: new tricks for old genes”. In: *Cell* 136.3 (2009), pp. 389–392.

[204] Eric R Kandel, Yadin Dudai, and Mark R Mayford. “The molecular and systems biology of memory”. In: *Cell* 157.1 (2014), pp. 163–186.

[205] Minoru Kanehisa and Susumu Goto. “KEGG: kyoto encyclopedia of genes and genomes”. In: *Nucleic acids research* 28.1 (2000), pp. 27–30.

[206] Konrad J Karczewski et al. “The mutational constraint spectrum quantified from variation in 141,456 humans”. In: *Nature* 581.7809 (2020), pp. 434–443.

[207] Orsolya Kardos, András London, and Tamás Vinkó. “Stability of network centrality measures: a numerical study”. In: *Social Network Analysis and Mining* 10.1 (2020), pp. 1–17.

[208] Ege T Kavalali and Lisa M Monteggia. “How does ketamine elicit a rapid antidepressant response?” In: *Current opinion in pharmacology* 20 (2015), pp. 35–39.

[209] Robert J Keenan et al. “The signal recognition particle”. In: *Annual review of biochemistry* 70.1 (2001), pp. 755–775.

[210] Kenneth S Kendler et al. “Prediction of onset of substance-induced psychotic disorder and its progression to schizophrenia in a Swedish national sample”. In: *American Journal of Psychiatry* 176.9 (2019), pp. 711–719.

[211] Phillip Kent. “Fluid intelligence: A brief history”. In: *Applied Neuropsychology: Child* 6.3 (2017), pp. 193–203.

[212] Samuel Kerrien et al. “Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions”. In: *BMC biology* 5.1 (2007), p. 44.

[213] TS Keshava Prasad et al. “Human protein reference database—2009 update”. In: *Nucleic acids research* 37.suppl_1 (2009), pp. D767–D772.

[214] Eric B Keverne. “An evaluation of what the mouse knockout experiments are telling us about mammalian behaviour”. In: *Bioessays* 19.12 (1997), pp. 1091–1098.

[215] Purvesh Khatri and Sorin Drăghici. “Ontological analysis of gene expression data: current tools, limitations, and open problems”. In: *Bioinformatics* 21.18 (2005), pp. 3587–3595.

[216] Purvesh Khatri, Marina Sirota, and Atul J Butte. “Ten years of pathway analysis: current approaches and outstanding challenges”. In: *PLoS Comput Biol* 8.2 (2012), e1002375.

[217] Jinhwan Kim, Sora Yoon, and Dougu Nam. “netGO: R-Shiny package for network-integrated pathway enrichment analysis”. In: *Bioinformatics* 36.10 (2020), pp. 3283–3285.

[218] Yang-Min Kim, Jean-Baptiste Poline, and Guillaume Dumas. “Experimenting with reproducibility: a case study of robustness in bioinformatics”. In: *GigaScience* 7.7 (2018), giy077.

[219] Sinead King et al. “Inflammation, Glutamate, and Cognition in Bipolar Disorder Type II: A Proof of Concept Study”. In: *Frontiers in psychiatry* 10 (2019), p. 66.

[220] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. “Optimization by simulated annealing”. In: *science* 220.4598 (1983), pp. 671–680.

[221] Hiroaki Kitano. “Systems biology: a brief overview”. In: *science* 295.5560 (2002), pp. 1662–1664.

[222] Maksim Kitsak et al. “Tissue specificity of human disease module”. In: *Scientific reports* 6.1 (2016), pp. 1–12.

[223] Adrienne Kitts and Stephen Sherry. “The single nucleotide polymorphism database (dbSNP) of nucleotide sequence variation”. In: *The NCBI Handbook. McEntyre J, Ostell J, eds. Bethesda, MD: US National Center for Biotechnology Information* (2002).

[224] Jon M Kleinberg. “Authoritative sources in a hyperlinked environment”. In: *Journal of the ACM (JACM)* 46.5 (1999), pp. 604–632.

[225] Frank Koopmans et al. “Comparative hippocampal synaptic proteomes of rodents and primates: differences in neuroplasticity-related proteins”. In: *Frontiers in molecular neuroscience* 11 (2018), p. 364.

[226] Frank Koopmans et al. “SynGO: an evidence-based, expert-curated knowledge base for the synapse”. In: *Neuron* 103.2 (2019), pp. 217–234.

[227] Esa R Korpi et al. “Mechanisms of action and persistent neuroplasticity by drugs of abuse”. In: *Pharmacological reviews* 67.4 (2015), pp. 872–1004.

[228] David Lamparter et al. “Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics”. In: *PLoS computational biology* 12.1 (2016), e1004714.

[229] Andrea Lancichinetti and Santo Fortunato. “Community detection algorithms: a comparative analysis”. In: *Physical review E* 80.5 (2009), p. 056117.

[230] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. “Benchmark graphs for testing community detection algorithms”. In: *Physical review E* 78.4 (2008), p. 046110.

[231] Melissa J Landrum et al. “ClinVar: public archive of interpretations of clinically relevant variants”. In: *Nucleic acids research* 44.D1 (2016), pp. D862–D868.

[232] Peter Langfelder and Steve Horvath. “WGCNA: an R package for weighted correlation network analysis”. In: *BMC bioinformatics* 9.1 (2008), p. 559.

[233] Melanie Laßek, Jens Weingarten, and Walter Volknandt. “The proteome of the murine presynaptic active zone”. In: *Proteomes* 2.2 (2014), pp. 243–257.

[234] Melanie Laßek, Jens Weingarten, and Walter Volknandt. “The synaptic proteome”. In: *Cell and tissue research* 359.1 (2015), pp. 255–265.

