Title
Global and Local Regulation of Gene Expression in the Human Brain

Author
Hartl, Christopher

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Global and Local Regulation of Gene Expression in the Human Brain

A dissertation submitted in partial satisfaction of the

requirements for the degree of Doctor of Philosophy

in Bioinformatics

by

Christopher Hartl

2019

ABSTRACT OF THE DISSERTATION

Global and Local Regulation of Gene Expression in the Human Brain

by

Christopher Hartl

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2019

Professor Daniel H. Geschwind, Chair

Neuropsychiatric disorders are behavioral conditions marked by intellectual, social, or emotional deficits that can be linked to diseases of the nervous system. Autism spectrum disorder (ASD), schizophrenia (SCZ), bipolar disorder (BP), major depressive disorder (MDD), and attention deficit and hyperactivity disorder (ADHD) are common, heritable diseases each with a prevalence exceeding 1% of the population, none of which can be characterized by discernable anatomical or neurological pathologies. Genetic association studies have identified mutations in hundreds of genes that contribute to risk for at least one of these disorders, and have shown that a substantial fraction of the genetic liability is shared between many of these neuropsychiatric diseases. It has long been hoped that with enough genetic evidence we will identify the biological pathways, developmental time points, and brain regions that, when disrupted, give rise to neuropsychiatric disorders. However, the cellular and functional complexity of the human

brain, as well as the genetic complexity of neuropsychiatric disease, make it difficult to search for such convergence.

In this thesis, I investigate global and local transcriptional regulation within and across 12 regions of the human brain in order to investigate the regional specificity of neuropsychiatric disorders. I develop novel bioinformatics methods – ranging from data processing to network construction – to identify whether the transcriptional regulation of a set of genes is shared or specific. I hypothesize that local, region-specific transcriptional regulation corresponds directly to cell types and processes that are specific to, or far more prevalent in, a given region; that cross-regional transcriptional regulation corresponds to cell types that show little heterogeneity across brain regions; and that genetic disruption of region-specific transcriptional programs results in regional susceptibility. I use a systems-biology approach to summarize transcriptional regulation into reproducibly co-expressed gene sets ("co-expression modules"), which can be analyzed statistically to identify common functions, pathways, and cell types. I then integrate data from genetic association studies to ascertain gene sets conferring outsized risk for neuropsychiatric disorders, thereby implicating the corresponding pathways for further investigation in disease etiology. Finally, I use the network structure itself to investigate the genetic architecture of ASD and SCZ in terms of omnigenics and network polygenics.

Chapter 1 presents the biological background for the studies and summarizes some of the major studies of neuropsychiatric disorders along with their principal methods and conclusions. In chapter 2, utilizing my multi-regional co-expression approach, I identify 12 brain-wide, 114 region-specific, and 50 cross-regional co-expression modules. Nearly 40% of expressed genes fall into brain-wide modules and correspond to major cell classes and conserved biological

processes, while region-specific modules comprise 25% of expressed genes and correspond to region-specific cell types. The detailed study in chapter 3 demonstrates that neuropsychiatric risk concentrates in both brain wide and multi-regional modules, implicating major core cell types in disease etiology but not region-specific susceptibility. Chapter 4 presents a new and more general framework for defining genetic networks. Using this framework, I show that the network pattern of ASD-associated rare loss-of-function mutations, as well as the large number of significant targets for *trans* master regulators in BP and SCZ, support a classical polygenic architecture with thousands of directly causal genes. These results suggest that a nontrivial component of risk for neuropsychiatric disease comes from the global polygenic disruption of neuronal function and neuronal maturation.

The dissertation of Christopher Lee Hartl is approved.

Michael Gandal

Jason Ernst

Bogdan Pasaniuc

Daniel H. Geschwind, Committee Chair

University of California, Los Angeles

2019

Table of Contents

ACKNOWLEDGEMENTS

This work would not have happened without the support of my marvelous and outsanding wife, Katherine, the guidance of my Ph.D. adviser, Dan Geschwind, and the advice of my parents, Dan and Christine.

Specific thanks to: William Pembroke and Gokul Ramaswami for many fruitful discussions, sharing of code, and their role as co-authors on the brain co-expression atlas manuscript resulting from the analyses presented in this thesis. Elizabeth Ruzzo and Jessica Rexach for floating valuble ideas for probing the biological meaning of gene networks. Sandrine Müller and Kasper Lage for discussions of region-specificity and protein-protein interactions, and their role as co-authors of the brain co-expression atlas manuscript. Ashis Saha, Princy Parsana, and Alexis Battle, also co-authors, for spirited debates on the appropriate treatment of RNA-seq data for co-expression analysis; and for performing the initial all-important bioinformatics (alignment, quantification, QC). To my lab-mates in the Geschwind lab, whose support of the bioinformatics internal meeting provided a forum for the dissemination and rapid development of the methods developed in this thesis.

VITA

Education

09/2005 to 06/2009	Harvard College B.A. Applied Mathematics
08/2014 to present	University of California, Los Angeles David Geffen School of Medicine at UCLA Bioinformatics Inter-Departmental Program

Research Articles

Ruzzo, E. K., Pérez-Cano, L., Jung, J.-Y., Wang, L., Kashef-Haghighi, D., **Hartl, C.**, Singh, C., Xu, J., Hoekstra, J. N., Leventhal, O., Leppä, V. M., Gandal, M. J., Paskov, K., Stockham, N., Polioudakis, D., Lowe, J. K., Prober, D. A., Geschwind, D. H. & Wall, D. P. Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks. Cell 178, 850–866.e26 (2019).

Won, H., Huang, J., Opland, C. K., **Hartl, C. L.** & Geschwind, D. H. Human evolved regulatory elements modulate genes involved in cortical expansion and neurodevelopmental disease susceptibility. Nature Communications 10, (2019).

Gandal, M. J., Haney, J. R., Parikshak, N. N., Leppa, V., Ramaswami, G., **Hartl, C.**, Schork, A. J., Appadurai, V., Buil, A., Werge, T. M., Liu, C., White, K. P., Horvath, S. & Geschwind, D. H. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. Science 359, 693–697 (2018).

Parikshak, N. N., Swarup, V., Belgard, T. G., Irimia, M., Ramaswami, G., Gandal, M. J., **Hartl, C.**, Leppa, V., Ubieta, L. de la T., Huang, J., Lowe, J. K., Blencowe, B. J., Horvath, S. & Geschwind, D. H. Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. Nature 540, 423–427 (2016).

Van der Auwera, G. A., Carneiro, M. O., **Hartl, C.,** Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S. & DePristo, M. A. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics 11.10.1-11.10.33 (2013). doi:10.1002/0471250953.bi1110s43

Walker R.L., Ramaswami G., **Hartl C.**, Mancuso N., Gandal M.J., de la Torre-Ubieta L., Pasaniuc B., Stein J.L. & Geschwind D.H. Genetic control of gene expression and splicing in the developing human brain. bioRxiv pre-print doi.org/10.1101/471193

Mercader, J. M., Liao, R. G., Bell, A. D., Dymek, Z., Estrada, K., Tukiainen, T., Huerta-Chagoya, A., Moreno-Macías, H., Jablonski, K. A., Hanson, R. L., Walford, G. A., Moran, I., Chen, L., Agarwala, V., Ordoñez-Sánchez, M. L., Rodríguez-Guillen, R., Rodríguez-Torres, M., Segura-Kato, Y., García-Ortiz, H., Centeno-Cruz, F., Barajas-Olmos, F., Caulkins, L., Puppala,

S., Fontanillas, P., Williams, A. L., Bonàs-Guarch, S., **Hartl, C.,** …, Flannick, J., Jacobs, S. B. R., Orozco, L., Altshuler, D. & Florez, J. C. A Loss-of-Function Splice Acceptor Variant in IGF2 Is Protective for Type 2 Diabetes. Diabetes 66, 2903–2914 (2017).

Manning, A., Highland, H. M., Gasser, J., Sim, X., Tukiainen, T., Fontanillas, P., Grarup, N., Rivas, M. A., Mahajan, A., Locke, A. E., Cingolani, P., Pers, T. H., Viñuela, A., Brown, A. A., Wu, Y., Flannick, J., Fuchsberger, C., Gamazon, E. R., Gaulton, K. J., Im, H. K., Teslovich, T. M., Blackwell, T. W., Bork-Jensen, J., Burtt, N. P., Chen, Y., Green, T., **Hartl, C.**, … Altshuler, D., McCarthy, M. I., Gloyn, A. L. & Lindgren, C. M. A Low-Frequency Inactivating AKT2 Variant Enriched in the Finnish Population Is Associated With Fasting Insulin Levels and Type 2 Diabetes Risk. Diabetes 66, 2019–2032 (2017).

Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D. J., Rivas, M. A., Perry, J. R. B., Sim, X., Blackwell, T. W., Robertson, N. R., Rayner, N. W., Cingolani, P., Locke, A. E., Tajes, J. F., Highland, H. M., Dupuis, J., Chines, P. S., Lindgren, C. M., **Hartl, C.**, …, Altshuler, D. & McCarthy, M. I. The genetic architecture of type 2 diabetes. Nature 536, 41–47 (2016).

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., **Hartl, C.**, Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D. & Daly, M. J. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43, 491–498 (2011).

Marth, G. T., Yu, F., Indap, A. R., Garimella, K., Gravel, S., Leong, W., Tyler-Smith, C., Bainbridge, M., Blackwell, T., Zheng-Bradley, X., Chen, Y., Challis, D., Clarke, L., Ball, E. V., Cibulskis, K., Cooper, D. N., Fulton, B., **Hartl, C.**, Koboldt, D., Muzny, D., Smith, R., Sougnez, C., Stewart, C., Ward, A., Yu, J., Xue, Y., Altshuler, D., Bustamante, C. D., Clark, A. G., Daly, M., DePristo, M., Flicek, P., Gabriel, S., Mardis, E., Palotie, A. & Gibbs, R. The functional spectrum of low-frequency coding variation. Genome Biology 12, R84 (2011).

Chapter 1 Systems biology approaches to neuropsychiatric disease

1.1 Systems biology approaches to neuropsychiatric disease

There are no known gross anatomical or neurological deficits that are hallmarks of common neuropsychiatric disorders (Bray2018). This contrasts with certain neurodegenerative diseases such as Alzheimer's disease (AD), characterized by the formation of tau tangles and amyloid plaques – or Parkinson's disease (PD), characterized by the aggregation of proteins into Lewy bodies. This means that, for neuropsychiatric disorders, there are no known specific mechanisms to investigate, and no disease-associated quantitative traits to measure.

Systems biology has the potential both to identify mechanisms with an important role in neuropsychiatric disease and to define molecular traits that correlate with disease risk and severity (Gandal2016). Since molecular interactions drive cellular behavior, which in turn determines brain tissue function, which ultimately gives rise to behavior and psychology, we take the most comminuted possible view: we model the brain as an enormous and dynamic collection of molecules, and hypothesize that gross differences in behavior correspond to molecular differences somewhere in the brain (Hawrylycz2016). This hypothesis has, over the past two decades, motivated multiple large-scale studies of spatio-temporal gene, protein, and epigenetic expression in the brain (Geschwind2000, Gong2003, Hawrylycz2012, Hernandez2012, Ramasamy2014, Zeisel2018). These studies provide overwhelming evidence in favor of the hypothesis, having identified molecular proxies for brain development (Parikshak2013), cell types, cellular processes (Oldham2008), and neuropsychiatric disease status (Parikshak2016, Gandal2018a, Radulescu2018). Modern quantitative genetic approaches – in particular TWAS (Gamazon2015) and heritability partitioning (Yang2011, Bulik-Sullivan2015) – provide confirmatory evidence by demonstrating that risk for certain neuropsychiatric diseases concentrates within a limited set of cells – e.g., in DNA regions that

are accessible in neuronal nuclei (Skene2018, Fullard2018, delaTorre2018, Polioudakis2018, Sullivan2019).

In spite of these successes, for neuropsychiatric disorders, disease etiologies remain opaque and endophenotypes elusive. One possibility is that some of these the deficits manifest in cell behavior not present in fresh-frozen post-mortem brain tissue at detectible levels (at current sample sizes). Another possibility – one which is rejected by the work in this thesis – is that the historical focus of expression studies on cortical regions may have missed disease signatures present in other regions. One final possibility is that functional perturbation in these diseases acts much like genetic risk: multiple dysregulated causal pathways exist even in normal individuals, but behavioral patterns only start to appear with a sufficient amount of dysregulation. Regardless, what we have learned from systems biology so far is the low-hanging fruit: the most-tightly co-regulated pathways, and those most severely altered in disease, that they can be identified in only dozens to hundreds of samples. As sample sizes grow, so too will out insight into the systems biology of neuropsychiatric disease.

1.1a Systems biology has implicated neuron-related pathways across neuropsychiatric disease

Since the inception of high-throughput RNA screening in 1995 (Schena1995) the number of expression profiles of human brain tissues or cells has grown exponentially: over 140,000 profiles are publicly available today and comparative disease studies now involve hundreds of individuals (Collado-Torres2019, Gandal2018a). Simultaneously, cohort sizes for both population-based (GWAS) studies and family-based studies have achieved substantial sample sizes (Sullivan2019). These data have supported multiple systems-biologic analyses aimed at identifying molecular pathways, cell types, and time-points involved in neuropsychiatric disease.

These studies have definitively linked both the pathology (e.g., observed differences in gene expression and histone modifications) and genetic risk (e.g., accumulation of risk-conferring mutations) of major neuropsychiatric disorders to genes expressed in cortical regions, to genes expressed in neurons, and to genes expressed in fetal and early postnatal time-points (Sullivan2019).

Contrastive studies such as differential expression analysis aim directly to identify molecular pathologies that differentiate brains from positively-diagnosed individuals from brains of normal individuals. The largest studies to date combine data from 8 prior studies with additional novel brain samples. Gandal *et al.* (shared pathology, Gandal2018a) looked across microarray and RNA-seq data in 407 brains representing 6 neuropsychiatric diseases and 293 controls, and identified a consistent pathology across ASD, BP, and SCZ: down-regulation of a large set of neuronal genes, and concomitant up-regulation of astrocyte genes. These results were recently replicated in an independent collection of brains wherein a small synapse-related module shows down-regulation across ASD, SCZ, and BP, while an adherens-junction related module shows up-regulation across these disorders (Guan2019). Gandal *et al.* (transcriptome dysregulation, Gandal2018b) examined more than 2,000 brains representing ASD, BP, and SCZ; identifying thousands of molecular differences at both the gene and isoform level, providing a more nuanced view of neuronal and glial dysregulation: modules enriched for trans-synaptic signaling and ribosomal turnover appear up-regulated across all disorders, while axonal, ion channel, and mitochondrial modules are downregulated. A recent single-nucleus sequencing study found a more complicated signature[1]: each neuronal class shows more up-regulation than down-regulation in ASD brains in terms both of numbers of genes and median log-fold-change

[1] Methods for within-cell-type single-nucleus differential expression have not yet been extensively evaluated, and the use of unsupervised methods for cell clustering may introduce inadvertent biases.

(with the exception of NRGN+ neurons; Velmeshev2019) suggesting that mRNA transport and maintenance in the cytoplasm may play an important role in defining tissue-level pathology. Yet in both cases, genes that compose the synapse or that are responsible for synaptic signaling were identified as differentially expressed in mature ASD brains.

Comparative *in vitro* studies have repeatedly established differences between normal and disease-model cultures in neurprogenitor cell (NPC) differentiation, neuronal migration, and neuronal maturation. Because it is impossible to obtain samples from developing ASD brains, comparative studies of developmental trajectories have all been *in vitro*. While these associations were initially observed in mouse models (Fukuda2005), they are recapitulated in recent studies of iPSC-derived cultures and neurospheres (Schafer2019, Lewis2019, DeRosa2018, Adhya2018). The observations for SCZ are different: noting a decrease in self-proliferation of NPCs, and defects in mature neurons, but no observation of deficits in maturation (Moslem2018). This may suggest that pathology may arise earlier in ASD than in SCZ, and even that early-developmental time-points may play a larger role in ASD than in SCZ. Gandal *et al.* (transcriptome dysregulation, Gandal2018b) examined more than 2,000 brains representing ASD, BP, and SCZ; identifying thousands of molecular differences at both the gene and isoform level

Integrative systems biology studies seek to identify and characterize functional pathways or networks in normal brains, and then use known genetic associations to identify those that carry more disease risk than would be expected by chance. This approach has been used to link ASD and SCZ risk to genomic regions that are accessible in mature neurons (Lake2017) and/or cortical plate (delaTorre2016), as well as to acetylated (active) histones within the cortex in fetal, infant, and adult brain (Li2018). Differential gene expression (e.g., between brain regions) can

also be used to identify likely important genes, regions, or cell types, and this approach has implicated neurons and adult cortical regions (where neurons are most abundant) across all neuropsychiatric disorders (Finucane2018), as well as both excitatory and inhibitory neurons in fetal brain for ASD and SCZ, and outer radial glia carrying outsized risk for SCZ and BP (Polioudakis2018). Finally, co-expression networks and pathways can be used to define where (e.g., what regions, cell types, or cell components) and when (e.g., early gestation, development, adult brain) risk genes act. These approaches have implicated genes involved in fetal and early-postnatal development in autism (Parikshak2013), synaptic genes and neuronal differentiation in schizophrenia (Schijven2018), and whole-brain (high-fidelity) neuronal genes across all neuropsychiatric diseases (Graham2018).

In spite of the beauty of the methodologies that predominantly converge onto cortical tissue, neuronal cells, and synaptic genes, it is difficult to resist an ironic interpretation. It does, after all, seem a great deal of work to confirm that common psychological disorders have a neuronal basis – a hypothesis for which there is evidence dating back to the 19th century (Weinstein1954). Instead, I think this is a consequence of methods outpacing data, particularly data about gene ontology. The ontologies for which we have power to detect enrichment are very broad (e.g. "synapse"), many ontologies are incomplete, and many more have yet to be catalogued. Later in this thesis I will identify co-expression modules associated with activity-dependent transcription and activity-dependent bulk endocytosis – a specific means of neurotransmitter-reuptake in highly-activated neurons. Notably, I identified this module prior to the time that gene sets corresponding to either process were published. Thus, the field of functional and integrative genomics has already reached a point where the information about certain gene relationships contained in expression data is not well represented in ontological

databases, such as Gene Ontology (Ashburner2000), especially with regards to fine grained aspects of neural function. As sample sizes increase, so too will the power to make fine-grained distinctions between gene sets, increasing this semantic gap – but at the same time empowering systems biology analyses to identify conclusively neurological pathways that drive and/or characterize disease etiology.

1.1b Brain co-expression network analysis: an overview and 15-year summary

Co-expression network analysis seeks to infer biological relationships between genes on the basis of their co-expression in relevant cells or tissues. These relationships can be grouped into functional units – termed modules –which in turn can be statistically analyzed to determine their likely biological role in the system, potential for disease relevance, or shared transcriptional regulators (vanDam2017). From a purely statistical viewpoint, co-expression network analysis is a form of unsupervised learning that aims to identify gene clusters (i.e., modules) and extract meaningful features (e.g., module membership scores: "kME"). The dozens of algorithms to do this have been extensively reviewed or compared elsewhere (Allen2012, Karmideen2012, Jay2012, Ballousz2015, Mahfouz2016, Saelens2018, Jha2019), but for well-controlled data where the biological system is the major driver of measured expression, even fundamentally different approaches yield functionally equivalent results. Indeed, later in this thesis I will develop two methods that eschew the network concept altogether, yet provide substantively equivalent modules to the most widely-used method, weighted gene co-expression network analysis (WGCNA; Zhang2005). Therefore for this section, I treat any unsupervised learning method applied to genes using gene expression – from hierarchical clustering to latent Gaussian processes – as an instance of co-expression network analysis.

Table 1 condenses the past 15 years of publications on co-expression and brain co-expression into short summaries of highly-cited papers. This collection shows steady progress in four major areas: i) methodology, ii) defining molecular processes that appear altered between different populations of brains, iii) identifying brain-specific, brain-region-specific, and cell-type specific genes, and iv) linking genetic risk to tissues and cell types on the basis of expression. Over this period, sample sizes have increased by roughly a factor of 10: Oldham2008 analyzed tissues in *N*=24 (CBL) to *N*=67 (CTX) brains; while Ramasamy2014 and Mostafavi2018 profile *N*=134 and *N*=418 brains respectively. Single-cell and single-nucleus studies are just now reaching the sample sizes of early systems-biology studies of the brain: Velmeshev2019 (not in table 1 as it is too recent) contains profiling of N=15 ASD brains and N=16 normal brains.

Combined meta-analyses and mega-analyses play a significant role in re-evaluating and extending prior studies: Gandal2018 utilizes microarray and RNA-seq data from nearly all previously published comparative expression studies between normal and neuropsychiatric brains, refining the co-expression relationships identified in Garbett2008, Maycox2009, Voineagu2011, Parikshak2016, and other publications not included in Table 1. Notably the largest directly-measured driver of expression variance in this study is diagnosis (though unmeasured biological variance – e.g., cell type heterogeneity – likely explains more). This study identifies robust co-expression modules with differential expression across all disorders (CD1, CD5, CD10, CD13 – neuron, CD4 – astrocyte) differential expression specific to (or much stronger in) ASD (CD11 – microglia) and depression (CD2 – receptor and hormone activity), or to alcoholism and SCZ (CD11 – endothelial cells), and establishes that the degree of similarity of (differential) expression patterns between disorders mirrors the similarity of genetic risk profiles. Because the pairwise differential expression correlation so strongly relates to correlation of risk

profiles, this study provides the first clear evidence of molecular phenotypes that are co-heritable with neuropsychiatric disease, although additional work is needed to convert the per-gene log-fold-change (log-FC) estimates into a phenotype that can be calculated in a single, undifferentiated and potentially unphenotyped cohort.

Kelley2018 (not in Table 1) is a meta-analysis of expression data from Ramasamy2014, Hawrylycz2016, and seven other studies not summarized by Table 1 (N=7221 samples from n=840 brains), with the aim of generating canonical cell type and neuronal subtype markers both within particular brain regions and across all brain regions, in order to reflect cell type diversity at the regional level. The genes identified by the simple approach of this study were entirely consistent with marker genes defined in single-cell sequencing studies, despite not using any single-cell data. A separate meta-analysis of five single-cell datasets identified novel genes that had not previously been associated with neurological cell types (McKenzie2018), and each of these genes are highly-ranked in the Kelly2018 resource for the appropriate cell type, which demonstrates that the bulk expression data from hundreds of brains can define marker genes at least as well as single-cell sequencing from tens of brains. These markers were then used to estimate cell type abundance within several CNS datasets, and establish that much of the age-related change in observed gene expression derives from overall cell type composition; also, when controlling for cell type abundance in AD, a large number of neuron-specific genes are up-regulated, suggesting that there are response pathways in neurons that are up-regulated in AD brains that have been missed, or identified as down-regulated, since neurons as a class are less abundant, and their markers thus under-expressed, in affected brains. It should be stressed that the cell type identity of a gene is not necessarily static, and may change as a result of disease,

which implies that there remains an under-appreciated role for disease-only co-expression network analysis.

Finally, advances in quantitative genetics – in particular partitioned heritability tools such as GCTA (Yang2011) and LD Score Regression (Bulik-Sullivan2015) or meta-analysis tools such as MAGMA (deLeeuw2015) or TWAS (Gamazon2015) – have prompted a wave of integrative analyses aimed at attributing genetic risk to tissues, cell-types, pathways, and to single genes. These approaches will be discussed in greater depth in the next section. One notable observation from Fromer2016, Parikshak2016, and Gandal2018 is that a large fraction of the co-expression signatures that show differences between cases and control also account for a disproportionate amount of genetic risk. It is important to stress that the expression differences that result from a disease need not be the same as the dysregulation that leads to the disease. The fact that genetic risk aggregates disproportionately in these gene sets suggests that they may have role both in the development of the disease as well as in its presentation. These modules are therefore the strongest candidates to investigate: should they give rise to measurable and heritable molecular signatures, they would form the basis for neuropsychiatric endophenotypes.