[235] Vito Latora and Massimo Marchiori. “Efficient behavior of small-world networks”. In: *Physical review letters* 87.19 (2001), p. 198701.

[236] Simon B Laughlin and Terrence J Sejnowski. “Communication in neuronal networks”. In: *Science* 301.5641 (2003), pp. 1870–1874.

[237] James J Lee et al. “Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment”. In: *Nature Genetics* 50.8 (2018), p. 1112.

[238] Younghee Lee et al. “Network models of genome-wide association studies uncover the topological centrality of protein interactions in complex diseases”. In: *Journal of the American Medical Informatics Association* 20.4 (2013), pp. 619–629.

[239] Christiaan A de Leeuw et al. “MAGMA: generalized gene-set analysis of GWAS data”. In: *PLoS computational biology* 11.4 (2015).

[240] Benjamin Lehne and Thomas Schlitt. “Protein-protein interaction databases: keeping up with growing interactomes”. In: *Human genomics* 3.3 (2009), p. 291.

[241] Monkol Lek et al. “Analysis of protein-coding genetic variation in 60,706 humans”. In: *Nature* 536.7616 (2016), pp. 285–291.

[242] Todd Lencz et al. “Molecular genetic evidence for overlap between general cognitive ability and risk for schizophrenia: a report from the Cognitive Genomics consorTium (COGENT)”. In: *Molecular psychiatry* 19.2 (2014), pp. 168–174.

[243] Jure Leskovec, Kevin J Lang, and Michael Mahoney. “Empirical comparison of algorithms for network community detection”. In: *Proceedings of the 19th international conference on World wide web*. 2010, pp. 631–640.

[244] Jure Leskovec et al. “Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters”. In: *Internet Mathematics* 6.1 (2009), pp. 29–123.

[245] Miao-Xin Li et al. “A knowledge-based weighting framework to boost the power of genome-wide association studies”. In: *PloS one* 5.12 (2010), e14480.

[246] Miao-Xin Li et al. “GATES: a rapid and powerful gene-based association test using extended Simes procedure”. In: *The American Journal of Human Genetics* 88.3 (2011), pp. 283–293.

[247] Yuxing Liao et al. “WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs”. In: *Nucleic acids research* 47.W1 (2019), W199–W205.

[248] Arthur Liberzon et al. “The molecular signatures database hallmark gene set collection”. In: *Cell systems* 1.6 (2015), pp. 417–425.

[249] Shin Lin et al. “Comparison of the transcriptional landscapes between human and mouse tissues”. In: *Proceedings of the National Academy of Sciences* 111.48 (2014), pp. 17224–17229.

[250] Esther S Lips et al. “Functional gene group analysis identifies synaptic gene groups as risk factor for schizophrenia”. In: *Molecular psychiatry* 17.10 (2012), pp. 996–1006.

[251] John Lisman. “The CaM kinase II hypothesis for the storage of synaptic memory”. In: *Trends in neurosciences* 17.10 (1994), pp. 406–412.

[252] John Lisman. “Glutamatergic synapses are structurally and biochemically complex because of multiple plasticity processes: long-term potentiation, long-term depression, short-term potentiation and scaling”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1715 (2017), p. 20160260.

[253] John Lisman, Ryohei Yasuda, and Sridhar Raghavachari. “Mechanisms of CaMKII action in long-term potentiation”. In: *Nature reviews neuroscience* 13.3 (2012), pp. 169–182.

[254] Jimmy Z Liu et al. “A versatile gene-based test for genome-wide association studies”. In: *The American Journal of Human Genetics* 87.1 (2010), pp. 139–145.

[255] Mengzhen Liu et al. “Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use”. In: *Nature genetics* 51.2 (2019), pp. 237–244.

[256] Patrick M Loerch et al. “Evolution of the aging brain transcriptome and synaptic regulation”. In: *PloS one* 3.10 (2008), e3329.

[257] Hugh Louch. “Personal network integration: transitivity and homophily in strong-tie relations”. In: *Social networks* 22.1 (2000), pp. 45–64.

[258] Katja Luck et al. “A reference map of the human binary protein interactome”. In: *Nature* 580.7803 (2020), pp. 402–408.

[259] Alicia Lundby et al. “Annotation of loci from genome-wide association studies using tissue-specific quantitative interaction proteomics”. In: *Nature methods* 11.8 (2014), p. 868.

[260] Hao Luo et al. “DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements”. In: *Nucleic acids research* 42.D1 (2014), pp. D574–D580.

[261] Hao Luo et al. “DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools”. In: *Nucleic Acids Research* 49.D1 (2021), pp. D677–D686.

[262] Jacqueline MacArthur et al. “The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)”. In: *Nucleic acids research* 45.D1 (2017), pp. D896–D901.

[263] Henryk Maciejewski. “Gene set analysis methods: statistical models and methodological differences”. In: *Briefings in bioinformatics* 15.4 (2014), pp. 504–518.

[264] Donna Maglott et al. “Entrez Gene: gene-centered information at NCBI”. In: *Nucleic acids research* 33.suppl_1 (2005), pp. D54–D58.

[265] Eirini Marouli et al. “Rare and low-frequency coding variants alter human adult height”. In: *Nature* 542.7640 (2017), pp. 186–190.

[266] Colin Mclean et al. “Improved functional enrichment analysis of biological networks using scalable modularity based clustering”. In: *J. Proteomics Bioinform* 9.10 (2016).

[267] Rupert McShane et al. “Memantine for dementia”. In: *Cochrane database of systematic reviews* 3 (2019).