The fundamental challenge remaining in the systems biology approach to neuropsychiatric disease is to translate observed molecular differences into clear, measurable, disease co-heritable endophenotypes. The recent publications cited in this section suggest strongly that this goal is within reach, and that the remaining work is largely technical: to distill disease-related and cell-related molecular signatures into a single, easily-calculable measure that i) distinguishes normal brains from diseased brains, ii) tracks with disease severity, and iii) is heritable within control samples alone. Recent work in our group has made progress on component (ii) by identifying multi-omic signatures that appear to correlate with ASD symptom severity. There is some

evidence that (iii) is achievable in that module eigengenes are known to be heritable (Leduc2012, Scott-Boyer2013), and Gandal2018b identifies a number of genes whose expression appears to correlate with polygenic risk for ASD or SCZ. Simultaneously, the field continues to progress with increasing sample sizes and new sources of data such as single-cell sequencing, enabling the refinement of broadly neuronal signatures into specific disrupted processes such as calcium gradient maintenance or bulk endocytosis.

Table 1 Selection and summary of highly-cited brain systems biology papers over the past 15 years

Reference	Year	Title	Cit/y	Summary
Caceres2003	2003	Elevated gene expression levels distinguish human from non-human primate brains	31	Neuronal, cell-growth-related, and chaperone genes are up-regulated in humans compared to NHPs
Lee2004	2004	Coexpression analysis of human genes across many microarray data sets	50	Reproducible co-expression across many tissues occurs within core pathways: transcription, translation, cell division, metabolism, cell adhesion, and immunity
Subramanian2005	2005	Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles	1214	Definition of efficient and rigorous statistical methodology for gene set enrichment
Zhang2005	2005	A general framework for weighted gene co-expression network analysis	173	Methodology of co-expression network construction, hub genes, and soft memberships
Oldham2006	2006	Conservation and evolution of gene coexpression networks in human and chimpanzee brains	38	Cross-tissue co-expression (brain patterning) differs between human and chimp. Cortical and striatal co-expression modules, reflecting energy metabolism, mitochnondrial respirtory chain, and synaptic genes, are more diverged between human and chimp than are cerebellar and white-matter modules
Carlson2006	2006	Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks	24	Highly-connected genes across three yeast co-expression networks are more likely to be evolutionarily conserved, and more likely to be essential
Lein2007	2007	Genome-wide atlas of gene expression in the adult mouse brain	265	Characterizes regional mRNA expression patterning, identifies cell-upregulated genes for major cell classes, and dendrite-transported mRNA species. Establishes coarse correlation between gross anatomy (brain regions) and molecular anatomy (expression measured within region)
Bansal2007	2007	How to infer gene networks from expression profiles	74	Review and comparison of methods for co-expression network inference and analysis. Notably finds that agreement between methods (consensus) for a co-expression edge does not increase the likelihood of the edge being "true."
Oldham2008	2007	Functional organization of the transcriptome in human brain	48	Establishes cellular heterogeneity as driver of co-expression modules; shows modules are re-producible across array studies; identifies shared modules across CBL, CDT, CTX; links modules to cell-types, neurogenesis, and sex
Miller2008	2008	A systems level analysis of transcriptional	28	Construction of unsupervised and supervised (pre-selected via trait correlation) modules in AD hippocampus and normal aging cortex;

		changes in Alzheimer's disease and normal aging		identifying commonalities, and also implicating mitochondrial and oligodendrocyte dysfunction in AD
Garbett2008	2008	Immune transcriptome alterations in the temporal cortex of subjects with autism	26	Small-cohort comparative study of expression in ASD vs normal brains, implicating inflammatory and autoimmune involvement in characterising ASD brains. Differentially expressed genes contain strong markers of astrocytes and microglia.
Johnson2009	2009	Functional and evolutionary insights into human brain development through global transcriptome analysis	44	Expression in mid-fetal (18W, 19W, 21W, 23W) brains across 5 regions. Identification of genes representing early brain patterning; observation that sequences under accelerated evolution in the human lineage fall disproportionately near to these genes.
Dobrin2009	2009	Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease	13	Joint network of cross-tissue and within tissue expression correlation (adipose, hypothalamus; nodes are tissue-gene pairs) in a mouse obesity model identifies gene clusters associated with feeding behaviors, circadian rhythm, leukotrine metabolism, heat shock, and ion transport as associated with body weight.
Maycox2009	2009	Analysis of gene expression in two large schizophrenia cohorts identifies multiple changes associated with nerve terminal function	13	Analysis of cortical differential expression in BA10 of two cohorts of SCZ and control. Consistent differential-expression (i.e. replicated in the other cohort) observed in synaptic vesicle function and signal transduction.
Miller2010	2010	Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways	34	Consensus networks of 18 human brain-tissue expression datasets (regardless of region or disease status) compared to 17 mouse datasets (regardless of region or strain). Gross expression patterning mouse co-expression is conserved in humans, human-specific co-expression identified. One human-specific module overlaps AD-progression genes, PSEN1 and MOG show human-specific properties.
Torkamani2010	2010	Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia	13	Differential expression and differential co-expression analysis (using module overlap) in SCZ and control brain. All SCZ modules significantly overlap CTL modules suggesting no dysregulation at the level of major modules. 5 modules (OxPhos, neurogenesis, neuron development, chromatin, and synaptic transmission) enrich for SCZ differential expression. Several modules which in controls decrease with age fail to do so in SCZ brains.
Voineagu2011	2011	Transcriptomic analysis of autistic brain reveals convergent molecular pathology	139	Profiling of 29 ASD/29 CTL brains across; 2 CTX regions, 1 CBL. Identifies synaptic and immune modules differentially expressed between ASD and CTL; and identifies cortical patterning modules differential in

				BA9 vs BA41 in CTL but not in ASD, implicating impaired cortical patterning.
Davies2012	2012	Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood	67	Methylation profiling of 6 brain regions and blood. The most extreme tissue methylation differences are mirrored by expression differences; brain/blood differences concentrate in nervous system development pathways. Cortical differences enrich for neurogenesis, neuronal function, and forebrain development. Co-methylation summarizes tissue-specific methylation patterns.
Ponomarev2012	2012	Gene coexpression networks in human brain identify epigenetic modifications in alcohol dependence	35	Profiling of 17 alcoholic and 15 control brains: AMY and CTX; H3K4me measured in CTX for 4 cases 4 controls. Co-expression networks and meta-networks corresponding to major cell classes and organelles (ribosome, mitochondria, nucleus). GC-rich modules (splicing) appear up-regulated in chronic alcoholism, and GC-poor modules (zinc finger, ubiquitination) down-regulated.
Konopka2012	2012	Human-specific transcriptional networks in the brain	25	Digital gene expression (3' profiling) of CDT, HIP, and CTX (frontal pole) in human, chimp, and macaque; meta-analyzed with prior array-based studies. Hundreds of genes show increased expression in humans within each region, with those upregulated in CTX enriched for neuron-related ontologies. Identifies primate-preserved co-expression modules (general CNS development) and human-specific co-expression modules, including a CLOCK and a FOXP2 module, implicating human-specific components of circadian rhythm and language pathways.
Parikshak2013	2013	Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism	82	Co-expression networks built from RNA-seq expression across brain development (8 PCW to 12 Mo post-birth) to identify developmental trajectories. 17 modules identified, of which 5, M2, M3, M13, M16, M17, enrich for ASD candidate genes or *de novo* PTVs. M2/M3 contain early-expressed genes related to chromatin modification and transcriptional regulation; M3 is particularly enriched in VZ and SVZ; M13/16/17 are late-expressed and implicate neuronal maturation and synapse production.
Willsey2013	2013	Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism	74	Combined analysis of 1,043+56 ASD families (559+56 quartets, 444 trios) to identify dnLOF genes. Uses developmental expression to build bipartite networks using 9 confident dnLOF "seed" genes. Of 52 modules, 4 enriched for likely ASD genes, related to deep layer glutamatergic projection neurons and the inner cortical plate.

Miller2014	2014	Transcriptional landscape of the prenatal human brain	98	Atlas of microarray-profiled and ISH gene expression across 300 regions from 4 fetal brains (2 15-16 PCW, 2 21PCS). Regions follow clear spatial expression patterns. WGCNA used to build brain-wide and germinal layer networks, identifying spatial and temporal patterning modules, as well as modules corresponding to fetal cell types; some of which overlap differentially-expressed genes in adult NPD brains.
Ramasamy2014	2014	Genetic variability in the regulation of gene expression in ten regions of the human brain	63	eQTL study of 134 brains within 10 brain regions, identifying tens of thousands of region-specific and cross-regional eQTLs; identified cases where the target gene of a cis-QTL appears to swithc between regions; and identified established neurodegerative GWAS hits that are QTLs for one specific gene in their region.
Goyal2014	2014	Aerobic glycolysis in the human brain is associated with development and neotenous gene expression	32	Resting FMRI data meta-analyzed and condensed into regional Aerobic Glycolisis score. BrainSpan developmental expression data to identify signatures of brain development ("neoteny scores"). Neoteny then computed within regions of the adult human brain atlas, and find strong correlation with AG scores; implying a continuous process of synapse growth and formation in the adult brain, predominantly in cortex.
Camp2015	2015	Human cerebral organoids recapitulate gene expression programs of fetal neocortex development	61	Single-cell sequencing in fetal neocortex and in iPSC-derived neurospheres. Cross-cellular expression patterns are largely similar, with most differences attributable to compounds present in the culture serum. Lineage trajectories are highly similar between NC and organoid.
Madabushi2015	2015	Activity-induced DNA breaks govern the expression of neuronal early-response genes	59	Initial observation that double-strand breaks induce expression of early-response genes was confirmed with Crispr-Cas9. Inihibition of DSB results in persistent early-response expression and no up-regulation; and Top2B knockdown precludes early-response gene expression, which is rescuable via targeted DSBs to the promoters.
Mo2015	2015	Epigenomic signatures of neuronal diversity in the mammalian brain	56	Single-cell methylation and RNA expression profiling in tagged nuclei (PV, Exc, VIP, mouse); identifying cell-specific patterns of non-CG hypomethylation that correspond to cell-specific expression; with stronger correlation than CG methylation or chromatin accessibility. ATAC-seq suggests TF-TF network rewiritng within neuronal subtypes
Prudencio2015	2015	Distinct brain transcriptome profiles in C9orf72-associated and sporadic ALS	37	Comparative RNA-Seq in cerebellum and cortex of C9orf72-expansion ALS (N=8), sporadic ALS (n=10) and control (N=8); identifying hundreds of differentially-expressed and differentially-spliced genes

				between groups. Differential modules implicate rRNA processing, neurons, and protein transport; and alternative polyadenylation is a major source of differential alternative splicing.
Fromer2016	2016	Gene expression elucidates functional impact of polygenic risk for schizophrenia	90	Comparative RNA-seq in DLPFC from N=258 SCZ subjects and N=279 controls used to identify differential expression, expression QTLs, and co-expression networks. One control module, M2c, shows enrichment for differentially-expressed genes and prior SCZ genetic associations; and appears to lose network connectivity in SCZ samples. This module relates to neuronal post-synaptic densities and activity-related cytoskeleton.
Lake2016	2016	Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain	83	Droplet-based RNA-seq of single NeuN+ nuclei, identifying 16 clusters which differ in the expression levels of canonical neuronal and cortical layer markers; among hundreds of newly-implicated marker genes.
Parikshak2016	2016	Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism	51	Comparative RNA-seq in cerebellum and cortex of ASD, Dup15q, and normal post-mortem brains, identifying differential expression and splicing. Co-expression networks implicate neuronal, astrocyte, and microglial patterns as differentially expressed, and within neurons implicating development/differentiation and synaptic transmission pathways.
Hawrylycz2016	2016	Canonical genetic signatures of the adult human brain	46	RNA-sequencing from 132 brain structures in 6 individuals; identifies ~8500 genes with consistent inter-region differences ("differential stability"); further grouped by WGCNA into 32 cross-regional co-expression modules, containing a broad distribution of cell type markers. Preservation analysis with mouse shows that many neuronal patterning modules are preserved, while glial patterning modules are not.
Bakken2016	2016	A comprehensive transcriptional map of primate brain development	40	Transcriptional atlas of Rhesus brain across developmental and aging timepoints, with corresponding MRI and ISH data. Shows that developmental processes (synaptogenesis, myelination) are slow-activating but sharply deactivating, and vary in onset and length across regions, and that cortical layer patterning shifts over time points. WGCNA groups expression patterns into modules that are expressed early, but persist well into adulthood
Cembrowski2016	2016	Spatial gene-expression gradients underlie prominent heterogeneity of CA1 pyramidal neurons	32	RNA-sequencing of CA1 hippocamal neurons, identifying spatial heterogeneity of transcription (primarily dorsal-ventral but also proximal-distal and superficial-deep); implying that spatial or connectivity cues may

				resolve CA1 neurons into subtypes, or drive within-type transcriptional heterogeneity. Several canonical binary marker genes showed gradients along one or more of these axes.
Gosselin2017	2017	An environment-dependent transcriptional network specifies human microglia identity	91	RNA-sequencing of purified human and mouse microglia ex-vivo and in-vitro; demonstrating significant transcriptomic changes due to in vitro culture conditions, impacting >50% of AD-associated genes. These differences appear to be due to the loss of brain-microenvironment signals, including TGF-beta. A wide range of culture conditions nevertheless failed to recapitulate ex-vivo transcriptional signatures.
Luo2017	2017	Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex	47	Single-cell methylation profiling of NeuN+ neurons using hypomethylation at marker genes to annotate 21 layer-, type-, and subtype- human neuronal clusters. Neuronal methylation is broadly conserved between human and mouse neurons, despite a slightly larger number of human clusters; with interneurons showing higher cross-species methylation conservation than excitatory neurons.
Mancuso2017	2017	Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits	33	TWAS study across 30 traits using GTEx gene expression data for weights (all tissues separately, multiple testing). Identifies 113 genes in a GWAS peak, and 24 genes not in a GWAS peak, that associate with SCZ.
Su2017	2017	Neuronal activity modifies the chromatin accessibility landscape in the adult brain	28	ATAC-seq and RNA-seq profiling of mouse granule neurons prior to and following electrical stimulation; identifying ~200,000 differences in chromatin accessibility between states; and ~1,200 differentially expressed genes. The chromatin differences appeared coherently to impact enhancer regions of differentially-expressed genes, and the newly-opened regions enriched for c-Fos binding motifs; and c-Fos knockdown significantly ablated activity-dependent chromatin and expression differences.
Galatro2017	2017	Transcriptomic analysis of purified human cortical microglia reveals age-associated changes	25	RNA-seq of purified microglia from 39 normal donors spanning a large adult age range, identifying ~500 genes that change in microglia with age; with many genes involved in actin dynamics being down-regulated, while adhesion, axonal guidance, and surface receptor genes showing mixed- up and down-regulation.
Nowakowski2017	2017	Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex	23	Single-cell RNA-seq of developing telencephalon; identifying cell types along the full timecourse from progenitor cells to postmitotic neurons. Identifies programs of radial glia and neuronal maturation, and

				emergence of cortical patterning signatures in maturing neurons.
Sousa2017	2017	Molecular and cellular reorganization of neural circuits in the human lineage	20	Profiling of 11 cortical regions, HIP, AMY, STR, MD, and CBC in 6 humans, 5 chimpanzees, 5 macaques. Identifies thousands of human up-regulated genes, both across and specific to tissues. Identifies TH+ interneurons as present in human and macaque yet absent from chimp cortex. Differences in cell type proportions, as well as within-cell-type differences on neurotransmitter gene expression, appear to drive species differences.
Gandal2018	2018	Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap	158	Meta-analysis of multiple comparative microarray and RNA-seq studies of neuropsychiatric disease, establishing that the degree of shared transcriptional signature mirrors the degree of co-heritability (high for ASD/SCZ, lower for others). Co-expression networks built from these data reflect differences in expression across all major cell types, with differences in neuronal modules shared across neuropsychiatric disease (but not alcoholism), and differenees in a microglia module apparently specific to ASD.
Finucane2018	2018	Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types	76	LD-score regression analysis of tissue-upregulated genes (top 10% one-vs-rest) for 48 diseases and traits. 34/48 show enrichment for at least one tissue-upregulated gene set; SCZ and BMI show enrichment for nearly all CNS tissues. SCZ and BP show enrichment in cortex-upregulated and neuron-upregulated genes (compared to the rest of the brain, and other cell types, respectively).
Lake2018	2018	Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain	75	Single-cell RNA-seq and chromatin accessibility mapping from nuclei in visual cortex, frontal cortex, and cerebellum; identifying the standard array of excitatory and inhibitory cell types classified by expression of subtype markers (SST, PVALB, etc). Known and nominal GWAS associations are enriched in open chromatin for excitatory neurons (SCZ, ASD, BP), microglia (BD, MS, AD), and endothelial cells (MS)
Barbeira2018	2018	Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics	59	Application of gene expression mediation analysis to 44 tissues and >100 phenotypes, with hundreds of genes implicated in SCZ by virtue of expression mediation in some tissue (possibly not brain).
Gandal2018b	2018	Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder	41	Comparative RNA-seq in prefrontal cortex across ASD, SCZ, BP, and controls; parsing differential expression into gene, isoform, and lncRNA; showing that isoform-level dysregulation encompasses a wider variety of cell types and processes than does gene-level

				dysregulation. Previously-observed correlations in differential expression are recapitulated at the splicing event level. Isoform-level and integrative modules identify indepentent RBFOX1 modules which enrich for different disease signatures and cellular components; and distinct isoform switching events between neuropsychiatric diseases.
Mostafavi2018	2018	A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease	31	RNA-sequencing of post-mortem cortex from longitudinal study of aging and cognitive decline. Co-expression networks used to identify 47 modules, several of which show differential expression in AD, and m109 as being associated with cognitive decline; knockdown of important hub-genes in this module reduces extracellular AB42 levels.
Li2018	2018	Integrative functional genomic analysis of human brain development and neuropsychiatric risks	23	Chromatin profiling, methylation profiling, bulk tissue and single-cell RNA-seq across the human brain lifespan (8PCW-64PY) and 6 brain regions. Analysis focuses on spatio-temporal gene expression differences, summarized by 75 co-expression modules, of which 10 enrich for brain-related traits or disorders, with prenatal-specific open chromatin most strongly enriching for neuropsychiatric disease (SCZ,BP) heritability. Cell type decomposition identifies expression trajectories of distinct fetal and adult excitatory and inhibitory neurons; and shows that oligodendrocyte growth and/or development occurs at different timepoints in different brain regions.
Tyssowski2018	2018	Different neuronal activity patterns induce different gene expression programs	22	Capture RNA-sequencing (257 activity-related genes) of neuronal cultures, induced with brief or sustained activity. Brief and sustained activity resulted in different patterns of expression of ARGs, with immediate-response, delayed-response, and sustained-response genes. Follow-up enhancer RNA-seq identifies rapid and delayed enhancers, and the induction of these rapid enhancers requires the MAPK/ERK pathway.

1.2 Heritability and etiology of complex neuropsychiatric disorders

Neuropsychiatric genetics seeks to identify genetic mutations that confer risk for one or more neuropsychiatric disorders. This objective is motivated by a simple observation: two relatives are

far more likely to both suffer from neuropsychiatric disease than two non-relatives. This increase – the relative risk: 50-100 for ASD, 7-10 for BP and SCZ, and 2-3 for MDD (Schultze2018) – provides evidence that natural genetic differences are partly responsible for the epidemiology of neuropsychiatric disease.

There are a number of reviews on the history of genetic studies of neuropsychiatric disease from early linkage studies through modern population studies (Sullivan2012, Malhotra2012, Huguet2013, Gratten2014, Geschwind2015, Bray2018, Sullivan2019). These provide a historical perspective on genetic findings and the role that particular mutation classes (e.g. large structural variants or *de novo* mutations) have played in the genetic understanding of neuropsychiatric disorders. The literature relevant to this thesis concerns genetic architecture: identifying properties – functional or population-genetic – of genetic mutations that correspond to neuropsychiatric risk. The collection of methods for interrogating genetic architecture fall under the loose heading of heritability partitioning. Such methods probe mutational classes by comparing the fraction of total mutations they represent to the fraction of total heritability that they explain – in effect asking whether the mutational class confers an outsized proportion of risk liability. These approaches are used in six of the fifteen papers in Table 1 since 2017, and are becoming standard in genetic association studies.

Chapter 3 of this thesis applies these methods to expression data drawn from the human brain in order to evaluate the regional or cell-type specificity of genetic risk. Chapter 4 is concerned with incorporating trans-regulation and co-expression networks into models of genetic architecture. These chapters presume a background knowledge of these methods and the statistical models that underlie them. To that end, the remainder of this chapter provides a brief

review of heritability, polygenic risk, heritability partitioning, and the polygenic and omnigenic models of genetic architecture.

1.2a Heritability and genetic complexity of common neuropsychiatric disease

Genetic heritability (H^2 or h^2) is defined as the proportion of trait variance that can be attributed to genetic factors. This single number characterizes the propensity for a disease to run in families. The distinction between H^2 and h^2 is a classical one: H^2 ("broad-sense heritability") represents an ideal where the unknown (potentially nonlinear) mapping function from mutations to phenotype is known, and h^2 ("narrow-sense heritability") represents the case where mutations are treated linearly ("additively"). Importantly, both H^2 and h^2 can be estimated from examining disease occurrence within many families, without any direct knowledge of underlying mutations. This family-based approach has provided estimates as high as 90% for ASD, 80% for SCZ, 65% for BP, and 30% for MDD (Schultze2018). It should be noted that disease heritability may differ between populations due to both genetic and environmental differences: a recent study noted a 20% difference in Type-2 Diabetes heritability between US and European cohorts (72% EUR+AUS, 52% US; Avery2019), demonstrating that the concept of "the" heritability of a common disease often belies the underlying environmental and genetic complexity.

One of the many goals of medical genetics is to identify the specific variant loci that give rise to genetic diseases. Modern genetic studies can measure, directly (sequencing) or indirectly (imputation), more than 20 million variant alleles in the human population (Vergara2018). Genome-wide association studies (GWAS) perform these measurements in large cohorts of affected and normal individuals – modern studies of neuropsychiatric disorders range from 18,381 (ASD) to over 200,000 (MDD) (Sullivan2019) – to identify genetic variants that are more common in the affected population. These studies have found dozens to hundreds of risk-

conferring variants: from 5 in ASD to 144 in SCZ. Yet, these mutations represent a tiny fraction of the total risk, implying that there are thousands of contributing genes and mutations for these disorders.

Genetic complexity is a feature of polygenic diseases. Figure 1 – a re-creation of figure 1 from Wray2018 with parameters consistent with ASD and SCZ: $h^2 = 75\%$ and prevalence 1% – illustrates precisely the challenges posed by genetic complexity: i) The impact of the average causal mutation is very small, making causal mutations difficult to identify in practice; ii) Phenotypically normal individuals carry a wide range of genetic liability – one that overlaps the affected population; iii) Affected individuals each carry a unique pattern of mutations, making it impossible to study any particular mutation in isolation. One response to these challenges has been to search for similarities – such as pathways, protein-protein interactions, or co-regulation – between genes harboring genetic mutations that confer disease risk.