[268] Rosa J Meijer and Jelle J Goeman. “Multiple testing of gene sets from gene ontology: possibilities and pitfalls”. In: *Briefings in bioinformatics* 17.5 (2016), pp. 808–818.

[269] Marta Melé et al. “The human transcriptome across tissues and individuals”. In: *Science* 348.6235 (2015), pp. 660–665.

[270] Fanrong Meng et al. “Coupled node similarity learning for community detection in attributed networks”. In: *Entropy* 20.6 (2018), p. 471.

[271] Huaiyu Mi et al. “Large-scale gene function analysis with the PANTHER classification system”. In: *Nature protocols* 8.8 (2013), p. 1551.

[272] Huaiyu Mi et al. “Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v. 14.0)”. In: *Nature protocols* 14.3 (2019), pp. 703–721.

[273] Enkelejda Miho, Victor Greiff, Sai T Reddy, et al. “Large-scale network analysis reveals the sequence space architecture of antibody repertoires”. In: *Nature communications* 10.1 (2019), pp. 1–11.

[274] Tarjei Mikkelsen et al. “Initial sequence of the chimpanzee genome and comparison with the human genome”. In: *Nature* 437.7055 (2005), pp. 69–87.

[275] Tijana Milenković et al. “Dominating biological networks”. In: *PloS one* 6.8 (2011).

[276] Stanley Milgram. “The small world problem”. In: *Psychology today* 2.1 (1967), pp. 60–67.

[277] Graeme Milligan and Evi Kostenis. “Heterotrimeric G-proteins: a short history”. In: *British journal of pharmacology* 147.S1 (2006), S46–S55.

[278] Aniket Mishra and Stuart Macgregor. “VEGAS2: software for more flexible gene-based testing”. In: *Twin Research and Human Genetics* 18.1 (2015), pp. 86–91.

[279] Markus Missler, Thomas C Südhof, and Thomas Biederer. “Synaptic cell adhesion”. In: *Cold Spring Harbor perspectives in biology* 4.4 (2012), a005694.

[280] Koyel Mitra et al. “Integrative approaches for finding modular structure in biological networks”. In: *Nature Reviews Genetics* 14.10 (2013), pp. 719–732.

[281] Cristina Mitrea et al. “Methods and approaches in the topology-based analysis of biological pathways”. In: *Frontiers in physiology* 4 (2013), p. 278.

[282] Michael A Mooney and Beth Wilmot. “Gene set analysis: A step-by-step guide”. In: *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 168.7 (2015), pp. 517–527.

[283] Vamsi K Mootha et al. “PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes”. In: *Nature genetics* 34.3 (2003), pp. 267–273.

[284] Gerard Muntané et al. “Analysis of synaptic gene expression in the neocortex of primates reveals evolutionary changes in glutamatergic neurotransmission”. In: *Cerebral Cortex* 25.6 (2015), pp. 1596–1607.

[285] Christopher JL Murray et al. “Global, regional, and national disability-adjusted life years (DALYs) for 306 diseases and injuries and healthy life expectancy (HALE) for 188 countries, 1990–2013: quantifying the epidemiological transition”. In: *The Lancet* 386.10009 (2015), pp. 2145–2191.

[286] James W Murrough, Chadi G Abdallah, and Sanjay J Mathew. “Targeting glutamate signalling in depression: progress and prospects”. In: *Nature Reviews Drug Discovery* 16.7 (2017), p. 472.

[287] Shigetada Nakanishi. “Molecular diversity of glutamate receptors and implications for brain function”. In: *Science* 258.5082 (1992), pp. 597–603.

[288] Zachary P Neal. “How small is it? Comparing indices of small worldliness”. In: *Network Science* 5.1 (2017), pp. 30–44.

[289] Mark Newman. *Networks*. Oxford university press, 2018.

[290] Mark EJ Newman. “The structure of scientific collaboration networks”. In: *Proceedings of the national academy of sciences* 98.2 (2001), pp. 404–409.

[291] Mark EJ Newman. “Assortative mixing in networks”. In: *Physical review letters* 89.20 (2002), p. 208701.

[292] Mark EJ Newman. “Finding community structure in networks using the eigenvectors of matrices”. In: *Physical review E* 74.3 (2006), p. 036104.

[293] Mark EJ Newman. “Modularity and community structure in networks”. In: *Proceedings of the national academy of sciences* 103.23 (2006), pp. 8577–8582.

[294] Mark EJ Newman. “Communities, modules and large-scale structure in networks”. In: *Nature physics* 8.1 (2012), pp. 25–31.

[295] Mark EJ Newman. “Spectral methods for community detection and graph partitioning”. In: *Physical Review E* 88.4 (2013), p. 042822.

[296] Mark EJ Newman. “Equivalence between modularity optimization and maximum likelihood methods for community detection”. In: *Physical Review E* 94.5 (2016), p. 052315.

[297] Mark EJ Newman and Aaron Clauset. “Structure and inference in annotated networks”. In: *Nature communications* 7.1 (2016), pp. 1–11.

[298] Mark EJ Newman and Michelle Girvan. “Finding and evaluating community structure in networks”. In: *Physical review E* 69.2 (2004), p. 026113.

[299] Mark EJ Newman, Duncan J Watts, and Steven H Strogatz. “Random graph models of social networks”. In: *Proceedings of the national academy of sciences* 99.suppl 1 (2002), pp. 2566–2572.

[300] Hung Nguyen et al. “A comprehensive survey of tools and software for active subnetwork identification”. In: *Frontiers in genetics* 10 (2019), p. 155.

[301] Tin Nguyen, Cristina Mitrea, and Sorin Draghici. “Network-Based Approaches for Pathway Level Analysis”. In: *Current protocols in bioinformatics* 61.1 (2018), pp. 8–25.