Genetic architecture describes the kinds of mutations or genes that drive the heritability of a disease. For instance, a disease may have a "rare variant" genetic architecture, where risk is conferred predominantly by rare variants of large effect size, in contrast to a "common variant" architecture, where most risk is conferred by many relatively common variants. For most studied traits (Schoech2019), including neuropsychiatric disease (Gaugler2014), rare variants have stronger individual effects than common variants – but due to their low frequency, they explain less heritability than common variants. Similarly, one can demonstrate that variants that alter gene expression levels, termed expression quantitative trait loci ("eQTL"), confer more heritability than all other types of variants across a wide variety of diseases and traits (Hormozdiari2018, Gamazon2019). These approaches have been adapted to search for functional genetic architecture: specific regions of the genome that are relevant to a biological function and

carry significant genetic risk for a given disease. For instance, Finucane2018 and Lake2018 from Table 1 respectively show that mutations near neuron-expressed genes, and mutations in neuronal open-chromatin regions, harbor a disproportionate amount of SCZ heritability.

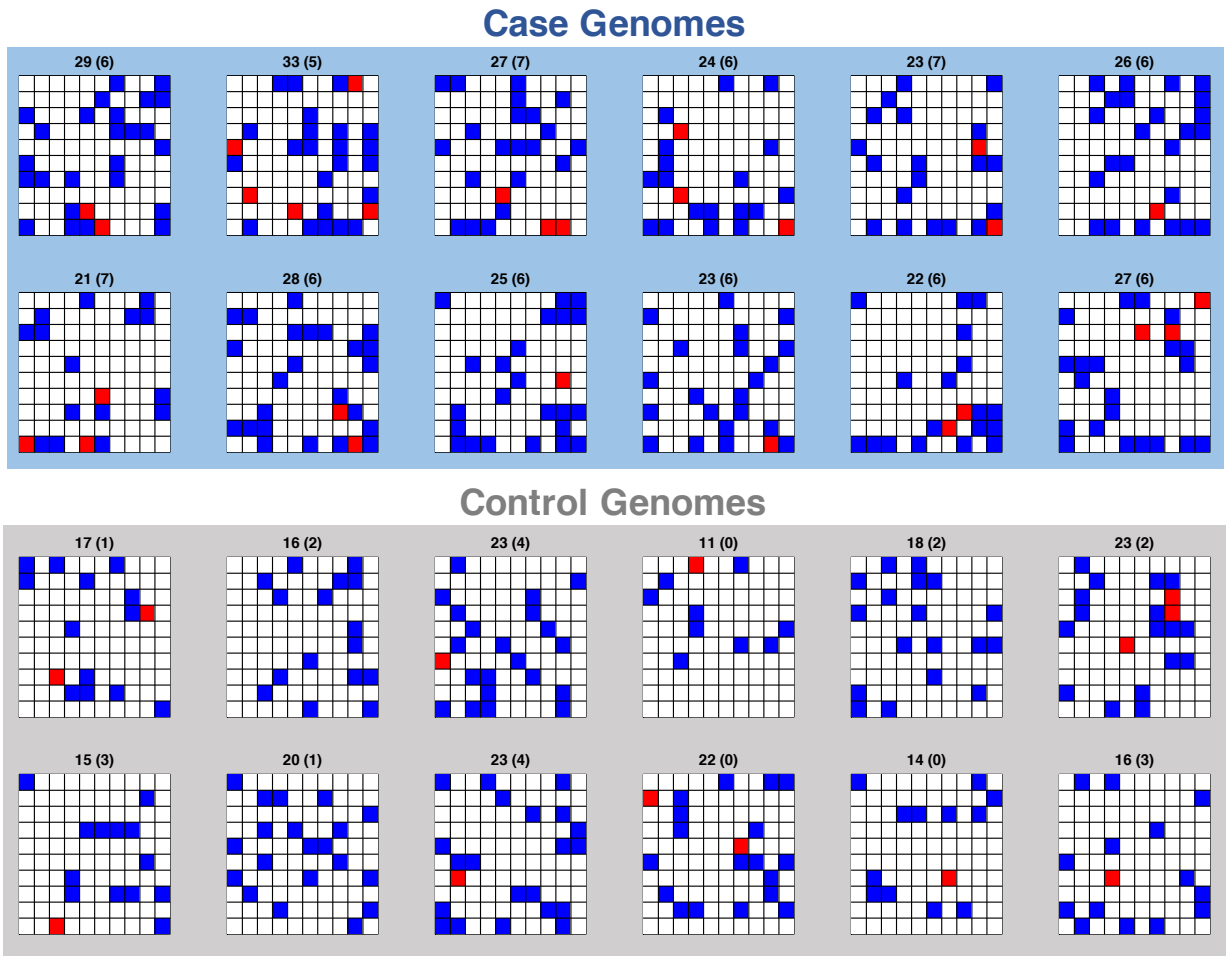

Figure 1.1: Genetic complexity under a polygenic model.

Simulation of polygenic architecture for a high-heritability trait, illustrating the complexity of genetic heterogeneity. Each box represents 100 causal loci in a genome and each square represents a genotype (blue for 1 risk allele, red for 2). Numbers at the top count the total number of risk alleles, and in parenthesis, the total number of risk alleles in the bottom row. Notably, non-diseased individuals carry many risk alleles. While the number of risk alleles is lower, on average, for control individuals, the distributions overlap, demonstrating the role that variant effect size and environmental effects have in complicating causal genetic factors.

1.2b Quantitative genetics: partitioning risk into regulatory elements and cell types

Functional genetic architecture can be investigated by the general method of partitioned heritability (Yang2011). Figure 1 summarizes the intuition behind heritability (or risk) partitioning. In this simulation, I assigned the last row of variants 10-fold higher disease risk than each of the first 9 rows, as though they represented variation in a fundamental disease pathway. There is a greater (average) difference in disease liability between the case and control

population in this row than for all other rows. The approach for partitioned heritability follows the same intuition: calculate the average liability difference between cases and controls for a given set of variants, and compare it to the variants not in the set. Because Chapter 4 utilizes and extends these models to investigate the genetic architecture of neuropsychiatric disease, the remainder of this chapter focuses on the prerequisite technical background.

The statistical model for disease genetics is the liability-scale linear model:

$$y^{(l)} = \alpha + X\beta + \varepsilon$$

Where $y^{(l)}$ is the disease liability[2] (a vector over all observed individuals), α is the average population liability, β are the per-variant risk coefficients (a vector over all observed mutations), and X is the genomic matrix, with each row containing the observed genotypes (0, 1, or 2) for each mutation in a single individual – normalized to 0 mean and unit variance. To map a previous concept into this model: a trait is complex (or polygenic) if a large number (more than 0.1%) of the entries of β are non-zero.

The full vector of variant effects, β, is only estimable when there are more observed individuals than there are variants.[3] Because of this, genetic associations use the "marginal" model, focusing only on a single variant:

$$y^{(l)} = \alpha + \beta_j x_j + \varepsilon$$

This treats each column of X (each variant) independently. Fitting this model to data produces a value, z_j, called the "marginal test statistic" for each variant. For sufficiently large (effective) sample sizes (n below), this statistic relates to the correlation:

[2] The liability is in general never observed; but instead a yes/no disease outcome with P[yes] increasing monotonically with $y^{(l)}$. The derivations for $y^{(l)}$ apply to binary outcome, but are more involved and will contain multiplicative terms dependent on the form of P.

[3] It is true that constrained likelihood and Bayesian estimation result in tractable models – but for efficient estimation the effective sample size should still exceed the number of non-zero entries of β.

$$z_j \sim N\left(\text{cor}\left(y^{(l)}, x_j\right)\sqrt{n}, 1\right)$$

Variants highly correlated to disease liability take on large values, while those showing no correlation to liability have small values. These z-statistics are the primary output of a GWAS study, and are commonly made publicly available as downloadable summary statistics files. Absolute Z-scores above a certain value (typically about 5.5; Johnson2010) pass the statistical threshold of association, thereby becoming known risk loci for the disease.

The goal of a functional genetic architecture screen is to identify patterns within these Z-scores by aggregating them into genetic pathways. The simplest of these approaches is gene set enrichment analysis (GSEA), which I apply in chapter 3. GSEA tests whether the average Z-score is higher in one pathway versus the background. Though conceptually simple, implementations such as MAGMA (deLeeuw2015), MAGENTA (Segre2010), and GSA-SNP2 (Yoon2018) use variations on this general idea to improve sensitivity and specificity. An implementation of GSEA defines three items: (i) How to aggregate contiguous variants within a single interval to a single score (variant aggregation), (ii) How to aggregate multiple intervals within a set to a single score (interval aggregation), and (iii) How to determine significance (significance). Chaining these steps together provides enrichment statistics and p-values for a hypothesized genetic pathway. Table 2 provides these implementation details for the above algorithms.

LD score regression is an alternative approach where the correlation between genotypes (known as linkage disequilibrium: "LD") is leveraged to partition risk across genomic intervals. Because the Z-scores from the GWAS marginal consider the impact of a single variant alone, two variants with a high correlation will necessarily have similar Z-scores. In cases where genomic risk is widespread – e.g., polygenic diseases variants with high total correlation will

tend to have extreme Z-scores. In fact, there is a linear relationship between the sum of a variant's r^2 values to all other variants (termed the "LD score"), and the slope of that line is the average risk. By computing the LD score with respect to a set of variants – rather than all variants – the average risk within that set can be computed, and contrasted to other sets. Letting $C_1, C_2, \ldots$ denote the variant sets, then:

$$E\left[z_j^2\right] = N \sum_C \tau_C \ell(j, C) + M$$

$$\ell(j, C) = \sum_{k \in C} \mathrm{cor}(x_j, x_k)^2$$

where M is an arbitrary constant and N is the number of samples. In this approach, every variant is associated with two values: its squared Z-score z_j^2 and its LD-score $\ell(j,C)$. The slope between these two values is directly proportional to the average disease risk conferred by a variant within the set C. In this way, risk can be partitioned and evaluated across different variant sets.

Algorithm	Naïve	MAGENTA	MAGMA	GSA-SNP2
Variant aggregation	mean(Z^2)	min(p) – lm(N,d)	Fisher's method*	min(p) – C(N)
Interval aggregation	$\frac{\mu_{set} - \mu_{all}}{\sigma_{all}} \sqrt{n_{set}}$	# intervals in top 5% of all intervals	None	$\frac{\mu_{set} - \mu_{all}}{\sigma^*_{all}} \sqrt{n_{set}}$
Significance	Standard normal	Matched permutation	Linear Mixed Model	Standard normal

Table 2 Implementations of gene set enrichment analysis (GSEA).

The Naïve approach to variant aggregation takes the mean of chi-square statistics over an interval, which can be deflated if the interval is large. MAGENTA and GSA-SNP2 instead take the log-minimum p-value; and adjust this value by the expected log-p value for intervals of size d containing N variants. MAGENTA uses a linear model, while GSA-SNP2 uses a cubic spline. MAGMA on the other hand uses Fisher's method, adapted to correlated test statistics. Both the Naïve method and GSA-SNP2 contrast the mean score of the interval set to the mean score of all intervals, and convert this to a Z-score (GSA-SNP2 uses a slightly inflated value for σ_{all}, and ignores adjacent genes). MAGMA uses the number of intervals whose score was in the top 5% of all interval scores. For the Naïve approach and GSA-SNP2, pathway significance is determined by converting the pathway Z-score directly to a p-value using the standard normal survival function. MAGENTA draws 10,000 gene sets of identical size and similar inter-gene distance, and re-computes the interval aggregation scores. MAGMA regresses pathway indicator variables against per-interval scores from Fisher's method [(*) Brown's extension to dependent tests] as a response variable, using the correlations from Fisher's method as the assumed residual covariance in a mixed model.

All three non-naïve methods hide additional complexity with regards to variant or interval filtering thresholds applied prior to model fitting.

1.3c Gene networks and genetic architecture: the omnigenetic model versus systems biology

Functional genetic architecture studies have identified broad cell types and gene ontologies related to various diseases, but have yet to identify highly-specific causal pathways consisting of at most dozens of genes. On the one hand, this may be a result of insufficient ontological knowledge, low sample sizes for many traits, and even lower sample sizes for tissue and single-cell expression data. On the other hand, it may be a property of the genetic architecture itself, where despite the proximate cause coming from one or two small functional pathways, the ultimate cause arises from hundreds to thousands of genetic perturbations that lead to aberrant operation of the proximate pathways. In such a case, we would expect to observe heritability to cluster not into the proximate causal pathways, but instead into the tissues and cell types where the proximate pathways operate.

The omnigenic model (Boyle2017) is a formalization of this idea, in which the genes within the proximate causal pathways are termed "core genes," while the genes that contribute indirectly are termed "peripheral genes." Importantly, the distinction between core and periphery is imagined to be a property of a *cellular gene network* that reflects co-regulation and protein-protein interactions within the relevant cell types. Mutations within the core of the gene network – implicitly defined as a set of communities containing a small proportion of the total genes – can achieve very high effect sizes for the trait; while mutations outside of the core cannot.

By relating effect size to the property of a variant – the network distance of the gene it impacts – the omnigenic model describes the *network* genetic architecture of a trait. Formally, a model of genetic architecture specifies a distribution for β. For example, to test the relationship between variant frequency (p) and risk effect (β), a common model is

$$\beta|p \sim N(0, \sigma_g^2(2p(1-p))^\gamma)$$

Where $\gamma < 0$ corresponds to an architecture where high-risk variants are more likely to have lower frequencies. In chapter 4, I define a model for network genetic architecture analogously by replacing frequency with *network distance*, d_G, which measures how far each gene is from the core set of genes:

$$\beta|d_G \sim N(0, \sigma_g^2(1 + d_G)^\gamma)$$

Where the value of d_G for a variant corresponds to the distance for the gene in which it is located. $\gamma < 0$ corresponds to architecture where high-risk variants are more likely to occur near to or within the network core. This includes the omnigenic model, where γ is assumed to be negative and large in magnitude. A limitation of the omnigenic model is that neither the graph G nor the core genes used to calculate d_G, are specified. As such, one can only test a specific choice of d_G. Nonetheless, chapter 4 derives a prediction from this model – that most genes identified from *de novo* screens should be near to core genes – and seeks to assess this prediction across a wide range of reasonable choices of gene networks and core gene sets.

1.3 Conclusions

Gene co-expression networks provide an organizing framework for summarizing transcriptomic signatures of species, function, and disease; as well as for defining gene sets that contribute to genetic disease liability. The 15-year arc of literature has conclusively demonstrated that this approach identifies functional groups of genes that reflect an underlying biological phenomenon such as cell type, some of which show an overabundance of neuropsychiatric disease risk.

However, regional comparisons of co-expression have received comparatively little attention, leaving many fundamental questions unanswered. For instance, can regional cell

subtypes such as Purkinje, neurons, spiny neurons, or hormone-secreting neurons be identified by virtue of co-expression? Are neuronal co-expression patterns largely the same across regions of the brain? These questions motivate the broader goal of classifying co-expression signatures as shared across all regions of the brain, specific to a major region (such as cortex, or striatum), or specific to a sub-region (such as occipital cortex, or the putamen). There are also the matters of what these shared and specific co-expression patterns represent biologically and what their potential role might be in neuropsychiatric disease. The remainder of this thesis addresses this broader goal, and presents the construction and analysis of the first human brain co-expression atlas.

Chapter 2 presents my construction of the co-expression atlas, validation in external datasets, investigation of regional-specificity, and annotation of brain-wide and region-specific modules (including region-specific neuronal-subtype modules). In this chapter, I also use the co-expression network atlas to identify cell-type specific long non-coding RNA, and cell-type specific gene isoforms, identifying several genes that produce different main isoforms for different cell types.

In chapter 3, I identify several modules, including three brain-wide and two regional modules, which enrich for genetic risk and exhibit differential expression or co-expression in ASD brains. I link these modules to adult neurogenesis and neuronal maturation, as well as to activity-dependent neuronal processes such as bulk endocytosis. I also evaluate the regional specificity of nearly all previously-published co-expression or PPI modules that have been linked to neuropsychiatric disease, finding that these studies are largely identifying the same brain-wide neuronal co-expression signature.

Chapter 4 concerns network genetic architecture. In this chapter, I introduce the network-distance genetic architecture model to derive a simple prediction from the model from section 1.3c, and I evaluate this prediction across co-expression in adult brain, developing brain, fetal brain, and whole blood as well as in two other kinds of gene networks. I generalize network genetic architecture to treat the network itself as a parameter, and establish that co-expression networks better capture more genetic heritability than gene modules alone. I derive a statistical test to identify *trans*-QTLs with a consistent effect on disease risk – i.e. alleles that up-regulate risk genes, and down-regulate protective genes. Applying this to a collection of 8 GWAS studies, I identify numerous such instances, representing thousands of downstream protein-coding genes, and conclude that these observations are inconsistent with a core-gene only ("omnigenic") network architecture.

Chapter 2 The human brain co-expression network atlas

2.1 Abstract

Gene networks have proven their utility for elucidating transcriptome structure in the brain, yielding numerous biological insights. Most previous analyses have focused on a particular brain region, and the applicability of the gene relationships identified to other brain regions is rarely explored. By leveraging RNA-sequencing in 864 samples representing 12 brain regions in a cohort of 131 phenotypically normal individuals, I create a structured atlas of regional co-expression, and partition the brain transcriptome into 12 brain-wide, 114 region-specific, and 50 cross-regional co-expression modules. Nearly 40% of expressed genes fall into brain-wide modules, corresponding to major cell classes and conserved biological processes. Region-specific modules comprise 25% of expressed genes and correspond to region-specific cell types and processes, such as oxytocin signaling in the hypothalamus, or addiction pathways in the nucleus accumbens. I further leverage these modules to capture cell type specific lncRNA and gene isoforms, which contribute substantially to regional synaptic diversity, but remain difficult to ascertain in single-cell sequencing studies.

2.2 Introduction

The human brain is a highly-structured, complex organ, comprising hundreds of regional structures (Hawrylycz2015) and hundreds of billions of cells (Azevedo2009). These are themselves heterogeneous, with neurons alone demonstrating hundreds of spatially-heterogenous subtypes (Zeisel2018, Lake2017), leading to unique regional patterns of functional connectivity and cellular composition across regions. Yet the general structure, connectivity, and cellular distribution of the brain remains generally consistent across individuals. Cellular identity and regional activity is organized around functional groups of genes or pathways. Be they members of a protein complex, components of a signaling cascade, or a collection of critical genes converging on a biological process, genes within these groups must be co-regulated so as to be expressed at the appropriate levels to permit the group or pathway to function consistently (Felix2015).

To inform our understanding of molecular mechanisms in human brain, and their potential relevance to disease, I create an unbiased atlas of co-expression networks across 12 human brain regions (GTEx Consortium 2017). I demonstrate that the co-expression relationships defined in these networks are robustly identified using alternative network methods and orthogonal brain data sets. Combined with previous networks built from fetal brain across developmental time-points (Kang2011), these networks comprise a new resource for understanding the core transcriptional pathways, their time-points, and their spatial extents, underlying the patterning of the human brain. I use this resource to address several core biological questions. I show that brain co-expression in the brain is hierarchically organized into brain-wide, cross-region and region-specific signals, which reflect shared and area specific signals. Brain wide/cross-regional networks correspond to a major component of gene expression

that tags global signatures of cell types and biological processes. By combining differential expression and co-expression, I show that region-specific modules capture regionally-upregulated genes, and reflect regionally distinct cellular subtypes.

2.3 Results

2.3a Estimating and validating co-expression from brain RNA-seq data

Summaries of gene co-expression, such as co-expression networks (Zhang2005) or factor analysis (Schreiber2008), seek to identify transcriptional patterns (e.g. modules or gene weights) that summarize functional relationships between genes. As in any system, measurement noise degrades the power to identify gene relationships resulting in false-negatives, while confounders can give rise to spurious relationships leading to false-positives (Freytag2015). Recent advances in the processing of RNA-seq data have addressed removing sources of unwanted variation (Leek2012, Stegle2012, Gerstner2016) and hidden confounders from analysis (Mostafavi2013). The standard approach for brain RNA-seq co-expression has been to correct for known technical confounders, but not for latent factors. Therefore, prior to the main analysis of co-expression, I sought to identify which of the two approaches performed best. In order to perform this analysis, I needed to develop two methods: i) a method to perform model selection across multiple technical confounders and tissues, ii) a method to assess the improvement of signal-to-noise.

The GTEx dataset is annotated with comprehensive individual-level and sample-level information. Allowing for the possibility of tissue-by-covariate and covariate-by-covariate effects, there are more potential confounders than there are observed samples. Therefore, I needed to develop an approach to automatically select relevant technical variables that capture a high proportion of total variance. I combined multivariate adaptive regression splines (MARS;

Friedman1991, Milborrow2011) with variance partitioning (Hoffman2016) to identify and visualize technical covariates and interactions (up to degree 3) that explain a high proportion of total gene expression.

The primary objective of covariate correction is to improve the ratio of biological signal to technical and measurement noise by removing variance due to non-biological factors such as library complexity, sequencing batch, or cDNA conversion efficiency. This primary objective motivates a number of surrogate objectives for model selection, such as the number of detected eQTLs (Saha2017), differentially-expressed genes (Gerstner2016), or gene annotation prediction (Long2016), which have been used to compare methods and select method parameters. In addition, a little-used statistic, the integrative correlation coefficient (ICC), provides direct estimates of reproducibility (Cope2014), and I reasoned that the ICC of the data with itself would provide a coarse proxy for the signal-to-noise ratio.

Figures 2.1 and 2.2 highlight the key findings of this analysis. Firstly, although the model selection procedure identified relevant covariate-by-covariate interactions that explain roughly 5% of expression variance, tissue-by-covariate effects are smaller, so technical factors tend to impact expression in the same way across all regions. This is likely because regions were evenly matched across library preparation and sequencing batches. Secondly, replicate correlations between GTEx RNA-seq and microarray data from the same samples– a direct measurement of the signal-to-noise ratio – are positively related to ICC estimates, confirming that this statistic is a valid measure of reproducibility within the GTEx brain data. Thirdly, covariate correction improves the ICC estimate for most genes, and especially for those with reasonable (> 0.5) initial reproducibility, consistent with intuition. Fourthly, that covariate correction using MARS model selection results in larger improvements to ICC than the HCP factor model, an observation that

holds across a wide range of HCP parameters. I confirmed an attenuation of biological signal in latent factor models by showing that in every tissue, the average AUC for the ontology prediction task was higher following covariate correction than following HCP correction. To explain these observations, I hypothesized that factor models were removing a significant amount of biological variation in addition to the technical noise. To test this, I examined how the post-correction principal components loaded onto canonical cell type markers. I observed that following covariate correction, the largest component of expression variance showed very strong loading onto canonical neuronal markers, consistent with a biological explanation. This stands in strong contrast to the explicit assumption in latent factor models that the first few significant factors – those that explain the largest proportion of variance – are largely non-biological. Indeed, I found that when HCP was used for correction, this pattern of cellular heterogeneity was severely attenuated.

These results demonstrate that in well-balanced and well-annotated brain data, the co-expression signal-to-noise ratio is best optimized by correcting for technical covariates alone, and not by including additional latent factors. They have since been confirmed for other tissue types within the same study (Somekh2019), suggesting that these results hold not only for the brain but for any heterogeneous tissue. It should be not that, for cases when co-expression is itself a source of noise, such as for eQTL studies, these methods do provide a marked improvement.

While it is not directly related to the main argument of the thesis, I feel it important to report the direct estimates of signal-to-noise ratio from samples across the whole brain (median 1:2) and cortical regions (median 1:4) that arise from the GTEx replicate samples. These findings are confirmed in Fromer2016 which assessed a mean correlation of r=0.451 between PCR and

RNA-seq in cortex (SNR=1:4, fig. S3 in Fromer2016). Notably, the largest fold change observed between diseased and normal cortex in a large multi-disease study was 2 (Gandal2018a), which is small in comparison to the median amount of noise in cortical gene expression measurements. At this noise level and with GTEx sample sizes, one has 80% power to detect genes with a true expression correlation of ≥0.75. The imprecision of RNA measurement likely explains its limitation as a diagnostic tool for complex disease, the challenge in identifying clear disease-specific expression endo-phenotypes, and the difficulty of reproducing specific gene-gene correlations. It also partly explains why observed case/control differences in chromatin and histone states only weakly (if at all) translate to measurable differences in expression. This probably remains the largest challenge: without more precise methods of measuring RNA expression, significantly larger sample sizes will be needed to thorougly identify eQTLs and functional signatures. It is a largely unappreciated feature of multivariate analysis that principal factors – and thus gross correlational structure – can be identified from small sample sizes. It is for this reason that network analysis and factor analysis of co-expression produces gene sets and gene scores that are fairly consistent across brain datasets in the face of low signal-to-noise of their individual components.

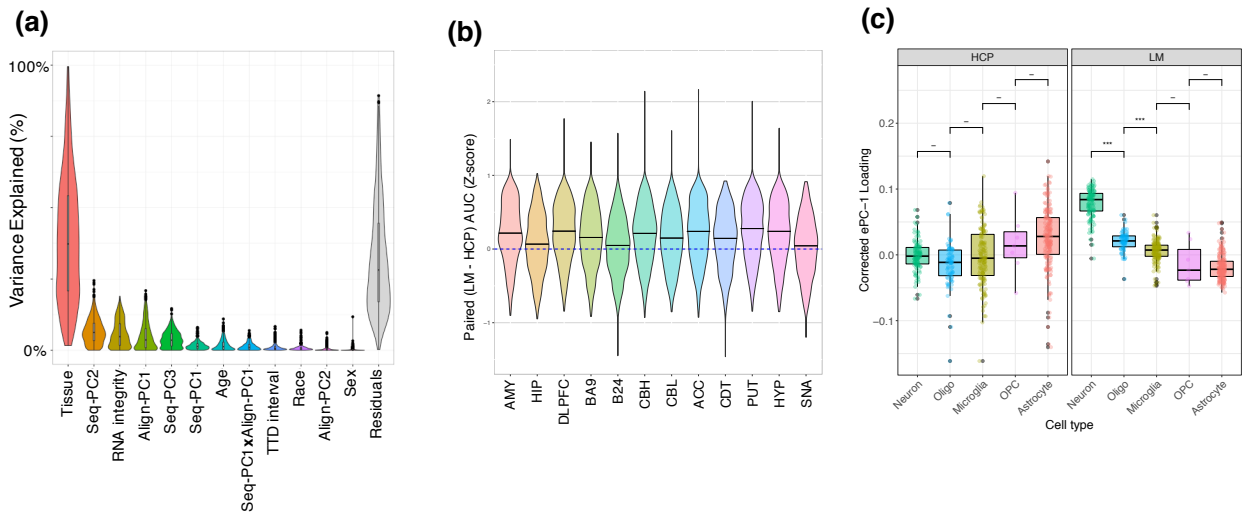

Figure 2.1 Technical covariate effects and comparison of correction methods

(a) Violin plot of % variance explained, per gene, of technical covariates and tissues, for the top factors identified, plus age, race, and sex as comparisons. Differences in mean expression across tissue drive the variance. No significant tissue x covariate interactions are identified. **(b)** Scaled differences in GO category prediction within each of the tissues; values > 0 imply that networks built from LM-corrected data are more accurate in the prediction task than networks built from HCP-corrected data. **(c)** Boxplot of top post-correction principal component loadings onto cell-type marker genes. Linear model correction maintains neuron/glia differences as the primary driver of expression, while HCP appears to lose this signal.

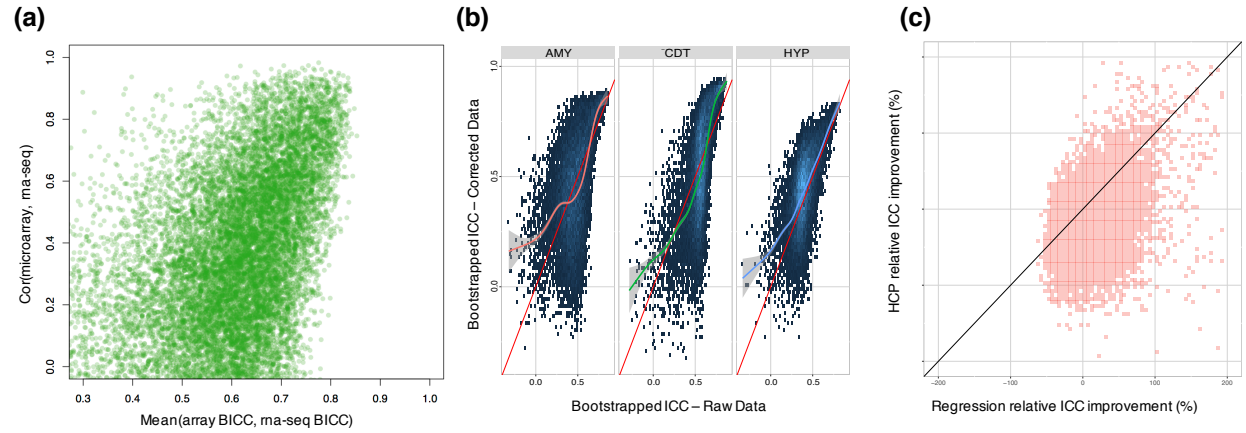

Figure 2.2 Signal-to-noise ratios for correction methods based on bootstrapped ICC

(a) Relationship between cross-platform correlation and average ICC. Each point is a gene, measured on the same samples across two platforms. X-axis: the average of the two BICC values within each platform; Y-axis: the cross-platform correlation for the gene, with higher within-platform BICC implying stronger cross-platform correlations; thus lower noise. **(b)** Bootstrapped ICC values or raw and LM corrected data. Higher density above the red y=x line implies improved signal-to-noise ratios after correction. **(c)** Comparison in hypothalamus between the relative ICC improvement for regression (x-axis) and HCP (y-axis). Values below the y=x line implies a stronger improvement in signal-to-noise ratio using regression to correct for covariates.

2.3b Identifying and verifying specific and shared network modules

I next asked how to structure a network analysis to incorporate multiple regions of the human brain. I recognized that the tissue hierarchy formed from average gene expression profiles corresponded exactly to the regional map, grouping individual regions into their higher-order structures. I therefore reasoned that this hierarchical structure could be used as a backbone for defining consensus gene relationships. For each tissue, I generated a bootstrapped-resampling version of weighted gene co-expression network analysis WGCNA (robust WGCNA; Langfelder2008), which reduces the impact of sample outliers. I then merged the resulting co-expression networks, forming consensus networks for each split of the tissue hierarchy, and arranged these regional co-expression networks into 20 hierarchical expression categories: 12 brain region specific categories (corresponding to each sampled region), 7 multi-regional categories (corresponding to multiple, structurally-linked regions), and a brain-wide category. Analyses of the networks within these categories identified 311 total modules, 199 from the base regions, and 112 from the hierarchical consensus regions. Of the 199 tissue-level modules, 173 (87%) replicate with strong support in at least one other expression dataset in matched tissue.

I asked how to group modules together that represent the same transcriptional program across multiple regions. For each pair of modules, I computed two similarity scores: the Jaccard similarity – the fraction of genes common to both modules – and the eigengene similarity – the correlation of module eigengenes – and formed a weighted average. Using this metric, I grouped the 311 modules hierarchically into *module sets*. The modules in a module set are by construction highly similar in terms of gene overlap and expression, and likely represent the same underlying co-expression relationship.

Figures 2.3 and 2.4 summarize this approach and demonstrate that the resulting module sets are *coherent*: module sets that contain a whole-brain module also contain modules from the other major brain structures, and show evidence for the co-expression relationships in the remaining regions. Similarly, module sets that contain a cerebellum consensus module also contain modules from the cerebellar body and cerebellar vermis, and tend to show weak to no evidence outside of the cerebellum. While the approach is not perfect – for instance Cerebellum-M1 shows strong evidence in regions outside the cerebellum – the majority of region-specific module sets only show evidence within that region.

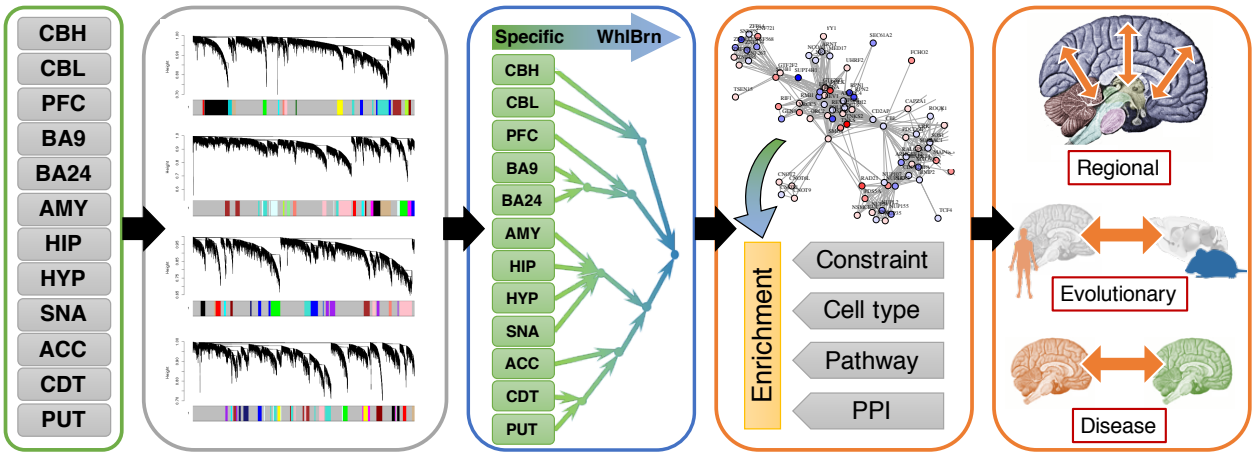

Figure 2.3 Overview of the brain expression atlas

To identify cross-regional and region-specific co-expression patterns in the human brain, 12 co-expression networks are built from base tissues. These are merged hierarchically according to regional expression similarity, resulting in co-expression networks at every level of the hierarchy. These can be investigated for the biological systems they represent through enrichment analysis, and used as a basis of comparison for regional, evolutionary, and disease comparisons.

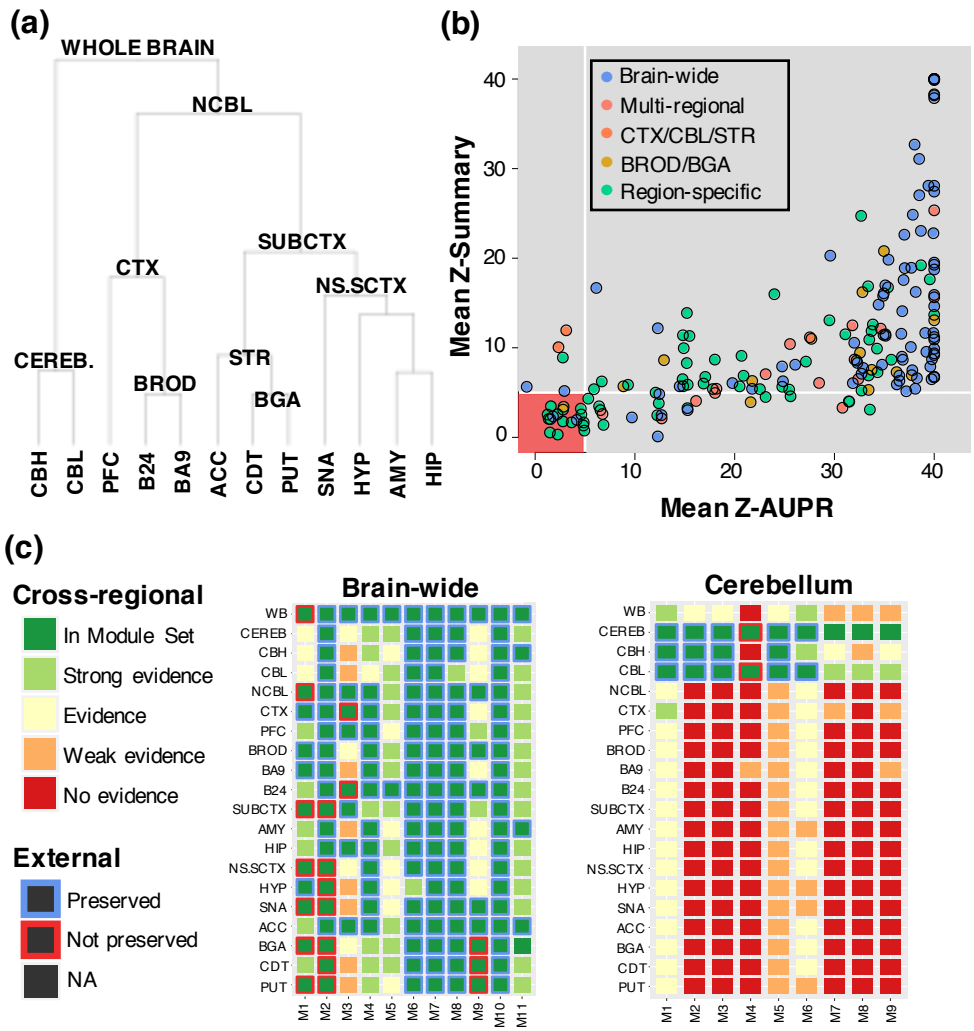

Figure 2.4 Module set definitions, replication in other datasets, and regional specificity

(a) Definition of regional and multi-regional module sets by position in the expression hierarchy. **(b)** Module-level preservation in other datasets, with the bulk of modules showing very strong reservation (Z>5) for both AUPR or the classical Z-summary. Modules not showing preservation tend to be from regions with limited microarray data (e.g. substantia nigra). **(c)** Module set evidence across all regions of the brain for brain-wide and cerebellar modules (Strong: Z>8, evidence; Z>5, weak evidence Z>3, no evidence Z<3). In module set: A module was identified that was identified as part of the module set using jaccard and kME similarity.

2.3c Methodology has little impact on identified region-level and whole-brain modules

More than a dozen algorithms have been developed for co-expression network construction and gene clustering over the past decade. Network-based methods – such as CAST (Ben-Dor1999), WGCNA (Zhang2005), ARACNe (Margolin2006), ARACNe-AP (Lachmann2016), GLASSO (Simon2013), MEGENA (Song2015), QCut (Ruan2010), RMT-threshold (Luo2007), or Bayesian Networks (Myers2009) – differ in their correlation measures and the topological constraints they place on the graph; while direct clustering methods –such as QUBIC (Li2009), convex biclustering (Chi2016), EPGMM (McNicholas2010), or simple k-means clustering – differ predominantly in their distance metric and cost function. The choice of algorithm may have a large potential impact on the ultimate results, and thus it is important to determine how much methodological choices effect the final results.

I therefore sought to establish that these various choices have little meaningful impact on the modules identified by my analysis at either the regional or whole-brain scale. To span the space of network and non-network models, I re-computed gene networks using algorithms that use a different metric of gene correspondence, ARACNe (mutual information) and GLASSO (partial correlation). I also implemented a Bayesian von-Mises-Fisher mixture model as a non-network probabilistic approach for gene clustering (**methods**). High overlaps between the modules resulting from these methods would indicate that our results are largely independent of methodology – reflecting a strong underlying signal discoverable by multiple approaches.

Figure 2.5 shows that this is indeed the case, with each WGCNA module corresponding to one or more modules from other methods – but in no case identifying a gene cluster which was not identified by at least one other method. The largest differences are in module size, with WGCNA tending to have the largest modules with other methods identifying only the most

confidently-clustered genes. The average co-clustering accuracy – the proportion of gene pairs placed in the same clusters across methods – exceeds 60%; and it can be observed from the overlaps that most modules identified by the other methods are wholly or nearly-wholly contained in a WGCNA module.

I next applied a down-sampling approach to estimate overall power for module and hub detection (**methods**), in order to estimate that WGCNA has power to comprehensively place all module hub genes into a cluster (i.e., not background "grey" genes). Further, module co-members can be identified with a lower, but reasonable precision (60%) recall (45%) and accuracy (55%), which increases to >90% for modules that are well-separated.

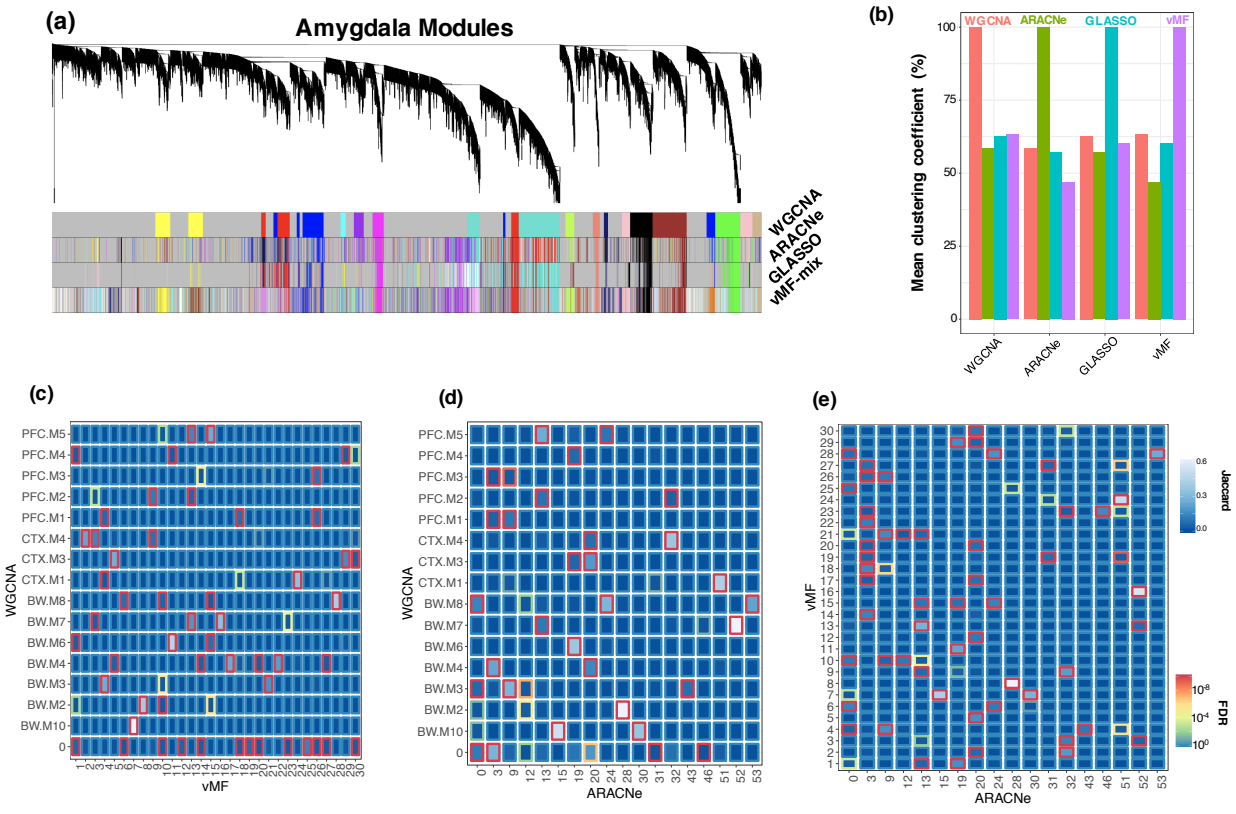

Figure 2.5 Comparison of co-expression network inference methods at the regional level

(a) Topological overlap dissimilarity dendrogram, with genes represented as colored vertical bars, colors corresponding to assigned network modules for WGCNA, ARACNe, GLASSO, and vMF-mix. **(b)** Pairwise co-clustering coefficients; X-axis represents which module is taken as 'reference' module for the co-clustring. Notably WGCNA shows >50% clustering coefficient with all methods. **(c-e)** Pairwise module overlaps, colored by the jaccard metric, and outlined by significance, demonstrating that a large number of modules overlap directly as one-to-one, or one-to-many fashion; and few modules are non-overlapping.

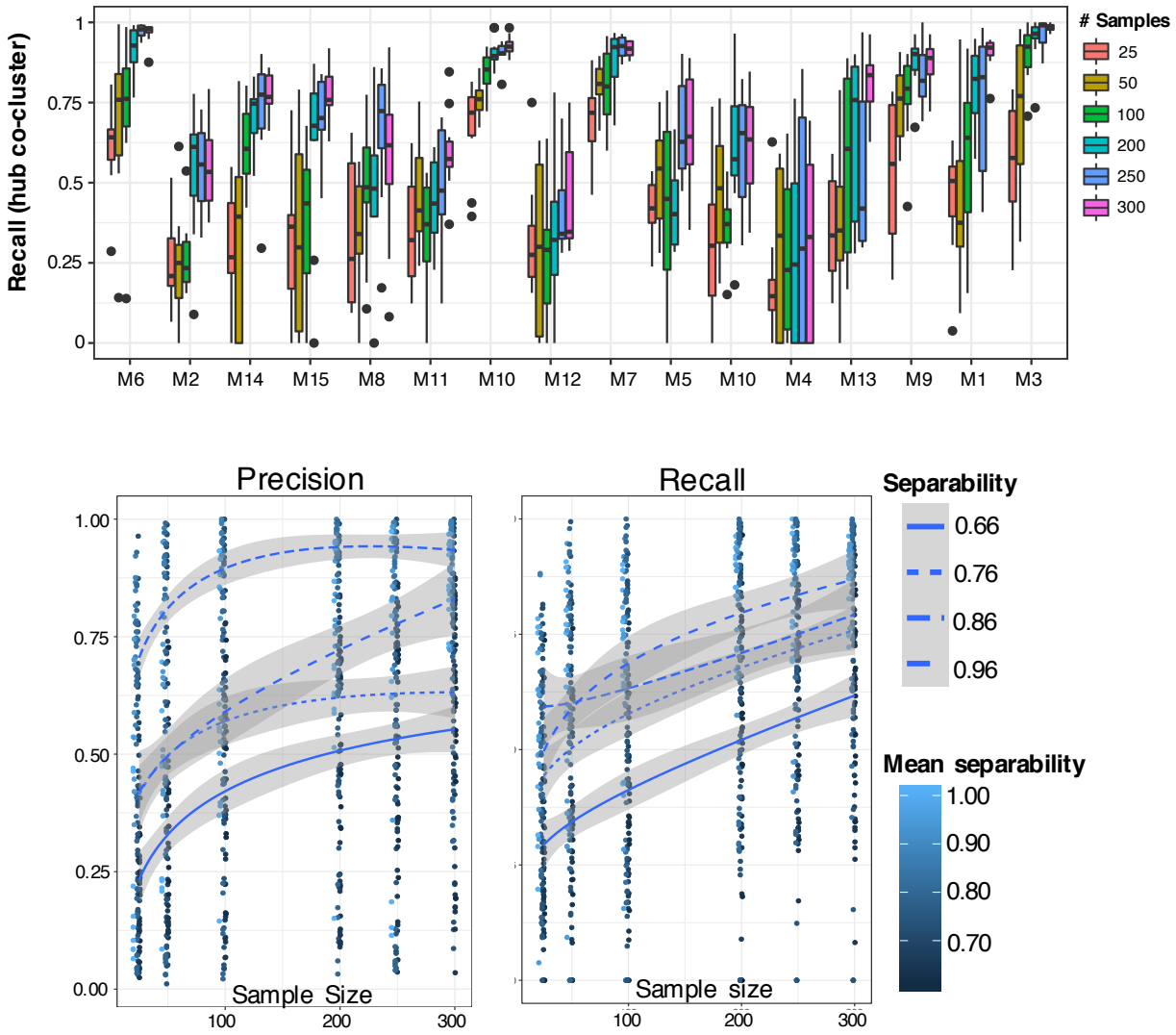

Figure 2.6 Power to detect modules as a function of separability and sample size

Top: Module accuracy, as measured by the faction of times a gene is co-clustered with its hub gene, as a function of sample size. Relationship is monotonic increasing, though the saturation point differs from module to module; this is a function of how separable the module is from the other modules. *Bottom*: Plot of cluster precision and recall as a function both of sample size and module separability. Separability is inversely proportional to the eigengene correlation between two modules, and low overall separability is indicative that the module correlates partly with at least one other module, leading to partly stochastic assignment of genes to modules. For instance, highly-distinct cell types such as microglia or distinct processes such as ribosomal turnover, correspond to well-separated modules (M2, M10; see figure 2.7), while the subtly-different neuronal subtypes generate modules with mutually-overlapping relationships (M3, M4, M5).