[302] Mark J Niciu, Benjamin Kelmendi, and Gerard Sanacora. “Overview of glutamatergic neurotransmission in the nervous system”. In: *Pharmacology Biochemistry and Behavior* 100.4 (2012), pp. 656–664.

[303] Roger A Nicoll. “A brief history of long-term potentiation”. In: *Neuron* 93.2 (2017), pp. 281–290.

[304] Colleen M Niswender and P Jeffrey Conn. “Metabotropic glutamate receptors: physiology, pharmacology, and disease”. In: *Annual review of pharmacology and toxicology* 50 (2010), pp. 295–322.

[305] Jess Nithianantharajah et al. “Synaptic scaffold evolution generated components of vertebrate cognitive complexity”. In: *Nature neuroscience* 16.1 (2013), p. 16.

[306] Marco S Nobile et al. “Graphics processing units in bioinformatics, computational biology and systems biology”. In: *Briefings in bioinformatics* 18.5 (2017), pp. 870–885.

[307] Rogier Noldus and Piet Van Mieghem. “Assortativity in complex networks”. In: *Journal of Complex Networks* 3.4 (2015), pp. 507–542.

[308] Kazuya Okamoto, Wei Chen, and Xiang-Yang Li. “Ranking of closeness centrality for large-scale social networks”. In: *International workshop on frontiers in algorithmics*. Springer. 2008, pp. 186–195.

[309] Aysu Okbay et al. “Genome-wide association study identifies 74 loci associated with educational attainment”. In: *Nature* 533.7604 (2016), pp. 539–542.

[310] Stuart Oldham et al. “Consistency and differences between centrality measures across distinct classes of networks”. In: *PloS one* 14.7 (2019).

[311] Stephen Oliver. “Guilt-by-association goes global”. In: *Nature* 403.6770 (2000), pp. 601–602.

[312] Sandra Orchard. “Molecular interaction databases”. In: *Proteomics* 12.10 (2012), pp. 1656–1662.

[313] Sandra Orchard et al. “Protein interaction data curation: the International Molecular Exchange (IMEx) consortium”. In: *Nature methods* 9.4 (2012), pp. 345–350.

[314] Sandra Orchard et al. “The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases”. In: *Nucleic acids research* 42.D1 (2014), pp. D358–D363.

[315] Günce Keziban Orman and Vincent Labatut. “A comparison of community detection algorithms on artificial networks”. In: *International conference on discovery science*. Springer. 2009, pp. 242–256.

[316] Günce Keziban Orman, Vincent Labatut, and Hocine Cherifi. “Towards realistic artificial benchmark for community detection algorithms evaluation”. In: *International Journal of Web Based Communities* 9.3 (2013), pp. 349–370.

[317] Arzucan Özgür et al. “Identifying gene-disease associations using centrality on a literature mined gene-interaction network”. In: *Bioinformatics* 24.13 (2008), pp. i277–i285.

[318] Annalise B Paaby and Matthew V Rockman. “The many faces of pleiotropy”. In: *Trends in genetics* 29.2 (2013), pp. 66–73.

[319] Bente Pakkenberg et al. “Aging and the human neocortex”. In: *Experimental gerontology* 38.1-2 (2003), pp. 95–99.

[320] Gergely Palla et al. “Uncovering the overlapping community structure of complex networks in nature and society”. In: *nature* 435.7043 (2005), pp. 814–818.

[321] A Pandey et al. “Epistasis network centrality analysis yields pathway replication across two GWAS cohorts for bipolar disorder”. In: *Translational psychiatry* 2.8 (2012), e154–e154.

[322] Neelroop N Parikshak, Michael J Gandal, and Daniel H Geschwind. “Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders”. In: *Nature Reviews Genetics* 16.8 (2015), pp. 441–458.

[323] Mikyoung Park. “AMPA receptor trafficking for postsynaptic potentiation”. In: *Frontiers in cellular neuroscience* 12 (2018), p. 361.

[324] Kartik Pattabiraman, Sydney Keaton Muchnik, and Nenad Sestan. “The evolution of the human brain and disease susceptibility”. In: *Current Opinion in Genetics & Development* 65 (2020), pp. 91–97.

[325] Leto Peel, Daniel B Larremore, and Aaron Clauset. “The ground truth about metadata and community detection in networks”. In: *Science advances* 3.5 (2017), e1602548.

[326] Tiago P. Peixoto. “The graph-tool python library”. In: *figshare* (2014). DOI: 10.6084/m9.figshare.1164194. URL: http://figshare.com/articles/graph_tool/1164194 (visited on 09/10/2014).

[327] Junmin Peng et al. “Semiquantitative proteomic analysis of rat forebrain postsynaptic density fractions by mass spectrometry”. In: *Journal of Biological Chemistry* 279.20 (2004), pp. 21003–21011.

[328] Jonathan Perreault, Jean-Pierre Perreault, and Gilles Boire. “Ro-associated Y RNAs in metazoans: evolution and diversification”. In: *Molecular biology and evolution* 24.8 (2007), pp. 1678–1689.

[329] Tune H Pers et al. “Biological interpretation of genome-wide association studies using predicted gene functions”. In: *Nature communications* 6.1 (2015), pp. 1–9.

[330] Ashley Petersen et al. “Assessing methods for assigning SNPs to genes in gene-based tests of association using common variants”. In: *PLoS One* 8.5 (2013), e62161.

[331] Milos M Petrovic et al. “Metabotropic action of postsynaptic kainate receptors triggers hippocampal long-term potentiation”. In: *Nature neuroscience* 20.4 (2017), pp. 529–539.