I then sought to investigate the robustness of brain-wide networks to methodology, to establish that the whole-brain networks would be identified by an orthogonal whole-brain approach. To do this, I leveraged the fact that all the samples from the various regions were drawn from the same set of brains, and performed a simple analysis of feature extraction through tensor decomposition, followed by dimensionality reduction on the genes via t-SNE (**methods**). This resulted in data with clear regions of high-density, but not clearly-defined clusters. As such, I identified these high-density regions as gene modules by applying DBSCAN. This resulted a set of clusters that strongly overlapped with my original whole-brain clusters (figure 2.7) despite the approaches sharing no common methodology.

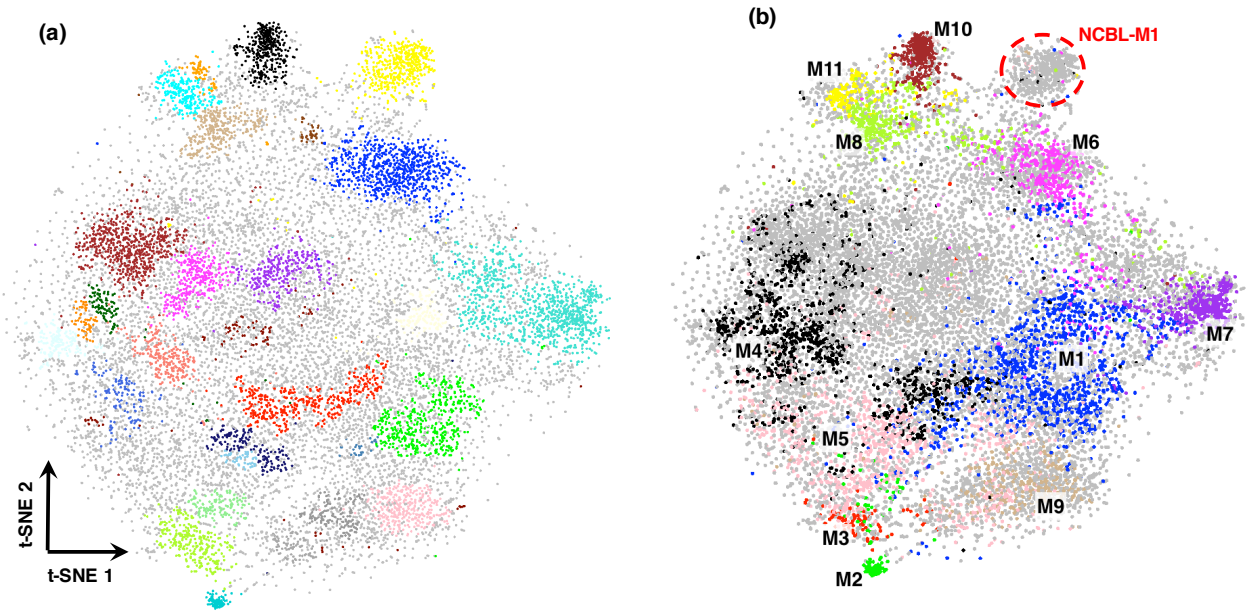

Figure 2.7 Comparison of brain-wide WGCNA modules to tensor-decomposition-based modules

Plot of the tSNE embedding built from tensor decomposition. Each point represents a gene, and is colored by the DBSCAN-assigned module **(a)** or the WGCNA whole-brain module **(b)**, demonstrating that several modules (M2, M6, M7, M8, M10, M11) have direct overlaps, while neuronal modules M3, M4, and M5 correspond to several DBSCAN modules each. The obviously missing cluster (yellow) corresponds to NCBL-M1 which has no presence in cerebellar tissue, illustrating one potential drawback of tensor decomposition (strong multi-regional signals can still appear to brain-wide).

While the above methods can readily define co-expression modules at the single-region level or the whole-brain level, the hierarchical approach I initially took is the only method that conveniently and efficiently produces consensuses for the intermediate levels. While alternatives could be evaluated for this approach – for instance using a hierarchical Bayesian model to estimate covariance matrices or von-Mises-Fisher parameters – I reasoned that validation at the root and leaves of the region tree implies validation for all subtrees. I thus concluded that, at the current sample sizes, the modules identified by various methods did not vary to a practical degre.

More speculatively, it seems that without the (biologically unwarranted) assumption of sparsity, the set of brain-expressed genes will not fall into discrete clusters. Instead, the tensor-

decomposition results make a fairly good case that genes can be placed relative to one another in a latent, regulatory space. Different biological pathways or functional classes should be distributed on different – but not exclusive – volumes of that space. Pleiotropic genes should inhabit regions of functional overlap, while unifunctional genes should not. If this hypothesis holds, then there is far more information in a co-expression network than in the modules. The last chapter of this thesis begins to explore this topic in just such a direction.

The tensor-decomposition approach merits additional analysis in the near future. The GTEx dataset is one of the few with expression measured in multiple regions of the same brain, and so can be arranged into a 3d-tensor. Decomposition then enables genes, regions, and individuals to be linked. The regional pattern can be examined for specificity, the gene loadings for ontology, and the individual loadings for a genotype or phenotype relationship (e.g. QTL or diagnosis). Because many new datasets use single-cell or single-nucleus sequencing from multiple individuals – which can then (by averaging within cell types) be arranged into individual×gene×type – this approach may have a wide application.

2.3d Hierarchical networks elucidate sources of regional and global brain co-expression

Cell type composition is a major driver of measured RNA expression in tissue (McKenzie2018, Kelley2018). I therefore hypothesized that whole-brain co-expression modules represent major cell classes (neurons, astrocytes, microglia, oligodendrocytes), and that multi-regional or regional modules represent specific cell subtypes. Using markers for primary brain cell types (Lein2006, Zhang2014) and cell subtypes (Heintz2004), I found that that modules at all levels are consistent with this hypothesis. Figure 2.8 presents the results of cell type enrichment and decomposition for whole brain and regional modules. Canonical markers for major cell types fall in a near one-to-one correspondence with whole brain modules: neuronal markers in M4, oligodendrocyte markers in M7, astrocyte markers in M6, microglial markers in M10, and endothelial markers in M11. Pathway analysis and analysis of cell states revealed additional roles of modules: M2 corresponds to components of the ribosome and translational machinery, M8 contains markers of reactive gliosis in astrocytes and microglia activation, and M1 enriches for markers of neural progenitor cells and neuronal maturation.

Because neurons span highly diverse cell populations in the brain, I then asked whether known region-specific cell types such as medium spiny neurons in the striatum or Purkinje cells in the cerebellum correspond to region-specific co-expression modules. Using published single-cell sequencing from human cortex (Lake2016), cerebellum (Lake2018), and mouse striatum (Zeisel2018), I identified three region-specific modules – BROD-M8, CEREB-M2, and STR-M1 – corresponding to interneurons, Purkinje neurons, and medium spiny neurons respectively. Figure 2.8 describes this relationship by showing the genes more central in the module (high

kME) have high relative expression in these cell types; and this relationship decays exponentially with module membership.

Brain-expressed genes, as a class, are known to be highly intolerant to loss-of-function mutations (LoFs). I question whether this intolerance was a broad feature across brain cell types, or whether this intolerance was a property of genes upregulated in a particular cell type. To address this, I evaluated whether genes with strong LoF intolerance scores (Petrovski2013, Lek2016), fall disproportionately into cell-type specific modules. Figure 2.9 shows that the neural progenitor module BW-M1, neuronal module BW-M4, and two un-annotated modules BW-M3 and BW-M5 all enrich significantly for LoF-intolerant genes, as well as the neuronal subtype modules BROD-M8, CEREB-M2, and STR-M1. Noting that, by construction, these modules cannot extensively share genes, I reasoned that this may indicate that LoF-intolerance may itself be indicative of neuronal modules. To test this, I used a set of ranked computationally-derived neuronal markers (Kelley2018) to extend cell-type annotations to high-confidence non-classical marker genes. Consistent with the correspondence between neuronal genes and LoF intolerance, both BW-M3 and BW-M5 significantly enrich for the top 300 to 500 neuronal markers, suggesting that they capture neuronal genes with a lesser degree of upregulation, lower overall expression (and thus noise), or processes that are not always co-expressed in all neurons. The ontologies for these modules reflect this possibility, as both module sets enrich for the Huntington's, Alzheimer's, and Parkinson's disease pathways, and for respiratory functions (BW-M5 enriches for mitochondrial respiratory chain I and BW-M3 for ATP metabolism, **figure 2.9**).

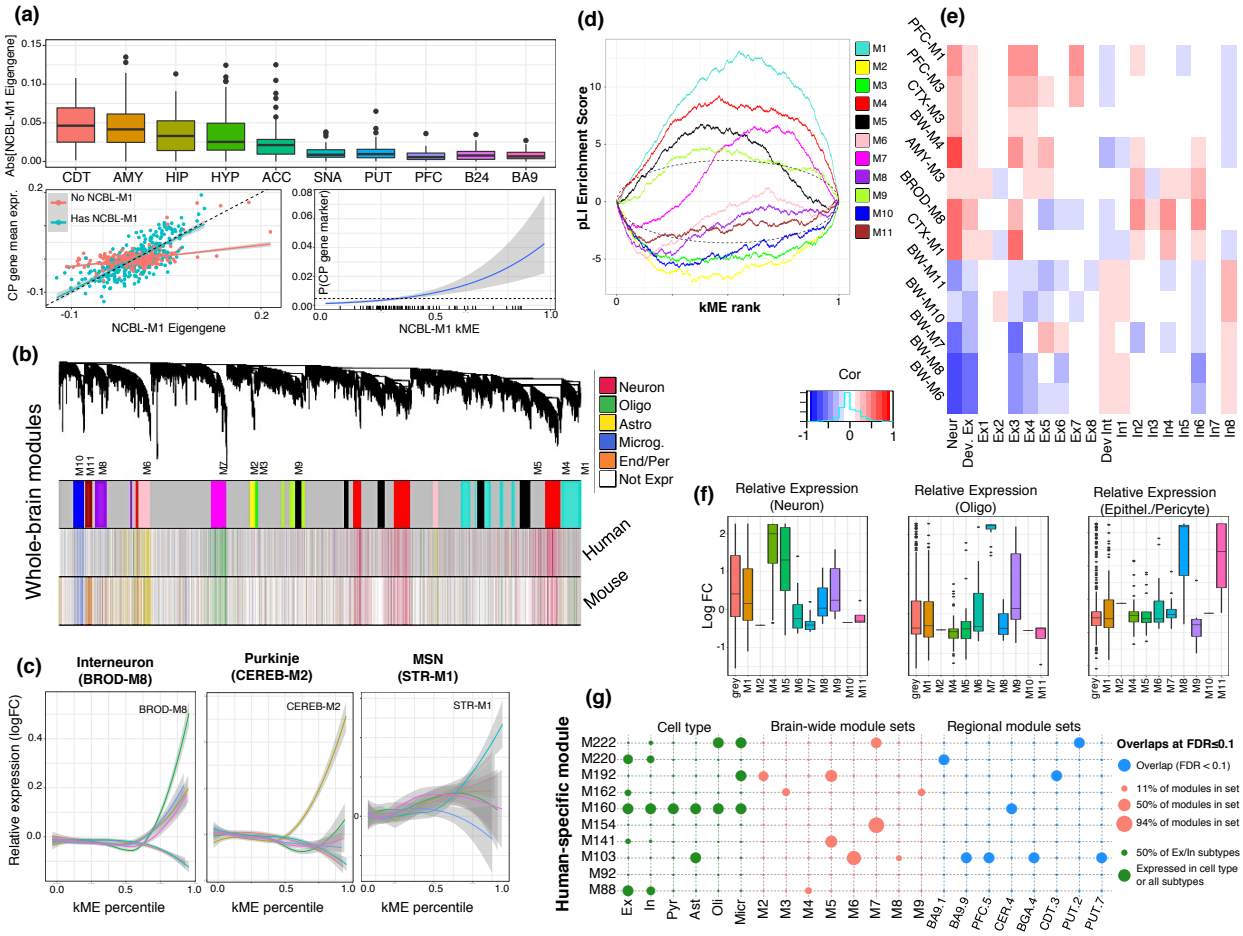

Figure 2.8: Cell-type heterogeneity relates to co-expression modules, gene intolerance, and evolution.

(a) *top*: Absolute value of the eigengene of module NCBL.M1 plotted across regions, showing higher variance in regions adjacent to or accessible through ventricles. *left*: Expression of NCBL.M1 eigengene and mean expression of choroid-plexus marker genes in regions with and without an NCBL-M1 module. *right*: Marginal probability of a gene being a choroid plexus marker, as a function of NCBL-M1 soft membership. **(b)** WGCNA dendrogram at the whole-brain level, colored by module (top), and canonical marker genes for major cell types in human and mouse. **(c)** Relative expression of neuronal marker genes for modules BW-M4, BROD-M8, CEREB-M2, and STR-M1 within interneurons from cortical SC-sequencing, Purkinje neurons from cerebellar SC-seq, and medium spiny neurons from striatal SC-seq, as a function of module kME. **(d)** GSEA enrichment plots for LoF-intolerant genes (pLI > 0.9) for all whole-brain modules. **(e)** Factorization-based decomposition of bulk expression (methods). Correlations for BW, CTX, and PFC modules come from decomposing DLPFC bulk expression; AMY from decomposing AMY bulk expression, and BROD from decomposing of B24 bulk expression. **(f)** lncRNA relative expression in single-cell data, grouped by the imputed module in riboZero data from BA9. **(g)** Cell type expression and significant module overlaps for human differentially-expressed modules from Sousa *et al.* (Sousa2017). Cell type assignments are as given in that publication.

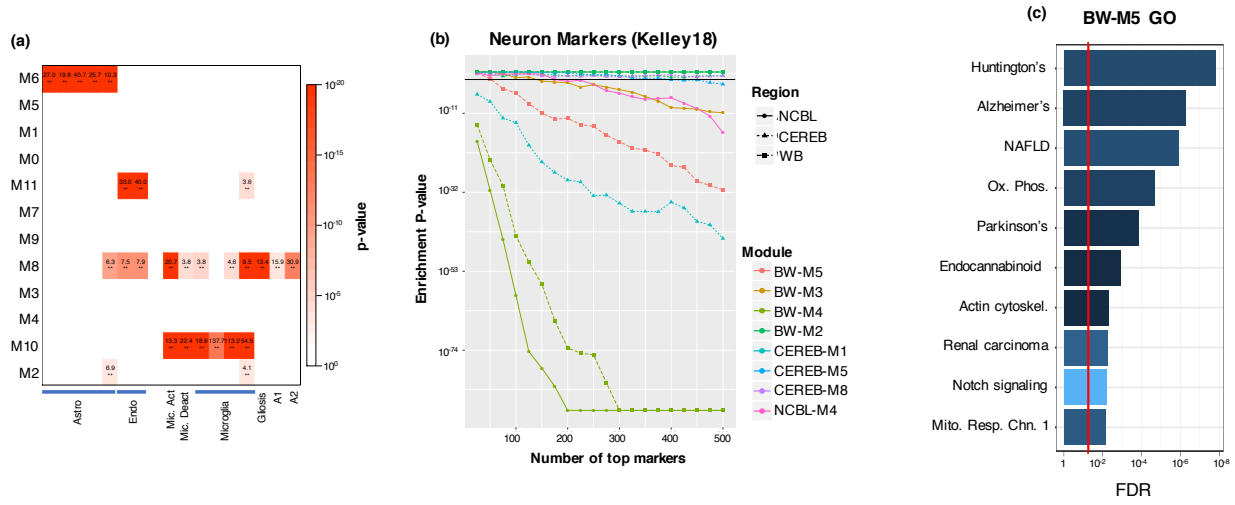

Figure 2.9 Glial activation and high-fidelity neuronal markers, related to figure 2.3.4

(a) Extended glial cell type enrichment for brain-wide modules; establishing BW-M8 as representing microglial activation, reactive gliosis, markers for both A1 (neurotoxic, microglia-induced) and A2 (neuroprotective, damage-indussed) astrocytes (Liddelow *et al.* 2017). **(b)** Plots of the marginal proportion of loss-of-function intolerant genes as a function of soft module membership for modules BW-M1 (most enriched) and BW-M2 (most depleted). **(c)** Gene ontology enrichments for module BW-M5.

Recent comparative expression studies have identified thousands of genes up-regulated in humans compared to non-human primates, and have implicated spatial differences in neuronal subtypes and neurotransmitter receptors in driving this divergence (Konopka2012, Sousa2017). Sousa *et al.* (Sousa2017) examine differential expression between human and non-human primates within 16 brain regions. They build modules across all species and regions together, identifying 27 modules showing differences between species (but not region), 37 modules showing differences between both species and region, and an additional 12 showing region-specific differences between species. Of these 76 modules, 10 show human-specific changes (that is, human does not match macaque or chimp). This approach, which has been referred to as "differential patterning" (Parikshak2016), may fail to capture differences in co-regulation across species. In other words, the human-specific modules identified by Sousa *et al.* should reflect brain-wide rather than region-specific changes in cell type composition. To test this hypothesis, I examined the overlap of the human-specific Sousa *et al.* modules with all modules in the human brain co-expression atlas, and found that 7/10 of human-specific modules overlap whole-brain modules, while only two modules (M160, M220) overlap regional modules alone (**figure 2.8g**). This result suggests that Sousa *et al.*, by virtue of both sample size and approach, may be underpowered to identify inter-species differences in regional co-expression.

2.3e Cell-type-specific lncRNA and isoforms in the human brain

Long non-coding RNA (lncRNA) are a diverse set of RNA species that modulate gene expression or protein function (Wei2018) across many CNS cell types (Chen2019), and several studies suggest that lncRNA dysregulation is a component of neuropsychiatric disease (Corgill2018, Parikshak2016, Ang2019, Zuo2016). Many brain-expressed lncRNA have roles in neurodevelopment (Clark2018), and the enhancer with the most accelerated substitution rate in the human genome, HAR1, modulates the expression of a neuronally-expressed lncRNAs now termed *HAR1A* and *HAR1B* (Pollard2006). Since lncRNA as a class tend to be expressed at a lower level than protein-coding RNA (Djebali2012), they tend to be difficult to profile and annotate through single-cell sequencing. Having identified co-expression networks corresponding to the major CNS cell types, I reasoned that they could be used to annotate human brain-expressed lncRNA in an untargeted, transcriptome-wide manner, in order to associate lncRNA with neurological cell types and processes.

Only 52 known lncRNA species were profiled in the initial GTEx data set, likely because GTEx used poly-A selection. Therefore, I expand the set of profiled lncRNA by projecting the whole-brain modules into non-polyA-selected data from 44 neuro-typical post-mortem brains (Parikshak2016) in which the whole-brain and cortical modules were preserved (preservation Z = 3 to 30). Using gradient boosted trees (Chen2016) to learn expression signatures of our module assignments in the new dataset and then classify lncRNA into the appropriate modules (**methods**), I identified 286 lncRNA belonging to major cell types and processes, the majority of which associate with neuronal module BW-M4 (66) or NPC module BW-M1 (109). Remarkably, slightly more than 20% (61/286) of these cell-type specific lncRNAs were previously shown to be dysregulated in neuropsychiatric disease (Gandal2018b). I cross-

referenced the inferred modules with published single-cell hippocampal and cortical RNA-seq (Habib2017), and validated that single-cell-expressed lncRNA belonging to the assigned cell-type modules are up-regulated within those cell types (**figure 2.8f**).

A previous study of ASD differential expression highlighted the differential expression of lncRNA as an integral component of the ASD transcriptomic signature (Parikshak2016). Since lncRNA do not encode proteins, I reasoned that the lncRNA signature might reflect different cell types from the protein coding signature. Therefore, I removed a set of protein coding genes – matched on length, mean-expression, and GC-content to the lncRNA – from the training set, and imputed modules for both lncRNA and matched protein-coding genes. I found that there were significant case-control differences in both expression and connectivity for modules M1, M6, and M8 ($p < 10^{-15}$, KS-test). These differences – which imply dysregulation in neurogenesis, astrocytes, and reactive glia – are mirrored in the matched protein-coding genes, confirming that the lncRNA signature is aligned with the overall ASD signature (**figure 2.10**).

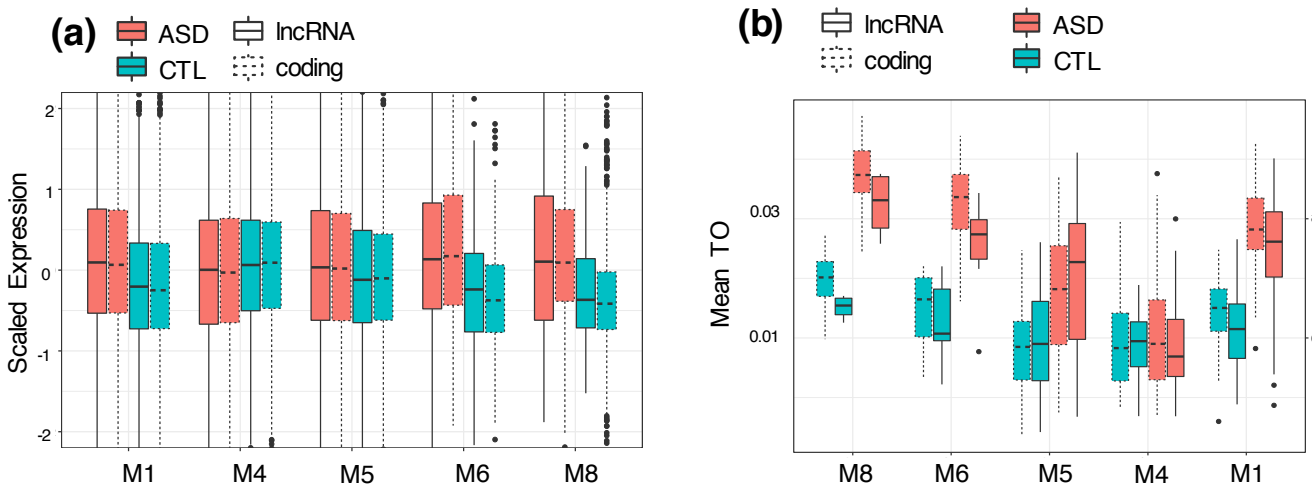

Figure 2.10 - lncRNA co-expression differences in brain-wide modules between cases and controls

(a) Boxplot of scaled gene expression in cases and controls across 5 major whole-brain modules, showing significant difference in mean expression or modules M1, M6, and M8. **(b)** As in (a), but gene mean topological overlap – a readout of co-expression – is plotted, recapitulating the trends from (a), suggesting that the expression and co-expression of lncRNA are disrupted in ASD; but not any differently from matched protein-coding genes.