[332] Joseph K Pickrell et al. “Understanding mechanisms underlying human gene expression variation with RNA sequencing”. In: *Nature* 464.7289 (2010), pp. 768–772.

[333] Jean-Philippe Pin and Bernhard Bettler. “Organization and functions of mGlu and GABA B receptor complexes”. In: *Nature* 540.7631 (2016), pp. 60–68.

[334] Janet Piñero et al. “DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants”. In: *Nucleic acids research* (2016), gkw943.

[335] Janet Piñero et al. “The DisGeNET knowledge platform for disease genomics: 2019 update”. In: *Nucleic acids research* 48.D1 (2020), pp. D845–D855.

[336] Robert Plomin and Ian J Deary. “Genetics and intelligence differences: five special findings”. In: *Molecular psychiatry* 20.1 (2015), pp. 98–108.

[337] Robert Plomin and Sophie von Stumm. “The new genetics of intelligence”. In: *Nature Reviews Genetics* 19.3 (2018), p. 148.

[338] AJ Pocklington, JD Armstrong, and SGN Grant. “Organization of brain complexity—synapse proteome form and function”. In: *Briefings in Functional Genomics* 5.1 (2006), pp. 66–73.

[339] Andrew J Pocklington, Michael O’Donovan, and Michael J Owen. “The synapse in schizophrenia”. In: *European Journal of Neuroscience* 39.7 (2014), pp. 1059–1067.

[340] Andrew J Pocklington et al. “The proteomes of neurotransmitter receptor complexes form modular networks with distributed functionality underlying plasticity and behaviour”. In: *Molecular systems biology* 2.1 (2006).

[341] Pascal Pons and Matthieu Latapy. “Computing communities in large networks using random walks”. In: *International symposium on computer and information sciences*. Springer. 2005, pp. 284–293.

[342] Derek J De Solla Price. “Networks of scientific papers”. In: *Science* (1965), pp. 510–515.

[343] Bruce M Psaty et al. “Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts”. In: *Circulation: Cardiovascular Genetics* 2.1 (2009), pp. 73–80.

[344] Shaun Purcell et al. “PLINK: a tool set for whole-genome association and population-based linkage analyses”. In: *The American journal of human genetics* 81.3 (2007), pp. 559–575.

[345] Shaun M Purcell et al. “A polygenic burden of rare disruptive mutations in schizophrenia”. In: *Nature* 506.7487 (2014), pp. 185–190.

[346] Filippo Radicchi et al. “Defining and identifying communities in networks”. In: *Proceedings of the national academy of sciences* 101.9 (2004), pp. 2658–2663.

[347] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. “Near linear time algorithm to detect community structures in large-scale networks”. In: *Physical review E* 76.3 (2007), p. 036106.

[348] Karthik Raman, Nandita Damaraju, and Govind Krishna Joshi. “The organisational structure of protein networks: revisiting the centrality–lethality hypothesis”. In: *Systems and synthetic biology* 8.1 (2014), pp. 73–81.

[349] Santiago Ramon y Cajal. “Histologie du système nerveux de l’homme et des vertébrés”. In: *Maloine, Paris* 2 (1911), pp. 153–173.

[350] Erin M Ramos et al. “Phenotype–Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources”. In: *European Journal of Human Genetics* 22.1 (2014), pp. 144–147.

[351] Oliver Ratmann, Carsten Wiuf, and John W Pinney. “From evidence to inference: probing the evolution of protein interaction networks”. In: *HFSP journal* 3.5 (2009), pp. 290–306.

[352] Uku Raudvere et al. “g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)”. In: *Nucleic acids research* 47.W1 (2019), W191–W198.

[353] David E Reich et al. “Linkage disequilibrium in the human genome”. In: *Nature* 411.6834 (2001), pp. 199–204.

[354] Jörg Reichardt and Stefan Bornholdt. “Detecting fuzzy community structures in complex networks with a Potts model”. In: *Physical review letters* 93.21 (2004), p. 218701.

[355] Jörg Reichardt and Stefan Bornholdt. “Statistical mechanics of community detection”. In: *Physical review E* 74.1 (2006), p. 016110.

[356] Jüri Reimand et al. “g: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments”. In: *Nucleic acids research* 35.suppl_2 (2007), W193–W200.

[357] Seung Yon Rhee et al. “Use and misuse of the gene ontology annotations”. In: *Nature Reviews Genetics* 9.7 (2008), pp. 509–515.

[358] Cornelius A Rietveld et al. “GWAS of 126,559 individuals identifies genetic variants associated with educational attainment”. In: *science* 340.6139 (2013), pp. 1467–1471.

[359] Cornelius A Rietveld et al. “Common genetic variants associated with cognitive performance identified using the proxy-phenotype method”. In: *Proceedings of the National Academy of Sciences* 111.38 (2014), pp. 13790–13794.

[360] Stephan Ripke et al. “Biological insights from 108 schizophrenia-associated genetic loci”. In: *Nature* 511.7510 (2014), pp. 421–427.

[361] Stuart J Ritchie and Elliot M Tucker-Drob. “How much does education improve intelligence? A meta-analysis”. In: *Psychological science* 29.8 (2018), pp. 1358–1369.

[362] Martin Rosvall and Carl T Bergstrom. “Maps of random walks on complex networks reveal community structure”. In: *Proceedings of the National Academy of Sciences* 105.4 (2008), pp. 1118–1123.

[363] Dina Ruano et al. “Functional gene group analysis reveals a role of synaptic heterotrimeric G proteins in cognitive ability”. In: *The American Journal of Human Genetics* 86.2 (2010), pp. 113–125.