Since single cell data does not yet provide similar isoform level coverage to bulk data, there are few known cell-type-specific isoforms. Given the successful annotation of lncRNAs using co-expression, I therefore sought to integrate isoform-level expression within cell type modules, both to understand cell-type specific splicing and to identify those specific isoforms likely involved in ontological pathways (**figure 2.11**). I hypothesized that isoforms whose expression showed high correlation with a module eigengene (isoform kME) are likely an integral component of that co-expression network. By thresholding on a kME > 0.55, I identified 3,764 isoforms showing specificity to major cell types (**methods**). To validate these findings, I obtained RNA-sequencing data from sorted cells (Zhang2015), quantified expression at the isoform-level, and ranked isoforms expression within cell types (**methods**). Consistent with my hypothesis, I observed a very strong correlation between isoform kME (to a cell-type module) and the rank of that gene's expression in the sorted cell data (Spearman's rho = 0.286 oligo 0.258 astro, $p < 10^{-15}$ for both).

Genes that show differential splicing between cell types are of particular interest, as they reveal protein domains with cell-specific roles, and potentially cell-specific protein binding partners. I reasoned that isoform-switch genes could be identified by finding genes with at least one daughter isoform assigned to a different module from her parent gene. Figure 2.9e shows that these occurrences are rare, with fewer than 1% of multi-isoform cell-type-related genes showing isoform switching between cell types. Using the same sorted-cell data from the previous section, I validated these cell-type isoform-switch genes. Figure 2.9f shows two examples of genes – *ANK2* and *SCP2* – with astrocyte/neuron switching, and demonstrates that the cell type with strong module kME is the cell type in which the isoform is up-regulated (**figure 2.10)**.

Noticing that *ANK2* and *SCP2* have both been previously implicated in ASD, I reasoned that, though isoform switching genes appear to be rare, the improper regulation of neuron/glia isoform switching may contribute to ASD. By cross-referencing the 11 neuron/glia isoform-switch genes (one gene is counted twice, as it has a daughter isoform in astrocytes and oligodendrocytes), I identified 4 genes with a AutDB (Basu2009) score of 4 or higher: *ERGIC3* (4), *PDE4DIP* (4), *SCP2* (9), *ANK2* (9). While this is a small number of genes, the overlap is significant ($p < 0.01$, Fisher Exact Test). Notably, *ANK2* and *SCP3* show differential splicing of at least one event in ASD vs CTL brains (FDR < 0.05, linear mixed-effects model), but *ERGIC3* and *PDE4DIP* do not. Gandal2018b identifies isoform switching in *ANK2* as differentiating between SCZ and ASD, and that this collection of isoform-switch genes between disorders is significantly enriched for syndromic ASD genes. However, the differentially spliced events in *ANK2* (both ASD/CTL and ASD/SCZ) do not match the primary difference between neuron and astrocyte transcripts: the inclusion of the 2,085 amino acid "giant exon." These observations suggest that while there may be a role for the disruption cell-type specific alternative splicing within ASD, it is likely more complicated than the up-regulation of certain glial isoforms within neurons.

Because regional co-expression networks correspond to regional cell types, I reasoned that the above approach could be used to build putative cell-specific isoform maps for D1/D2 medium spiny neurons, Purkinje cells, basket cells, and inhibitory neurons, all of which show regional specificity in the data. Repeating the above analysis, I identified between 300 and 500 putative cell-type-specific isoforms, finding that very few (< 5%) of the resulting isoforms show high kME to the broad neuronal module BW-M4, consistent with the interpretation that these marker isoforms are cell-type specific (**figure 2.9c**). All isoforms enrich for synapse-related

functions, with additional regional variability: MSN isoforms enrich for the oxytocin signaling pathway, consistent with their role as downstream targets of oxytocin in the nucleus accumbens, whereas Purkinje isoforms enrich for AMPA receptor regulators, and inhibitory neuron isoforms enrich for the NMDA receptor activity (**figure 2.9d**). These results demonstrate how the atlas networks can be used to identify cell-type-specific splicing differences from bulk expression data, and highlight the synapse as a nexus of gene regulatory complexity and isoform heterogeneity. Further, they provide additional evidence for the importance of isoform-level analysis compared with gene expression alone in defining cell-type specific transcriptomes.

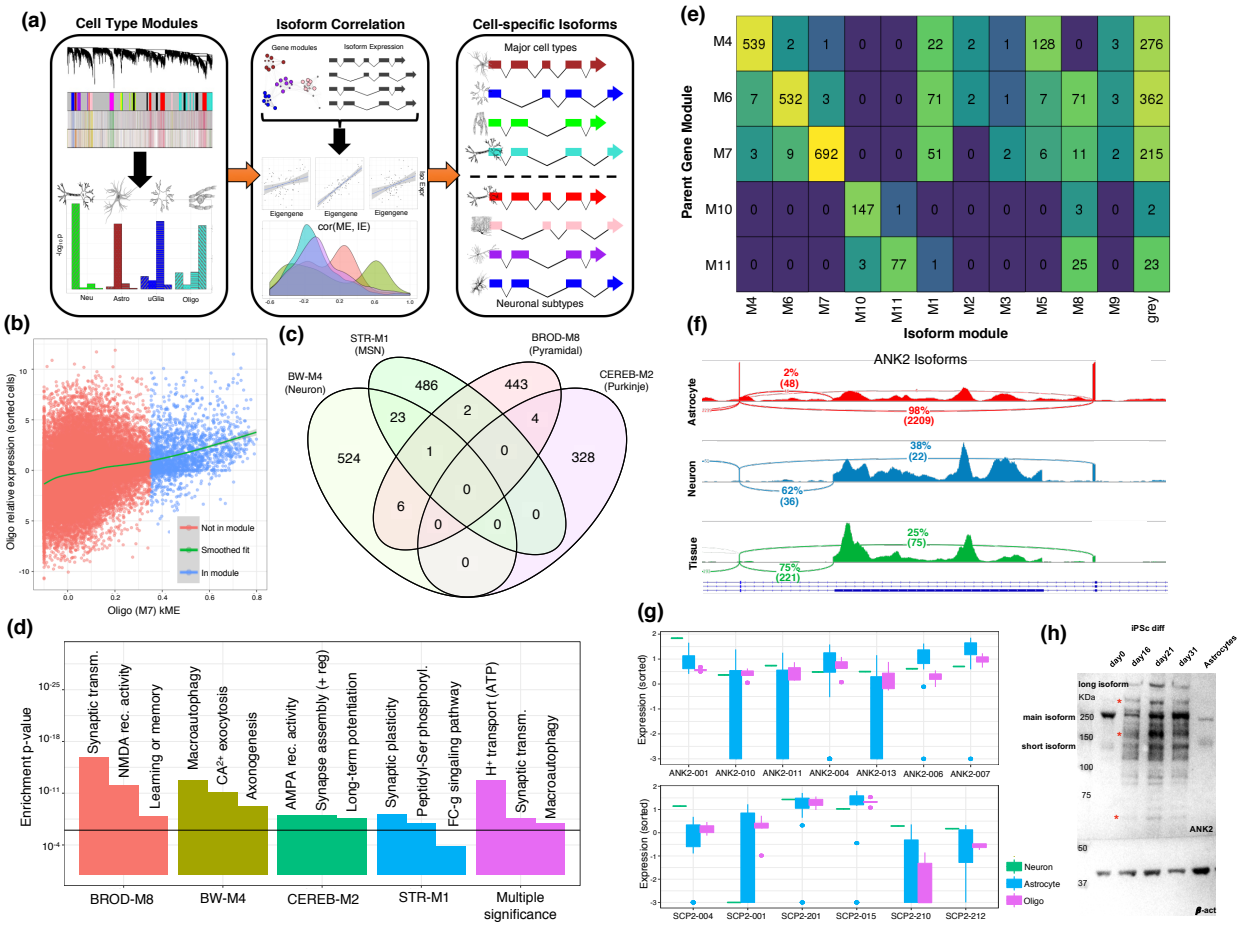

Figure 2.11: Cell-type-specific isoforms reflect receptor heterogeneity

(a) Overview of isoform assignment on the basis of kME to cell-type modules. **(b)** Isoform relative expression (log-FC of TPM) in oligodendrocytes plotted against isoform kME to BW.M7 showing significant positive relationship ($p<10\text{-}6$, linear regression). **(c)** Venn diagram of isoforms assigned to neuronal subtypes **(d)** GO enrichment of parent genes of subtype-specific isoforms. Top module-specific terms are shown, followed by terms which are significant across multiple subtypes (min p-value shown). **(e)** Assignment of daughter isoforms of genes with membership to a whole-brain cell type module, showing that most daughter isoforms are either assigned to the parent gene module, or to the grey (unclustered) module. **(f)** IGV visualization of the event differentiating the astrocyte and neuron isoforms of ANK2, the inclusion of the giant exon, in sorted cell data. **(g)** Expression of ANK2 and SCP2 transcripts in sorted-cell data, showing isoform switching between neuron and astrocyte. **(h)** Western blot of ANK2 across iPSC differentiation into neurons, and within astrocytes, demonstrating the presence of a long isoform specific to neurons.

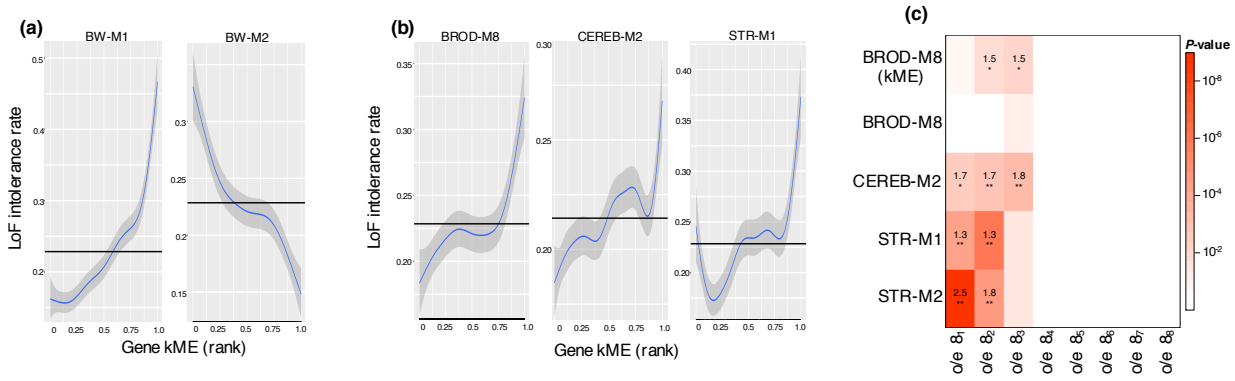

Figure 2.12 pLI enrichments for brain-wide and neuronal subtype modules

(a) pLI rates as a function of gene kME, showing that genes with a high kME to BW-M1 have a far higher than background pLI rate (0.4 vs 0.23); while those with a high kME to BW-M2 have a far lower than background pLI rate. **(b)** As **(a)**, but for the neuronal subtype modules BROD-M8 (interneuron), CEREB-M2 (Purkinje), and STR-M1 (medium spiny). **(c)** Enrichment statistics for loss-of-function genes calculated by Fisher's exact test, o/e ratio (inversely related to pLI) is cut into 8 bins. Low o/e implies intolerance.

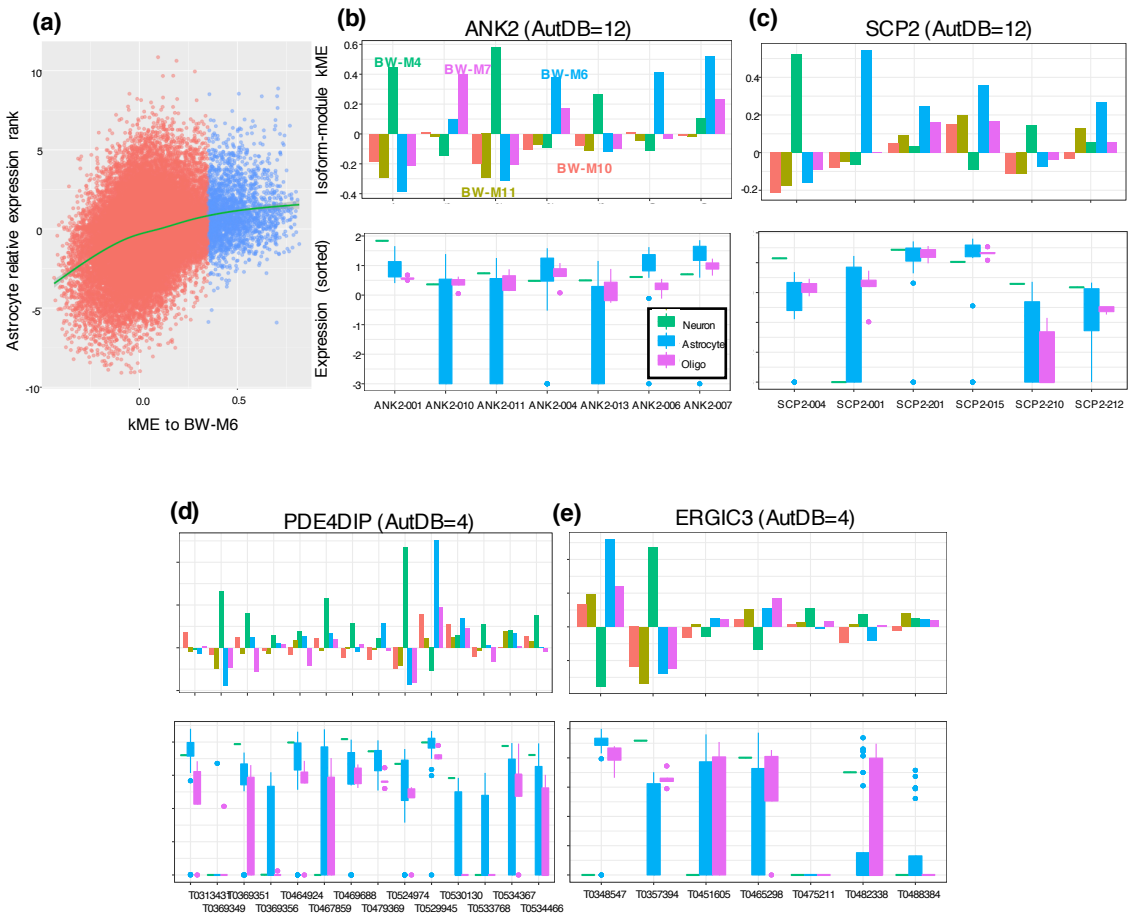

Figure 2.13 Isoform switch gene validation in single-cell data, related to 2.12

(a) Replicate of figure 2.11(b) for astrocytes, showing a strong positive relationship between astrocyte module membership, and relative expression in astrocytes. **(b-e)** Relationship between module kME and cell type relative expression for transcripts across 4 neuron/astrocyte isoform switch genes, demonstrating concordance between high kME, and high relative expression.

2.3f Region-specific upregulation: increased protein turnover in subcortical brain regions

I next sought to incorporate regional differential expression with co-expression to provide a more refined view of modules across brain regions. While differential expression has been used to identify differences between brain regions (Negi2017), this approach is too broad: nearly every gene expressed in brain shows differences in expression across brain regions (n = 15616/15894, FDR < 10^{-3}, likelihood ratio test). I reasoned that genes with "extreme" expression profiles (i.e., significantly up-regulated or down-regulated in a region-specific or multi-region-specific manner) are more likely to have a specific role; and that those differing substantially in their expression levels within brain-wide cell-type modules may reflect regional differences in cell function. I developed a Regional Contrast Test (RCT, **2.14a**), which assigns Z-score per group per gene, whereby Z-scores > 4 represent over-expression within that group, and those with Z-scores < -4 represent under-expression (**methods**). This approach identifies e.g. genes where the minimum expression across all cerebellar regions is still larger than the maximum expression across cortical regions. To account for the presence of multiple regions, I performed this analysis with several different backgrounds: the whole brain, non-cerebellar tissue, telencephalophalic regions, and cortical vs subcortical regions.

Figure 2.14 presents a summary of these results. I first examined the set of genes up-regulated in subcortical regions (striatum, hippocampus, and amygdala) versus cortex, and observe that these differences enriched for non-neuronal cell type modules (p < 1e-10 for BW-M11, BW-M6, BW-M8, BW-M10, and BW-M7), consistent with a higher glia/neuron ratio in the striatum. Expanding this approach to non-cerebellar regions and finally the whole brain region, I observed that the region-specific upregulated genes within each comparison largely reflect neuronal heterogeneity: striatum-upregulated genes reflect MSN markers, cholinergic and

noradrenergic neurons are over-represented in hypothalamus-upregulated genes, while genes expressed at the highest level in cortex (versus the striatum, amygdala, or hippocampus) reflect pyramidal and GABAergic neurons.

When comparing cortical and subcortical expression, I observed module BW-M4 (neuronal) to enrich for the genes up-regulated in the cortex. Perplexingly, I also observed a significant ($p = 4.89e-3$) enrichment in BW-M2, a module dominated by small- and large-ribosomal subunit RNA for sub-cortical upregulated genes. This suggests that subcortical regions may show higher translational demand, more ribosomes, or faster ribosome turnover than cortical regions. Recent work has demonstrated that protein turnover – particularly the large and small ribosomal subunits – drastically increases in cultures with high glial proportion (Dorrbaum2018), providing an explanation for increased abundances of these ribosomal mRNA in regions of high glial proportion in the brain.

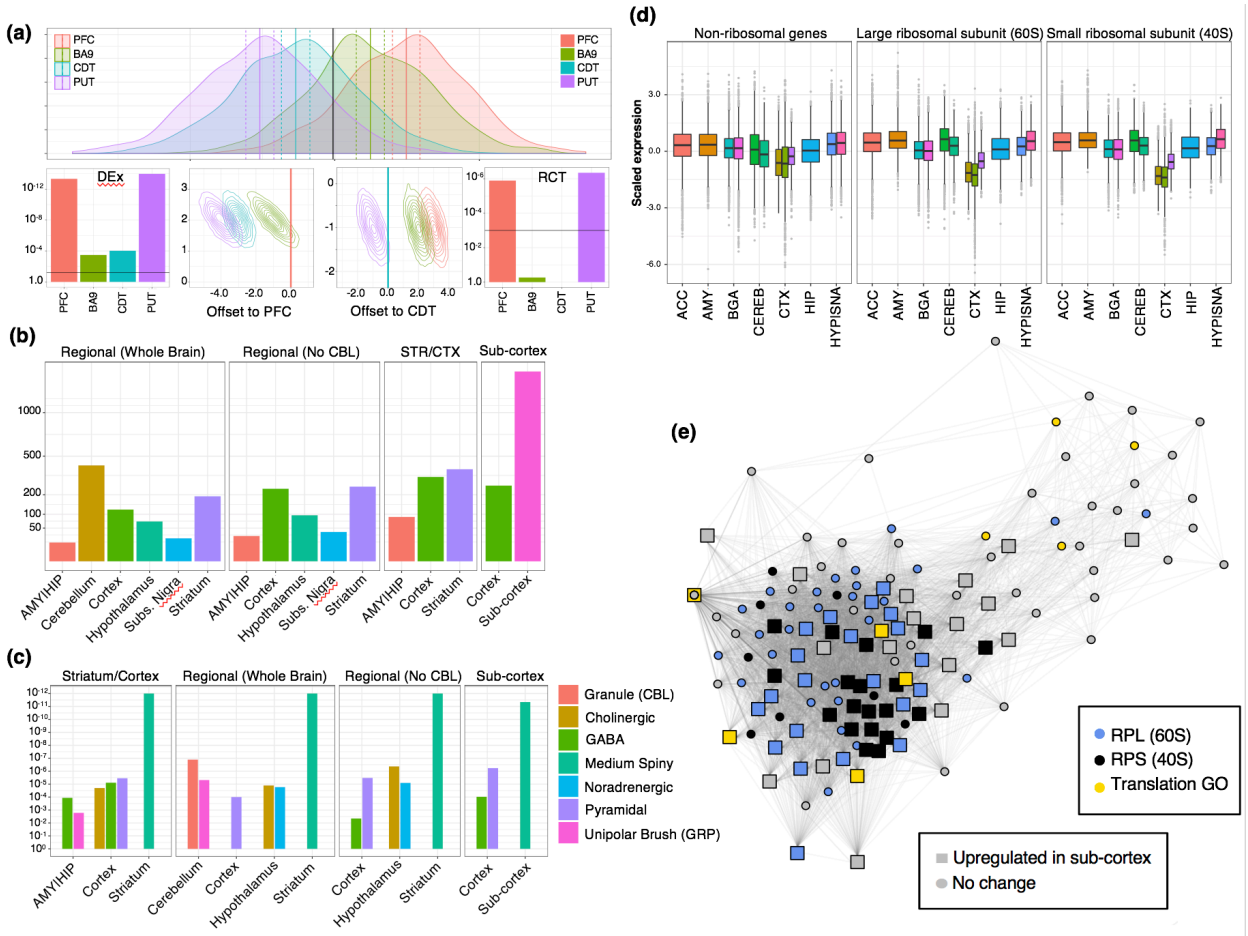

Figure 2.14 Region-specific gene up-regulation reflects region-specific cell types and ribosomal turnover.

(a) Overview of the regional contrast test: *top* Example of heterogeneous data where mean expression within each region differs from the global mean. *left* With only 50 samples, all regions are significantly differentially expressed in a global manner. *middle* Visualization of the PFC statistic for PFC and CDT: The PFC mean (set to 0) overlaps only a small amount of the confidence region for another region, while confidence regions straddle the CDT mean. *right* The RCT statistic identifies the two most extreme tissues as differentially up- and down-expressed compared to all other regions. **(b)** Count of genes which are significantly up-regulated within brain regions, across four backgrounds. **(c)** For the gene sets identified in (b), enrichment p-values for markers of neuronal cell subtypes. **(d)** Plot of scaled expression (per gene across tissues) for all genes in BW-M2, showing CTX-specific down-regulation of ribosomal subunits. **(e)** PPI-coexpression network for genes in BW-M2, showing a sizeable fraction of the module core, a substantial fraction RPL and nearly all RPS mRNA are up-regulated in sub-cortical regions.

2.4 Discussion

Gene co-expression networks have provided a powerful organizing framework for transcriptomic studies of the nervous system. Here, I have described the construction of a robust, hierarchical, co-expression resource aimed at establishing common and region-specific aspects of gene co-expression within the brain and within a single hierarchical framework. I identified 11 major whole-brain signatures represented in co-expression modules, corresponding to common cellular components such as major neuron and glial types. I also captured region-specific signatures dominated by regional cell subtypes. By using a consistent framework that allows for module relationships to be grouped by spatial scale, I demonstrated that the spatial extent of co-expression networks relates to the spatial extent of underlying cell types, that cell-type-specific lncRNA and isoforms are identifiable from these networks, and that isoform-level analysis is likely essential to interpret certain disease associations.

The relationship between bulk tissue co-expression modules and cell-type-expressed genes has been well-established for markers of major cell classes (Oldham2008, Kelley2018), but the quantitative link between module membership and cell-type relative expression appears to be novel. This link establishes bulk tissue co-expression networks as a valid method for marker discovery and ranking, both for gene and isoform expression. Since there are no published single-cell datasets from human striatal tissue, the kME values for striatal modules M1 and M2 may represent the most salient information regarding both genes and isoforms with high relative expression in human spiny neurons. As sample sizes grow, we expect to identify new subdivisions of co-expression modules, corresponding to ever finer distinctions between underlying cell types.