[364] Dmitri A Rusakov et al. “Shaping the synaptic signal: molecular mobility inside and outside the cleft”. In: *Trends in neurosciences* 34.7 (2011), pp. 359–369.

[365] Lukasz Salwinski et al. “The database of interacting proteins: 2004 update”. In: *Nucleic acids research* 32.suppl_1 (2004), pp. D449–D451.

[366] Keiko Satoh et al. “Identification of activity-regulated proteins in the postsynaptic density fraction”. In: *Genes to Cells* 7.2 (2002), pp. 187–197.

[367] Jeanne E Savage et al. “Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence”. In: *Nature genetics* 50.7 (2018), pp. 912–919.

[368] Martin H Schaefer et al. “HIPPIE: Integrating protein interaction networks with experiment based quality scores”. In: *PloS one* 7.2 (2012), e31826.

[369] Thomas Schank and Dorothea Wagner. “Approximating clustering coefficient and transitivity.” In: *Journal of Graph Algorithms and Applications* 9.2 (2005), pp. 265–275.

[370] Nicky Scheefhals and Harold D MacGillavry. “Functional organization of postsynaptic glutamate receptors”. In: *Molecular and Cellular Neuroscience* 91 (2018), pp. 82–94.

[371] Susanne Schoch and Eckart D Gundelfinger. “Molecular organization of the presynaptic active zone”. In: *Cell and tissue research* 326.2 (2006), pp. 379–391.

[372] Jochen Schwenk et al. “High-resolution proteomics unravel architecture and molecular diversity of native AMPA receptor complexes”. In: *Neuron* 74.4 (2012), pp. 621–633.

[373] Thomas W Sedlak et al. “The glutathione cycle shapes synaptic glutamate activity”. In: *Proceedings of the National Academy of Sciences* 116.7 (2019), pp. 2701–2706.

[374] Stephen B Seidman. “Network structure and minimum degree”. In: *Social networks* 5.3 (1983), pp. 269–287.

[375] Fekrije Selimi et al. “Proteomic studies of a single CNS synapse type: the parallel fiber/purkinje cell synapse”. In: *PLoS Biol* 7.4 (2009), e1000083.

[376] Matteo Serafino et al. “True scale-free networks hidden by finite size effects”. In: *Proceedings of the National Academy of Sciences* 118.2 (2021).

[377] S Ali M Shariati and Bart De Strooper. “Redundancy and divergence in the amyloid precursor protein family”. In: *FEBS letters* 587.13 (2013), pp. 2036–2045.

[378] Jitendra Sharma, Alessandra Angelucci, and Mriganka Sur. “Induction of visual orientation modules in auditory cortex”. In: *Nature* 404.6780 (2000), pp. 841–847.

[379] M Sheerington C S in Foster. *A Text-book of Physiology, part 3*. 7th. Macmillan and Co.Ltd, 1897.

[380] Mikihito Shibata et al. “Hominini-specific regulation of CBLN2 increases prefrontal spinogenesis”. In: *Nature* 598.7881 (2021), pp. 489–494.

[381] Alcino J Silva et al. “Impaired spatial learning in alpha-calcium-calmodulin kinase II mutant mice”. In: *Science* 257.5067 (1992), pp. 206–211.

[382] Barry Smith et al. “The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration”. In: *Nature biotechnology* 25.11 (2007), pp. 1251–1255.

[383] Suzanne Sniekers et al. “Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence”. In: *Nature genetics* 49.7 (2017), p. 1107.

[384] Ithiel de Sola Pool and Manfred Kochen. “Contacts and influence”. In: *Social networks* 1.1 (1978), pp. 5–51.

[385] Nadia Solovieff et al. “Pleiotropy in complex traits: challenges and strategies”. In: *Nature Reviews Genetics* 14.7 (2013), pp. 483–495.

[386] Anatoly Sorokin, Oksana Sorokina, and J Douglas Armstrong. “RKappa: Statistical sampling suite for Kappa models”. In: *International Workshop on Hybrid Systems Biology*. Springer. 2014, pp. 128–142.

[387] Oksana Sorokina et al. “A simulator for spatially extended kappa models”. In: *Bioinformatics* 29.23 (2013), p. 3105.

[388] Fabiano Berardo de Sousa and Liang Zhao. “Evaluating and comparing the igraph community detection algorithms”. In: *2014 Brazilian Conference on Intelligent Systems*. IEEE. 2014, pp. 408–413.

[389] Rafael T de Sousa et al. “Genetic studies on the tripartite glutamate synapse in the pathophysiology and therapeutics of mood disorders”. In: *Neuropsychopharmacology* 42.4 (2017), pp. 787–800.

[390] Elizabeth K Speliotes et al. “Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index”. In: *Nature genetics* 42.11 (2010), pp. 937–948.

[391] Michael G Stewart et al. “Structure and complexity of the synapse and dendritic spine”. In: *The synapse*. Elsevier, 2014, pp. 1–20.

[392] Vincent Strehlow et al. “GRIN2A-related disorders: genotype and functional consequence predict phenotype”. In: *Brain* 142.1 (2019), pp. 80–92.

[393] Aravind Subramanian et al. “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles”. In: *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550.

[394] Thomas C Südhof. “The presynaptic active zone”. In: *Neuron* 75.1 (2012), pp. 11–25.

[395] Cathie Sudlow et al. “UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age”. In: *Plos med* 12.3 (2015), e1001779.

[396] Chad J Swanson et al. “Metabotropic glutamate receptors as novel targets for anxiety and stress disorders”. In: *Nature reviews Drug discovery* 4.2 (2005), pp. 131–144.

[397] Damian Szklarczyk et al. “STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets”. In: *Nucleic acids research* 47.D1 (2019), pp. D607–D613.

[398] Erdogan Taskesen and Marcel JT Reinders. “2D representation of transcriptomes by t-SNE exposes relatedness between human tissues”. In: *PloS one* 11.2 (2016), e0149853.

[399] Saeed Tavazoie et al. “Systematic determination of genetic network architecture”. In: *Nature genetics* 22.3 (1999), pp. 281–285.

[400] Scott M Thompson et al. “An excitatory synapse hypothesis of depression”. In: *Trends in neurosciences* 38.5 (2015), pp. 279–294.

[401] Aurelie Tomczak et al. “Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations”. In: *Scientific reports* 8.1 (2018), pp. 1–10.

[402] Roberto Toro et al. “Key role for gene dosage and synaptic homeostasis in autism spectrum disorders”. In: *Trends in genetics* 26.8 (2010), pp. 363–372.

[403] Stephen F Traynelis et al. “Glutamate receptor ion channels: structure, regulation, and function”. In: *Pharmacological reviews* 62.3 (2010), pp. 405–496.

[404] JC Trinidad et al. “Phosphorylation state of postsynaptic density proteins”. In: *Journal of neurochemistry* 92.6 (2005), pp. 1306–1316.

[405] Jonathan C Trinidad et al. “Quantitative analysis of synaptic phosphorylation and protein expression”. In: *Molecular & Cellular Proteomics* 7.4 (2008), pp. 684–696.

[406] Kimberly A Tryka et al. “NCBI’s Database of Genotypes and Phenotypes: dbGaP”. In: *Nucleic acids research* 42.D1 (2014), pp. D975–D979.

[407] Patrick Turley et al. “Multi-trait analysis of genome-wide association summary statistics using MTAG”. In: *Nature genetics* 50.2 (2018), pp. 229–237.

[408] “UniProt: the universal protein knowledgebase”. In: *Nucleic acids research* 45.D1 (2017), pp. D158–D169.

[409] Thomas W Valente et al. “How correlated are network centrality measures?” In: *Connections (Toronto, Ont.)* 28.1 (2008), p. 16.

[410] S Van der Sluis et al. “Power in GWAS: lifting the curse of the clinical cut-off”. In: *Molecular psychiatry* 18.1 (2013), pp. 2–3.

[411] Stijn Marinus Van Dongen. “Graph clustering by flow simulation”. PhD thesis. 2000.

[412] Sunita Venkateswaran et al. “Whole-exome sequencing in an individual with severe global developmental delay and intractable epilepsy identifies a novel, de novo GRIN2A mutation”. In: *Epilepsia* 55.7 (2014), e75–e79.

[413] Petra E Vértes and Edward T Bullmore. “Annual research review: growth connectomics–the organization and reorganization of brain networks during normal and abnormal development”. In: *Journal of Child Psychology and Psychiatry* 56.3 (2015), pp. 299–320.

[414] Nguyen Xuan Vinh, Julien Epps, and James Bailey. “Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance”. In: *The Journal of Machine Learning Research* 11 (2010), pp. 2837–2854.

[415] Peter M Visscher and Jian Yang. “A plethora of pleiotropy across complex traits”. In: *Nature genetics* 48.7 (2016), pp. 707–708.

[416] Peter M Visscher et al. “10 years of GWAS discovery: biology, function, and translation”. In: *The American Journal of Human Genetics* 101.1 (2017), pp. 5–22.

[417] Bert Vogelstein, David Lane, and Arnold J Levine. “Surfing the p53 network”. In: *Nature* 408.6810 (2000), pp. 307–310.

[418] Quang H Vuong. “Likelihood ratio tests for model selection and non-nested hypotheses”. In: *Econometrica: Journal of the Econometric Society* (1989), pp. 307–333.

[419] Shinichiro Wachi, Ken Yoneda, and Reen Wu. “Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues”. In: *Bioinformatics* 21.23 (2005), pp. 4205–4208.

[420] Randall S Walikonis et al. “Identification of proteins in the postsynaptic density fraction by mass spectrometry”. In: *Journal of Neuroscience* 20.11 (2000), pp. 4069–4080.

[421] Joseph Walpole, Jason A Papin, and Shayn M Peirce. “Multiscale computational models of complex biological systems”. In: *Annual review of biomedical engineering* 15 (2013), pp. 137–154.

[422] Kai Wang, Mingyao Li, and Maja Bucan. “Pathway-based approaches for analysis of genomewide association studies”. In: *The American Journal of Human Genetics* 81.6 (2007), pp. 1278–1283.

[423] Lily Wang et al. “Gene set analysis of genome-wide association studies: methodological issues and perspectives”. In: *Genomics* 98.1 (2011), pp. 1–8.

[424] Tao Wang et al. “Disease Module Identification Based on Representation Learning of Complex Networks Integrated From GWAS, eQTL Summaries, and Human Interactome”. In: *Frontiers in Bioengineering and Biotechnology* 8 (2020), p. 418.

[425] Tim Wang et al. “Identification and characterization of essential genes in the human genome”. In: *Science* 350.6264 (2015), pp. 1096–1101.

[426] Yu Wang et al. “Comparison of different generalizations of clustering coefficient and local efficiency for weighted undirected graphs”. In: *Neural computation* 29.2 (2017), pp. 313–331.

[427] Zhiqiang Wang et al. “An approach to cold-start link prediction: Establishing connections between non-topological and topological information”. In: *IEEE Transactions on Knowledge and Data Engineering* 28.11 (2016), pp. 2857–2870.

[428] Kyoko Watanabe et al. “Functional mapping and annotation of genetic associations with FUMA”. In: *Nature communications* 8.1 (2017), pp. 1–11.