I showed, through analyzing lncRNA from a separate collection of brains – sequenced in a different location and with different technology – that the modules in the atlas can be imputed in a separate and smaller bulk dataset, yet retain the same cell-type signals. This approach is potentially very powerful: it allows every gene in a new expression dataset to be identified as region-specific or cell-type specific, without requiring large sample sizes for co-expression analysis or samples from multiple regions.

I used a similar approach to create isoform networks, utilizing module distance to assign isoforms into modules, which in many cases are cell-type specific. This is important, as single cell sequencing approaches do not yet capture full length transcripts and thus will at best incompletely represent isoform expression profiles. I identified a first generation set of 1,987 cell-type specific isoforms for major cell classes in the brain – including 549 neuron, 543 astrocyte, 696 oligodendrocyte. Remarkably, several of these isoforms, including 4 ASD risk genes, manifest isoform switching between neurons and glia. One of these, *ANK2*, has been recently described and validated as having different isoforms in neurons and glia, which manifest distinct protein-protein interactions (PPI; Gandal2018a). As long-read sequencing matures and is applied at larger scale, more complete cell-type specific isoform networks can be constructed using this approach. These data indicate that this will be of substantial value in understanding disease-relevant variation.

2.5 Methods

Expression quantification, QC, and covariate correction

Reads were aligned using STAR (Dobin2013) in standard two-pass fashion. Gencode v25 transcripts (hg19/b37) were used as the reference transcriptome and genome for alignment. Transcripts were quantified using RSEM to produce gene and isoform level TPMs. The analyzed TPMs are log-transformed log(0.005 + x) resulting in approximate normality.

Sample and individual-specific covariates were downloaded from the GTEx (GTEx Consortium 2017) website, and supplemented with technical alignment information from the STAR alignment and PicardTools QC of the resulting .bams.

Individuals were excluded if they were positive for any of the following phenotypes: 'MHALS', 'MHALZDMT', 'MHDMNTIA', 'MHENCEPHA', 'MHFLU', 'MHJAKOB', 'MHMS', 'MHPRKNSN', 'MHREYES', 'MHSCHZ', 'MHSEPSIS', 'MHDPRSSN', 'MHLUPUS', 'MHCVD', 'MHHIVCT', 'MHCANCERC', 'MHPNMIAB', 'MHPNMNIA', 'MHABNWBC', 'MHFVRU', 'MHPSBLDCLT', 'MHOPPINF'. The individual-specific covariates 'GENDER', 'AGE', 'RACE', 'ETHNCTY', 'TRISCH', 'TRISCHD', 'DTHCODD', 'SMRIN', 'SMNABTCH', 'SMGEBTCH', 'SMTSISCH', 'SMTSPAX' were extracted. The `DTHCODD` variable was binned into the following categories: ‘UNKNOWN’, ‘0to2h’, ‘2hto10h’, ‘10hto3d’, ‘3dto3w’, ‘3wplus’.

STAR alignment metrics and PicardTools QC metrics were subset to non-excluded samples, outliers were flagged and removed via a chi-squared test ($p < 10^{-5}$). The PicardTools metrics were log-scaled, and the top 5 principal components extracted using the PCA class from

scikit-learn (Pedregosa2011) (“seq-PC”). The STAR alignment covariates were subset to those with “splice” in the feature name, and the top 3 principal components similarly extracted (“STAR-PC”).

Given the gene expression and covariate matrices, features that explain a significant proportion of expression variance in a non-trivial subset of genes were extracted using a forward-backward regression approach (see supplemental methods). This approach identified the features "seq_pc1", "seq_pc2", "seq_pc3", "SMRIN", "SMEXNCRT", "Number_of_splices_GT/AG", “TRISCHD” and “DTHCODD” as significant features, with no significant interactions between these features or between any of these covariates and tissue type.

Because there were no significant cross-terms between tissue and covariate, all tissues were combined for the removal of covariate effects. A linear model (expr ~ tissue + covariates - 1) was applied, with a separate intercept (mean) for each tissue. The covariate effects were removed, while the estimates of mean expression per tissue were retained.

Forward-backward covariate selection using MARS (earth)

A key step in the treatment of RNA-seq data is identifying what technical or biological covariates are strong drivers of measured expression. RNASeqQC produces a large set of alignment metrics derived from the aligned RNA-seq bams. I combined these with the splicing metrics output by STAR. Separately, each of these data were scaled and the top 5 PCs calculated to summarize the bulk of the technical covariate distribution, producing an additional 10 potential covariates. This final set of technical covariates are combined with the sample-level

individual-level information provided by GTEx (ischemic time, age, biological sex, RIN, ethnicity, race).

I then used the `earth` package in R to select covariates that explained a large amount of expression variance across many genes. We set the parameters so that no non-linear splines were used, but that cross terms up to degree 3 were allowed, enabling the model to select tissue-by-covariate or covariate-by-covariate effects.

'earth' builds a forward model by selecting the covariate (or cross term) which most improves the total R^2 across all genes considered; and when a diminishing-returns threshold is reached (for us, an improvement of 0.01), prunes the terms using a penalized R^2 heuristic.

I ran earth 100 times on a random sample of 1,000 genes; each run producing an estimate of variance explained for all covariates (covariates *not* included in the model are assumed to explain 0% of expression variance). We summarized the impact of each covariate by taking the upper 20% of the variance explained (**figure S1a**). Any covariate whose summary estimate was >5% variance explained was included in our final model for covariate correction. For group variables (such as tissue); if any subgroup exceeded the variance explained threshold, then the entire group variable was selected.

Notably, no cross-terms exceeded the threshold for variance explained, suggesting that I could perform covariate correction simultaneously across all tissues. The lack of region-by-covariate effects may be due to the fact that the library preparation batches and sequencing batches are well-balanced across brain regions.

Tissue hierarchy

The median expression of all genes across a given tissue is taken as the *exemplar* of said tissue. These exemplars (12 in all) are hierarchically clustered into the tissue hierarchy observed in figure 2.3.1 using Euclidean distance and single-linkage hierarchical clustering.

Module construction

Robust WGCNA:

Robust rWGCNA (Langfelder2008) was applied to each brain tissue independently. Briefly, the power parameter is selected as the smallest power (between 6 and 20) which achieves a truncated r^2 of >0.8 and a negative slope. Then, 50 signed co-expression networks are generated on 50 independent bootstraps of the samples; each co-expression network uses the same estimated power parameter. These 50 topological overlap matrices are then combined edge-wise by taking the median of each edge across all bootstraps.

The topological overlap matrices are then clustered hierarchically using average linkage hierarchical clustering (using `1 – TOM` as a dis-similarity measure). The boostraps are used to determine cut height as follows: multiple cut-heights are considered (0.9 to 0.999, by 0.005); and for each cut the within-module correlation of TOMs is considered. For the top 8 modules by size (fewer if fewer modules are produced), the consensus and each bootstrap TOM is subset to the genes within each module, and the correlation between bootstrap and consensus is computed. The median (within module, across bootstraps) of these consensuses is computed, and the mean of these summaries is taken to be a measure of `goodness` for the cut. The cut height which maximizes this metric is taken to define the initial modules.

These initial modules are then merged via `mergeCloseModules` in WGCNA, which hierarchically re-clusters modules based on the module eigengenes, using the correlation-based adjacency as a dis-similarity matrix. Modules with a distance of < 0.35 are merged together into a combined module.

Aggregating co-expression:

At each merge of the hierarchy, a single round of consensus topological overlap is performed. Each pair of genes has two descendent edges, and the parent edge is estimated as the 80^{th} percentile between the two (i.e. for $x < y$; $p = 0.2\ x + 0.8\ y$)). This process proceeds up the tissue hierarchy until a single network TOM remains.

Consensus labeling:

After construction of co-expression networks from all tissues and splits, modules have been defined for a total of 21 groups (BRNACC-BRNSNA, BROD, CTX, CBL, BGA, STR, NS-SCTX, SCTX, NCBL, WHOLE-BRAIN), yielding over 300 overlapping modules. The overlapping nature of these modules motivates labeling each module in terms of a hierarchy group, allowing one to identify (say) BRNHYP-M2 and BRNCTX-M7 with the module group WHOLE-BRAIN-M3.

To perform this labeling, similarity matrices are computed. First, the module eigengenes for all modules (regardless of origin) are computed within every tissue, and the correlation matrix (using `bicor') is computed for each module for each tissue. This produces an (all modules) x (all modules) matrix for each tissue. The consensus eigengene similarity ("E") between two modules is chosen as the component-wise maximum of all of these matrices. The second similarity matrix is the standard Jaccard similarity ("J") between module gene lists. These

similarities are combined into a dis-similarity matrix D = 1 - (E + 3*J)/4, which is used to hierarchically cluster (average linkage) these modules.

Module groups are defined by cutting the dendrogram at a height of 0.35. This process results in a set of module clusters, each of which has a "level" in the brain tissue hierarchy (for instance, a cluster of BRNCTXBA9-M4, BRNCTXB24-M2, CTX-M7 would have the level "CTX" as the top-level of the tree represented is CTX). The "representative" of the module group is taken to be the module at the highest (most rootward) level of the tree – and if there are two, the larger of the two. A second round of clustering is performed by removing all modules in the group (except for its representative) from the dissimilarity matrix, and re-clustering only the group representatives. This process repeats until there are no additional merges. Finally, each module is labeled with its group representative; for instance "BRNCTXBA9-M4" would receive the label "CTX-M7", because it shares its highest similarity with the consensus cortex module M7.

In addition, I re-named and abbreviated modules: "BW" for brain-wide, "NCBL" for non-cerebellar, "NS.SCTX" for non-striatal subcortex, "CEREB" for Cerebellum; and the GTEx tissue names were abbreviated to clear region codes: ACC, AMY, B24, BA9, CBH, CBL, CDT, HIP, HYP, PFC, PUT, SNA.

Preservation

I consider two module preservation statistics: the classical Z-summary (Langfelder2011) and a leave-one-gene-out neighbor statistic. For the classical Z-summary; module statistics such as the mean gene-gene correlation in the module, the correlation-of-correlations across datasets,

the variance explained by the first module PC, and other metrics are computed for each module (in both the original and comparison dataset); and compared to 100 random (via permutation) modules of identical size. Each observed statistic is converted to a Z-score, and these are averaged to generate a final summary, for which large Z-scores are indicative of replication of the underlying biological signal.

The neighbor statistic ("Z-AUPR") is strongly influenced by the single-cell statistic MetaNeighbor (Crow2018). Briefly, a *k*-nearest-neighbor network is built in the comparison dataset (we use k=15), and we impose the module labels from the reference dataset. For each gene, we compute the proportion of its neighbors (again, in the comparison dataset) whose labels match its own. Note that if this proportion is > 0.5, then this gene *would* be assigned the same label in the comparison dataset as the reference dataset under a neighbor-voting scheme. Using these scores, we can compute an AUPR for each module. We repeat this approach for 100 permuted modules (and, unlike the WGCNA permutation, we split genes into connectivity deciles, and permute only within decile), and use this baseline to convert observed AUPR to Z-scores. As with the classical Z-summary, high Z-AUPR is indicative of replication of underlying biological signal.

Module comparisons

I considered three alternatives to WGCNA for network building and module identification: ARACNe, GLASSO, and von-Mises-Fisher clustering.

ARACNe was run with default settings (10 permutations, FDR of 0.05); and genes filtered by ARACNe (for having no significant edges) were placed into a background 'grey'

module. The resulting network was imported into iGraph (Csardi2006) and modules identified by Louvain clustering.

As sparse inverse-covariance estimation is computationally intensive, I took an approximate approach. First, we partitioned the genes into initial groups of approximate size 1000 using k-medioids clustering. GLASSO was applied independently to each group to estimate a blockwise precision matrix. Within each block, the penalty parameter was selected using StARS (Liu2010), targeting an edge instability of between 0.05 and 0.1. Genes with no partial correlation to any others were grouped into a background 'grey' module. The remaining network was imported into iGraph and modules identified by Louvain clustering.

vMF mixture modeling, unlike the other approaches, does not build a network, but seeks to identify gene clusters directly. Gene expression vectors were pre-processed by transforming their values into ranks (across samples) and normalizing them to unit norm. In this way, an inner product between two gene vectors is effectively their Spearman correlation. The resulting data is modeled as a collection of draws from an n-dimensional mixture of k von-Mises-Fisher distributions (where n is the number of samples). The model was fit using the R package movMF (Hornik2014) for k varying from 8 to 50. The final choice of k came from the model that maximized likelihood – 2 * ndim * k; and module assignments were determined from the most likely mixture probability (or 'grey' if that probability was less than 0.8).

Whole-brain module comparisons

Beyond comparing modules within each tissue, I sought to compare the hierarchical WGCNA modules with an orthogonal approach for building consensus modules. As consensus

modules built from methods already similar to WGCNA would certainly produce similar consensus modules, we considered an alternate approach: tensor decomposition.

First, I built a fully imputed (gene x brain x region) tensor by using probabilistic PCA to impute missing samples within every (brain x region) submatrix for each gene. I then applied CANDECOMP to this tensor to produce 150 feature triplets: {(gene x 1), (brain x 1), (region x 1)}. We treated the gene-level features as a (gene x 150) feature matrix, and ran t-SNE to embed the genes in a 2-dimensional space.

While this embedding did not show distinct visual clusters, it clearly showed regions of high and low density, likely corresponding to modules. Given this intuition, I applied the DBSCAN clustering algorithm, producing a set of 30 whole-brain modules.

I found that the ribosomal, glial, and choroid-plexus modules were in one-to-one correspondence with TD-DBSCAN modules, and that the neuronal WGCNA modules correspond to multiple TD-DBSCAN modules, with statistically significant overlaps. Visually, the WGCNA modules are localized in the embedded tensor-decomposed space, strongly suggesting that the modules are not driven by the specifics of WGCNA, nor are they induced by the structure of hierarchical merging; but rather that these genes are grouped together by disparate approaches because of an underlying biological signal.

Learning curves

To examine how module identification and specificity changes as a function of the number of samples, I combined samples from similar tissues to increase the maximum N: we

combined the cerebellar samples into one larger group (N=122), and we also grouped the cortical samples (PFC, B24, BA9) together with hippocampal samples into a second group (N=304).

"Reference" modules for these groups were determined by applying rWGCNA to the full dataset. We down-sampled the group to a smaller set of samples of size n = 25, 50, …, N and performed rWGCNA on the smaller set. I repeated this process 10 times, generating 10 networks and module assignments for each sub-sampling of the full dataset.

Because two clusterings should be considered identical up to renaming the labels in one or the other datasets, we use module co-clustering as a measure for accuracy, precision, and recall. Within the reference (whole group) dataset, we extract the top 'hub' gene from each of the modules, and the list of genes co-clustered with that hub gene (i.e. the other members of its module). For a given reference module, within a sub-sampled dataset, one has

Recall = (# ref hub co-clustered genes also co-clustered in subsample)/(# ref hub co-clustered genes)

Precision = (# ref hub co-clustered genes also co-clustered in subsample)/(# subsample co-clustered genes)

In effect, these are precision/recall statistics for the hub gene co-clustering indicators. If two reference modules fail to separate in a sub-sample (a typical failure mode), the result is slightly higher recall, but far worse precision.

Single-cell data

Quantified single-cell data was downloaded from http://mousebrain.org (mouse; Zeisel2018) and subset to only cells from the CNS (without spinal chord); and GEO GSE97942 (Lake2018) was downloaded for human. These data were log-transformed log(1 + x) for counts and log(0.005 + x) for TPM; and the cell type labels from the respective publications were used for all subtype analyses. Absolute expression values were taken as the mean expression of a cluster; and relative expression was obtained via `*Relative = absolute – background*` where the background expression is the average expression of a gene over all cells. To incorporate gene variance information into relative expression, the *relative expression rank* is defined as the lower end of a small confidence-interval for the difference in means:

$$\text{rank} = (\mu_a - \mu_b) - \ 0.5 * \sqrt{\frac{v_a}{n_a} + \frac{v_b}{n_b}}$$

kME enrichments are based on the correlation between module kME and the relative expression rank within a given cell type.

Cell-type enrichment and single-cell data

For kME-based enrichments (such as those in figure 2), the shaded region of the figure represents the standard error around the estimated functional relationship between kME and relative expression rank. In all cases it is visually apparent that these lines deviate from 0 by a factor far exceeding 2.5 times their standard error ($p \sim 0.006$).

For gene-set based enrichments such those presented in the text, and those in figure 3, cell type markers were obtained from several sources (Zhang2014, Zhang2016, Miller2010, Mancarci2017, Romanov2016, Tasic2016, Heintz2004, Kelley2018) representing various studies performed both in mouse and in human. The statistical test is a logistic regression using the model

is.cell.marker ~ 1 + is.in.module + gene.length + gene.gc

adjusting for gene length and GC. I test that the coefficient for module presence is significantly different and greater than zero, implying an enrichment (as opposed to depletion) of cell-type related genes. This test is performed independently on cell type markers from the various studies, and FDR adjusted across all tests.

Defining mouse orthologs to human genes

The ensembl API was used, through biomaRt, to query human genes with associated mouse orthologs and the type of orthology; and visa versa. These queries enabled defining genes as one-to-one orthologs, one-to-many orthologs, many-to-many orthologs, or non-orthologous. The ensembl API was also used to obtain human-mouse dN and dS values; and the ratio dN/dS calculated, with 0/0 treated as 0.

Module Imputation

For the lncRNA analysis, I imputed whole-brain modules into an independent RNA-seq dataset (Parikshak2016) by i) splitting the data into BA9 and BA41-42-22 regions, ii) Calculating module kMEs within each region, and iii) Averaging across the two regions. This generates a set of 11 features (average within-region kME to each module) for each gene. The overlapping genes between the GTEx modules and control brain expression were used as labels to fit a boosted trees classifier (using the R package xgboost with 2000 trees and a learning rate of 0.025). Non-overlapping genes (which contain most lncRNA and a set of held-out, matched protein-coding genes) are assigned to modules via the prediction of the fitted classifier. Using cross-validation we estimate that the sensitivity and specificity of this approach are 0.63 and 0.53 for BW-M6, with all misclassifications resulting from assigning a 'grey' gene as in the module, or a BW-M6 gene as 'grey'.

Human-specific modules

To define modules exhibiting human-specific differential expression, I obtained the modules and human-specific differentially-expressed gene list from Sousa2017. I subset only to modules flagged as showing inter-species heterogeneity, and computed enrichment p-values and FDR values by Fisher's exact test, using the intersection of all GTEx-ascertained genes and Sousa2017-ascertained genes. This resulted in a set of 25 modules with enrichment FDR<0.1 for human-specific differentially expressed genes.

Using the same statistical approach and background gene set, I then tested for significant overlaps between the 311 GTEx modules and the 25 human-specific modules, identifying 10 with an enrichment FDR < 0.1. These overlaps are plotted in figure 2.3.2. Not every module in

brain-wide module sets necessarily overlapped at FDR<0.1; so the figure reflects the proportion of modules within brain-wide module sets that show such an overlap. Furthermore, because hypothalamus and substantia nigra were not profiled in Sousa2017, these regions (and the NS.SCTX region) were excluded from this fraction calculation (but not from the initial overlap tests and FDR correction).

Sousa2017 also lists cell types in which these modules are expressed. These are summarized in figure 2.8(g). Expression is listed for cell types Ex1-Ex8 and In1-In8; for space this is collapsed to the fraction of Ex and In in which the module is expressed, so a gene expressed in In4 and In2 would receive a value of 0.25 for the "In" group.

GO enrichment

Gene ontology enrichment is performed competitively, with covariate correction, using logistic regression. Briefly, each GO category is treated as a binary variable (1 for genes in the category, 0 for genes not in the category – only genes ascertained in our gene expression matrix are part for the regression). Modules are also treated as binary. I include as covariates the average gene expression across all tissues in the brain, the gene GC content, and the log gene length. The GO enrichment model is then

GO ~ module.1 + … + module.k + mean.expr + GC + log.gene.length

and is fit using logistic regression. If convergence fails, an L2-regularized logistic regression is instead applied (using `brglm`). The enrichment p-values are taken to be the statistics that reject

($\beta_i \leq 0$) for all β_i corresponding to a module indicator. The enrichment p-values are adjusted to FDR values across all ontology categories.

Meta-GSEA

To aggregate enrichment results (such as GO) from the module level to the module set level, the GO p-values are treated as independent p-values, and Fisher's method is applied: For a given ontology category, a χ^2 value is calculated as $-2 * \log(p_1 * p_2 * \ldots * p_k)$, where the product is taken across modules in the set. In the case of independence, this statistic has $2*k$ degrees of freedom; allowing a p-value to be calculated. Because the modules in a set overlap by construction, the resulting statistics are not calibrated probabilities, and are referred to as "scores" or "rankings," and should not be interpreted as reflecting significance. In nearly all cases, the highly-ranked consensus ontology had been significant in one or more of the modules within the set.

Meta-GSEA was applied the genes within the regional BW-M4 modules (e.g. PFC-BW-M4) with MAGMA Z-scores > 3.0 (SCZ) or 2.5 (ASD). This generated an indicator variable which was then used to perform gene ontology, using the BW-M4 genes as a background; generating p-values for each ontology. Meta-GSEA was applied to these p-values, generating a score for each ontology.

pLI enrichment

Gene pLI scores were downloaded from the ExAC consortium release [cite], and a gene was considered likely to be LoF-intolerant if its pLI score was 0.9 or higher. Enrichment for "hard" module membership (i.e. comparing two gene lists) is performed via Fisher's exact test on the contingency table between module membership and LoF-tolerance/intolerance. "Soft" module enrichment (i.e. based on kME) is computed via a Brownian Bridge statistic.

The genes were ranked by their module membership (kME); and the proportion of all genes which are likely LoF-intolerant (the pLI rate, $r=P/M$) is computed. At a given quantile q of genes, I tabulate how many of the first $q * M$ genes are LoF-intolerant; and denote this cumulative sum by $Cs(q)$. The expected number of LoF-intolerant genes is $Ne(q) = q * P = q * r * M$. For large M, this cumulative sum converges to a scaled Brownian motion with drift r; and has variance $V(q) = q * (1 - q) * M * r * (1 - r)$. Z-scores for this cumulative sum at each q are given by $Z(q) = (Cs(q) - Ne(q))/\sqrt{V(q)}$. An excess of LoF-intolerant genes occurs when $min_q \Phi(Z(q)) < 0.05$. For clearer visualization, we plot $(Cs(q) - Ne(q))$ and $2.17 * \sqrt{V(q)}$ as functions of q.

I also used a generalized additive models ("GAM") and a generalized linear models ("GLM") to verify findings of constraint. In these cases I applied the (logistic) model

is.constrained ~ rank(kME)+ gene.length + gene.GC

and found that, for the whole-brain modules, these enrichments were so strong that the three methods were in 100% concordance. The results of the linear models did not change substantively when using competitive as opposed to marginal enrichments.

For method validation (binary enrichment in pLI and o/e bins), the odds ratio and p-values were computed using a Fisher Exact Test between module membership, and bin membership.

PPI enrichment

We use InWeb PPI database (Li2016; brain tissue) for a source of defined PPI, with a confidence threshold of 0.2 used as a cutoff for a particular interaction. PPI prediction is treated as edge-related data, where the response variable is binary (presence/absence of PPI), and the predictors the following collection of data relevant to that edge: the (PPI) connectivity of its first vertex, the (PPI) connectivity of its second vertex, the product of kMEs of its vertices (for each module), the product of the GCs of its vertices, and the product of the reproducibilities of its vertices. Or:

$$E_{ij} \sim C_i + C_j + kME\_M1_i * kME\_M2_j + \ldots + kME\_Mk_i * kME\_Mk_j + GC_i * GC_j$$

This equation encodes the model that gene pairs which are mutually close to a given module are more likely to physically interact. The logistic model is fit using `statsmodels` in python, and the hypotheses $\beta i \leq 0$ is assessed for each βi corresponding to a module. By testing the PFC modules, I found perfect concordance for PPI enrichment (<0.05) between this method and DAPPLE (Rossin2011).