[429] Duncan J Watts and Steven H Strogatz. “Collective dynamics of ‘small-world’networks”. In: *nature* 393.6684 (1998), p. 440.

[430] Gregory P Way et al. “Implicating candidate genes at GWAS signals by leveraging topologically associating domains”. In: *European Journal of Human Genetics* 25.11 (2017), pp. 1286–1289.

[431] Danielle Welter et al. “The NHGRI GWAS Catalog, a curated resource of SNP-trait associations”. In: *Nucleic acids research* 42.D1 (2014), pp. D1001–D1006.

[432] Hal Whitehead et al. “The reach of gene–culture coevolution in animals”. In: *Nature Communications* 10.1 (2019), pp. 1–10.

[433] Andrew R Wood et al. “Defining the role of common variation in the genomic and biological architecture of adult human height”. In: *Nature genetics* 46.11 (2014), p. 1173.

[434] Katherine A Wood et al. “The role of the U5 snRNP in genetic disorders and cancer”. In: *Frontiers in Genetics* 12 (2021).

[435] Christina Wraw et al. “Intelligence in youth and health at age 50”. In: *Intelligence* 53 (2015), pp. 23–32.

[436] Stefan Wuchty and Eivind Almaas. “Peeling the yeast protein network”. In: *Proteomics* 5.2 (2005), pp. 444–449.

[437] Ioannis Xenarios et al. “DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions”. In: *Nucleic acids research* 30.1 (2002), pp. 303–305.

[438] Lei Xiao et al. “Adaptive neural information processing with dynamical electrical synapses”. In: *Frontiers in Computational Neuroscience* 7 (2013), p. 36.

[439] Feng Xie and David Levinson. “Measuring the structure of road networks”. In: *Geographical analysis* 39.3 (2007), pp. 336–356.

[440] Gang Xu et al. “Module detection in complex networks using integer optimisation”. In: *Algorithms for Molecular Biology* 5.1 (2010), p. 36.

[441] Jianzhen Xu and Yongjin Li. “Discovering disease-genes by topological features in human protein–protein interaction network”. In: *Bioinformatics* 22.22 (2006), pp. 2800–2805.

[442] Hsin-Chou Yang et al. “Genome-wide gene-based association study”. In: *BMC proceedings*. Vol. 3. 7. BioMed Central. 2009, pp. 1–5.

[443] Jaewon Yang and Jure Leskovec. “Overlapping community detection at scale: a nonnegative matrix factorization approach”. In: *Proceedings of the sixth ACM international conference on Web search and data mining*. 2013, pp. 587–596.

[444] Jaewon Yang and Jure Leskovec. “Defining and evaluating network communities based on ground-truth”. In: *Knowledge and Information Systems* 42.1 (2015), pp. 181–213.

[445] Jian Yang et al. “GCTA: a tool for genome-wide complex trait analysis”. In: *The American Journal of Human Genetics* 88.1 (2011), pp. 76–82.

[446] Zhao Yang, René Algesheimer, and Claudio J Tessone. “A comparative analysis of community detection algorithms on artificial networks”. In: *Scientific reports* 6.1 (2016), pp. 1–18.

[447] Soon-Hyung Yook, Zoltán N Oltvai, and Albert-László Barabási. “Functional and topological characterization of protein interaction networks”. In: *Proteomics* 4.4 (2004), pp. 928–942.

[448] Yoshiyuki Yoshimura et al. “Molecular constituents of the postsynaptic density fraction revealed by proteomic analysis using multidimensional liquid chromatography-tandem mass spectrometry”. In: *Journal of neurochemistry* 88.3 (2004), pp. 759–768.

[449] Haiyuan Yu et al. “The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics”. In: *PLoS Comput Biol* 3.4 (2007), e59.

[450] Haiyuan Yu et al. “High-quality binary protein interaction map of the yeast interactome network”. In: *Science* 322.5898 (2008), pp. 104–110.

[451] Hongjie Yuan et al. “Ionotropic GABA and glutamate receptor mutations and human neurologic diseases”. In: *Molecular pharmacology* 88.1 (2015), pp. 203–217.

[452] Wayne W Zachary. “An information flow model for conflict and fission in small groups”. In: *Journal of anthropological research* 33.4 (1977), pp. 452–473.

[453] Huiying Zhao and Dale R Nyholt. “Gene-based analyses reveal novel genetic overlap and allelic heterogeneity across five major psychiatric disorders”. In: *Human genetics* 136.2 (2017), pp. 263–274.

[454] Shanrong Zhao, Zhan Ye, and Robert Stanton. “Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols”. In: *RNA* (2020), rna–074922.

[455] Jie Zheng et al. “LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis”. In: *Bioinformatics* 33.2 (2017), pp. 272–279.

[456] Jinwei Zhu, Yuan Shang, and Mingjie Zhang. “Mechanistic basis of MAGUK-organized complexes in synaptic development and signalling”. In: *Nature Reviews Neuroscience* 17.4 (2016), p. 209.

[457] Xiaowei Zhu, Mark Gerstein, and Michael Snyder. “Getting connected: analysis and principles of biological networks”. In: *Genes & development* 21.9 (2007), pp. 1010–1024.

[458] Elena Zotenko et al. “Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality”. In: *PLoS Comput Biol* 4.8 (2008), e1000140.

[459] Benoît Zuber et al. “The mammalian central nervous synaptic cleft contains a high density of periodically organized complexes”. In: *Proceedings of the National Academy of Sciences* 102.52 (2005), pp. 19192–19197.

[460] Joanna Zyla et al. “Ranking metrics in gene set enrichment analysis: do they matter?” In: *BMC bioinformatics* 18.1 (2017), p. 256.