Regional contrast test

The Regional Contrast Test is a multivariate test of significance for

H_0: $\beta_i \leq \max(\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_n)$

H_a: $\beta_i > \max(\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_n)$

This statistic corresponds to a multidimensional integral, with infinite limits on all coefficients other than β_i, and taking $\max(\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_n) < \beta_i < \infty$. Because of the large numbers of degrees of freedom in this regression, I treat the variance-covariance matrix ($\Sigma_\beta^{(ML)}$) of the β vector as giving the true sampling covariance of these parameters, and perform Monte-Carlo integration by drawing 50,000,000 samples from the multivariate normal distribution $N(\beta, \Sigma_\beta^{(ML)})$ using the R package *fastmvn*.

The above statistic works for testing each tissue against all others. A grouped version of the test is a simple extension, which considers several β in tandem. For simplicity we assume the indexes for the group are the first *k* coefficients, then the comparison becomes:

H_0: $\min(\beta_{1, \ldots,} \beta_k) \leq \max(\beta_{k+1}, \ldots, \beta_n)$

H_a: $\min(\beta_{1, \ldots,} \beta_k) > \max(\beta_{k+1}, \ldots, \beta_n)$

This only changes the integration limits to (for $j \leq k$) to $\max(\beta_{k+1}, \ldots, \beta_n) < \beta_j < \infty$; and I use the same Monte-Carlo approach as before.

Post-hoc tests for module enrichment use Fisher's exact test on the contingency table

	Significant (RCT)	Not significant (RCT)

In module	A	B
Not in module	C	D

Isoform specificity from sorted cell data

RNA-sequencing data was obtained from GSE73721 (SRA project SRP064454) and quantified at the isoform level with Kallisto (mouse gencode release M16). These data included sorted populations of astrocytes, oligodendrocytes, endothelial cells, a single neuronal population, and a whole-tissue background. Relative isoform expression were obtained as described in "Single-cell data," with the background set to be the average expression across the whole-tissue background samples.

Isoform switching and validation

Isoform-level TPM values (produced by RSEM) were corrected using a linear model with the same covariates used for correcting gene expression TPMs. Subsequently, each isoform expression (within tissue) was correlated to brain-wide module eigengenes computed within the tissue, and the mean correlation across tissues taken as an estimate of module membership for the isoform.

To determine an appropriate kME threshold, I evaluated the impact of thresholding on cell type enrichments. Each threshold produces a set of isoforms within a module; and each isoform can be annotated with the cell type marker status of its parent gene. Fisher's Exact Test

produces an odds ratio and p-value for cell-type enrichment at each threshold. I found that a threshold of 0.45 produced a 15-fold enrichment for both astrocyte and oligodendrocyte markers when looking at kME to their respective modules (M6 and M7); but that when increasing this threshold the odds ratio for oligodendrocytes did not substantially change, while the astrocyte odds ratio increased. Based on this I defined the threshold for isoform module membership at 0.45 kME. In the case where an isoform has >0.45 kME to multiple modules, module with highest kME is selected.

An "isoform switch" is defined as two sister isoforms having membership to different modules.

Western Blot Analysis

Human iPS cells were differentiated into cortical glutamatergic-pattern neurons (GPiN) according to Nehme2018, and samples extracted at days 0, 16, 21, and 31. Human astrocytes were used as an outgroup. Western blot was run as per Wu *et al.,* 2015, using the G-11 antibody sc-365757 from Santa Cruz Biotechnology.

Acknowledgements

Western blot analysis was performed by Greta Pintacuda (Eggan Lab, Broad Institute of MIT and Harvard). ARACNe network construction was adapted from code and initial analysis provided by Gokul Ramaswami (Geschwind Lab).

Chapter 3 Linking neuropsychiatric disease to regional brain processes

3.1 Abstract

Genomic and transcriptional disruptions in neuropsychiatric disease have been shown to impact neuronal pathways and cortical co-expression. Whether these pathways reflect neurobiology common to the whole brain, or instead reveal a region-specific pathology, is unknown. Using the human brain co-expression atlas developed in chapter 2, I identify enrichment of neuropsychiatric disease risk variants in brain wide and multi-regional modules, consistent with their impact on major core cell types – primarily neurons. Nearly all previously-published disease modules overlap whole-brain modules, implicating nonspecific pathology. I also identify regional modules that are both intolerant to loss of function mutations and enrich for neuronal-activity-dependent processes that are disrupted in neuropsychiatric disease.

3.2 Introduction

High-throughput genomics has engendered rapid progress in understanding the genetic signature of and genomic basis for neuropsychiatric disorders such as autism (ASD), schizophrenia (SCZ), bipolar disorder (BP), and major depression (MDD). These prevalent genetic diseases are now known to be highly polygenic and to exhibit a high degree of genetic overlap (Anttilla2018): both genetic risk factors and transcriptional signatures converge onto a selection of neuronal and neurodevelopmental pathways (Gandal2018a). Many of these points of convergence have been defined by co-expression networks built from cortical gene expression, and may reflect either regional vulnerability, or brain-wide disruption.

Using the structured brain co-expression atlas built in chapter 2, I interrogate previous neuropsychiatric disease associations for regional specificity, and investigate the atlas itself to identify signatures in the normal brain that may be disrupted in neuropsychiatric disease. Finally,

using the deep annotation of the co-expression atlas, I identify potential biological processes that may contribute to the etiology of neuropsychiatric disease.

3.3a Qualifying regional specificity of previously-identified neuropsychiatric disorder co-expression networks

The previous chapter of this thesis developed an atlas of co-expression in the human brain, organizing co-expression relationships into those shared across or specific to regions of the brain. The obvious structural differences between humans and non-human primates led to a historical focus on neocortical regions for the study of human neurological disease – both neuropsychiatric and neurodegenerative. I therefore reasoned that the co-expression atlas could be used to provide regional localization of previously identified disease-associated co-expression signatures.

I therefore re-evaluated gene modules identified in post mortem tissue from 11 publications – normal brain (Konopka2012, Hawrylycz2015) , ASD (Parikshak2016), SCZ (Fromer2016, Radulescu2018), cross-psychiatric (Gandal2018a), Alzheimer's disease (Wang2016), epilepsy (Johnson2016), and developing brain (Parikshak2013, Hormozdiari2015, Mahfouz2015) – with the objective of identifying: i) whether or not those previously discovered modules, typically based on analysis of only one or two brain regions were related to any the modules we identified from the normal individuals in GTEx, and ii) whether those previously discovered modules are indeed region-specific.

Johnson2016 profiled expression in resected hippocampi from 122 epileptic patients, and used WGCNA to identify 24 co-expression modules. Using GSA-SNP, they identified of modules M1 and M3 as enriched for genetic association to cognitive ability; and using Fisher's exact test, they identified module M3 as enriched for *de novo* mutations in intellectual disability

(ID) probands, and for a combined ID+ASD+SCZ cohort. This small module (150 genes) strongly overlaps (OR=6, $p < 10\text{-}5$) HIP-BW-M4 and no others, strongly suggesting that this is part of the brain-wide neuronal signature. Similarly, Fromer2016 generated co-expression networks separately in SCZ (n=278) and normal (n=254) prefrontal cortex; identifying a single module, M2c, showing enrichment for differentially-expressed genes, GWAS signal, and genes within SCZ-associated rare structural variants. This large 1,411-gene module is preserved in our data; and within pre-frontal cortex most significantly overlaps PFC-BW-M4 (OR=4.2, $p<10\text{-}4$) followed by PFC-CTX-M3 (OR=2.9, $p<10\text{-}3$). Though these overlaps comprise a small (159 BW-M4 + 87 CTX-M3) proportion of the module, the implication is that a substantial proportion of the genes in FromerM2c reflect brain-wide relationships. This is underscored by the fact that CTX-M3 itself is not confidently region specific, but shows moderate to strong evidence of preservation across the brain.

Figure 3.1 shows the overlaps between published disease-relevant modules and the atlas BA9 modules. Of the 59 published modules, there are only 2 instances of an overlap with a region-specific module that did not *also* show a significant overlap with a whole-brain module: Parikshak2016-M4 and Harlywycz2016-M19, both of which show overlap with BA9-M8. However, as the module preservation statistics show strong evidence for this module across all non-cerebellar tissues, it is likely that it reflects a brain-wide process. BA9-M8 enriches for markers of noradrenergic and cholinergic neurons, both of which are found in all non-cerebellar brain regions (**figure 3.2**). This gene set likely represents a population of neuronal subtypes which are variable enough in BA9-M8 to have a co-expression signature, but which otherwise contribute to the brain-wide neuronal module BW-M4. Figure 3.1 also shows a common set of brain-wide modules involved in overlaps across every study: BW-M1, BW-M4, BW-M6, and

BW-M10, corresponding to the major cell types in the brain. While region-specific effects may play a role in neuropsychiatric disease, these findings show that no clear region-specific signature has yet been identified, but instead that a consistent, brain-wide signature of neuronal dysfunction and glial dysregulation underlies previous findings. Furthermore: while the pathways involved in genetic risk are broad, the impact of neuronal dysfunction may be more acute in certain regions, such as the cortex where neurons are more prevalent. Indeed Parikshak *et al.* (Parishak2016) observe the same dysregulated genes in cortex and cerebellum, but that the cortex is more severely dysregulated. In this way, we might expect to identify other instances of regional pathology that do not involve region-specific molecules or pathways.

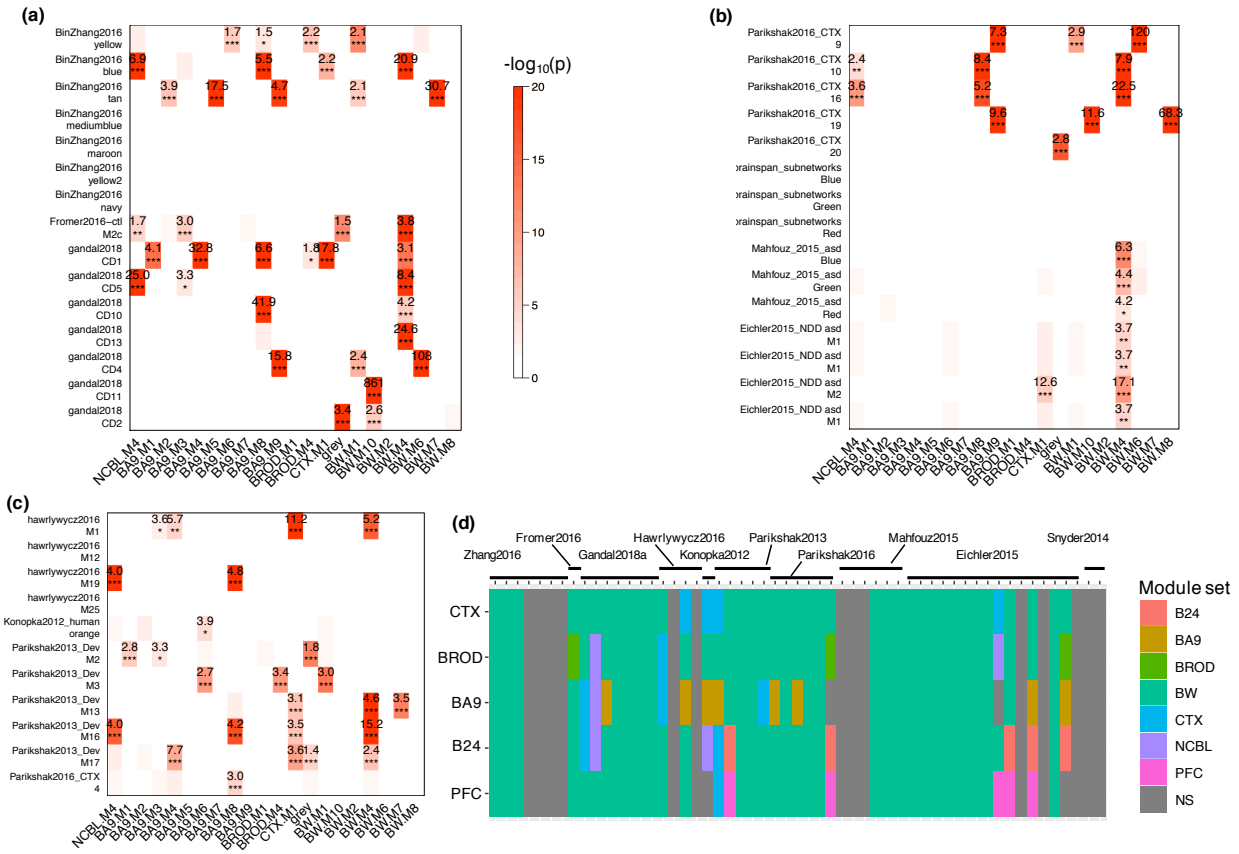

Figure 3.1 Whole-brain co-expression drives most neuropsychiatric disease modules

(a-c) Module overlaps, labeled by odds ratio and colored by *p*-value, between disease-implicated modules from other studies, and the brain-wide and regional (cortical) modules in GTEx. All heatmaps share the same color key. The vast majority of overlaps are with neuronal modules (BW-M3, BW-M4, BW-M5, BW-M9, CTX-M1, BA9-M8). **(d)** Summary of overlaps by region, demonstrating a large amount of overlap with brain-wide module sets.

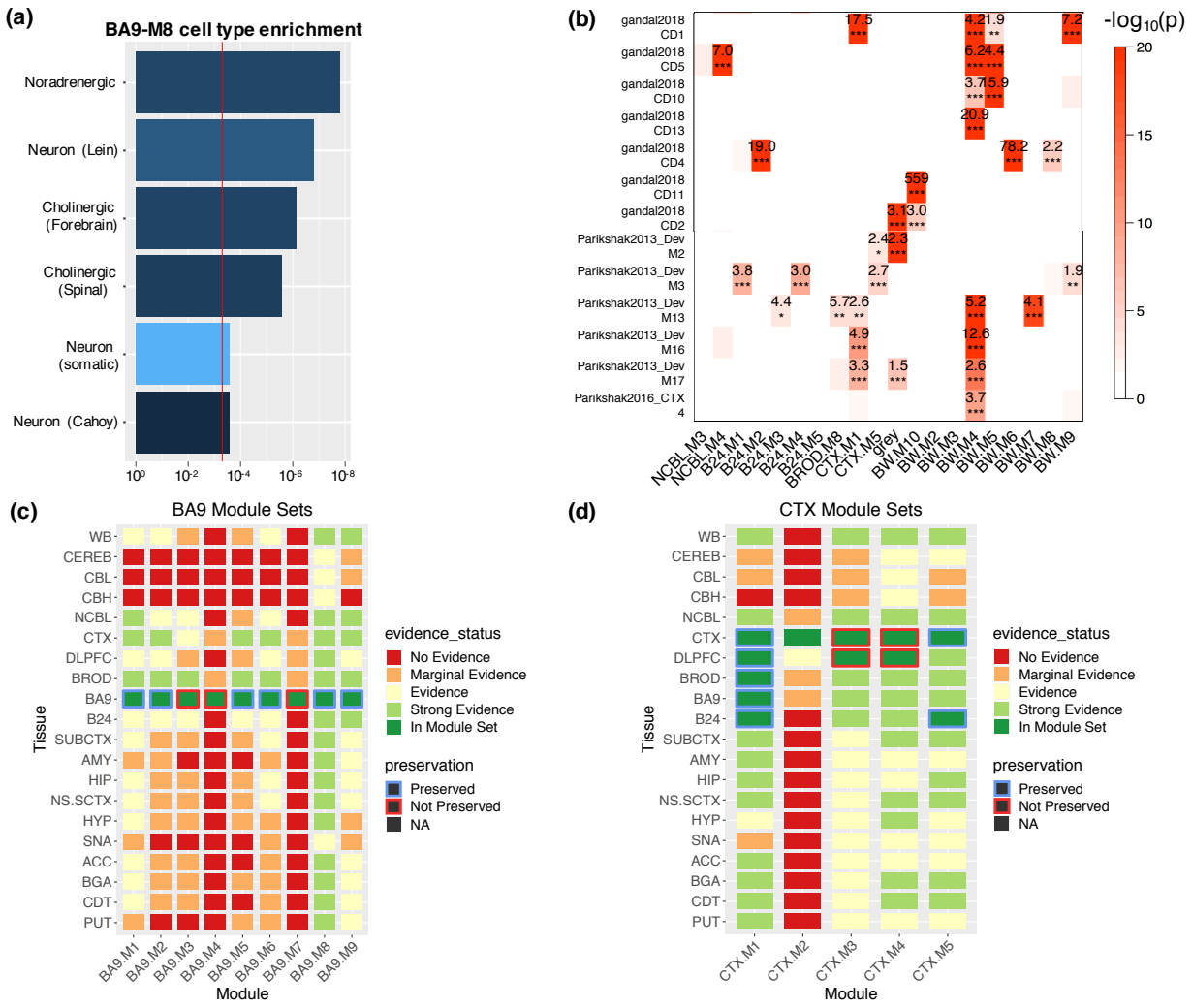

Figure 3.2 BA9-M8 and CTX-M1 show whole-brain preservation, related to 3.1

(a) Cell-type enrichment for module BA9-M8 showing enrichment for neuronal subtypes that are present throughout the brain. Lein, Cahoy represent published neuronal gene markers (Lein2007, Cahoy2008); somatic represents genes expressed on somatic neuronal components, and Spinal and Forebrain represent markers of neurons from these brain regions. **(b)** Zoom-in on overlaps for cross-disorder and developing brain modules colored by p-value and labeled by odds ratio. Significance at 0.005, 0.001, and 0.0001 given as *, **, *** respectively. **(c, d)** Region-specificity metrics for BA9 (c) and CTX (d) specific modules, demonstrating that BA9-M8 and CTX-M1 show strong evidence in the telencephalon and striatum.

3.3b Convergence of molecular signatures of neuropsychiatric disease onto brain-wide neuronal modules

I next investigated whether genetic perturbations in neuropsychiatric disease converge onto region-specific or cross-regional modules. Utilizing databases of *de-novo* variants implicated in ASD and SCZ (Tychele2016), GWAS summary statistics (PGC2013, PGC2014, PGC2017, Grove2019) and RNA-sequencing in post-mortem ASD and normal brains (Parikshak2016), I identified two whole-brain modules, BW-M4 (neuron) and BW-M1 (neural progenitor), that simultaneously enrich for ASD-linked rare variants, enrich for SCZ GWAS signal, and that manifest disrupted expression in ASD post mortem brain. I also observed two regional modules, CTX-M3 (activity-dependent regulation and endocytosis) and CEREB-M1 (mRNA binding) that show ASD rare-variant and SCZ GWAS enrichment. While the co-expression relationships for CTX-M3 and CEREB-M1 are distinct, the genes do overlap more often than expected by chance (Jaccard=383/1938, OR=9.5, $p<10^{-20}$ Fisher's Exact), and both show significant preservation in control brain, but not in ASD post mortem brain, suggesting that they are disrupted in ASD.

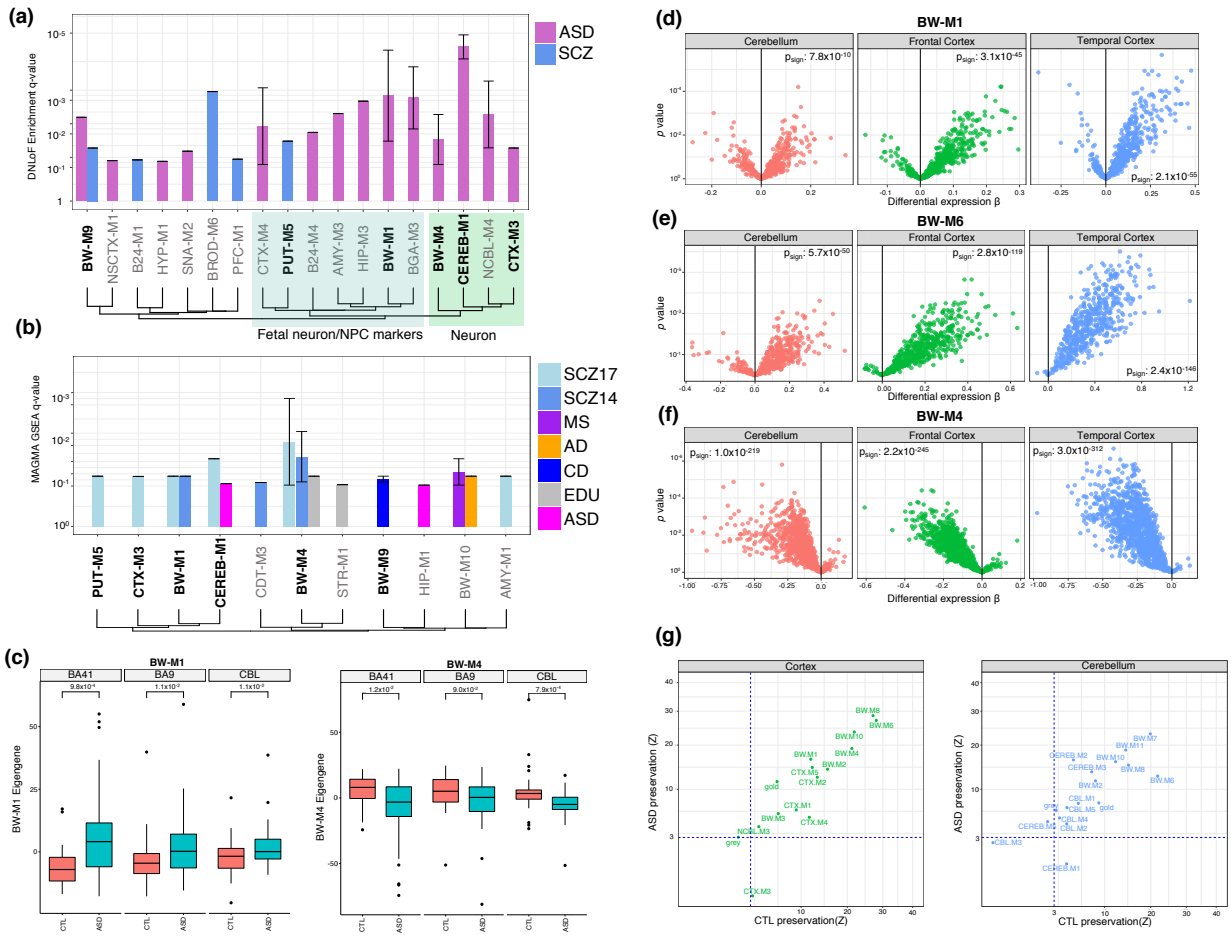

Figure 3.3 Gene-level module enrichments for de novo PTVs, GWAS summary statistics, and differential expression

(a) FDR values (Fisher exact test) for enrichment of de-novo loss-of-function variants within modules, summarized to module sets. Bar height gives geometric mean of FDR, and whiskers the range of (significant) FDR values for modules within the module set. Bold modules also show enrichment from GWAS summary statistics. Module sets are ordered by Jaccard similarity between their index modules. Green region: These modules enrich for neuronal markers. Blue region: These modules enrich for fetal neuron, mitotic progenitor, or outer radial glia markers. **(b)** FDR values (MAGMA) for GWAS summary statistics within modules. Method of ordering identical to (a). **(c)** Module eigengene expression for BW-M1 and BW-M4 in ASD cases and control brains across three regions and associated p-values from a T-statistic (linear model including covariates as in Parikshak2016). **(d-f)** Volcano plots and sign-test P-values for genes in NPC, astrocyte, and neuronal modules. **(g)** Module preservation statistics separately in ASD and control brains, suggesting differential preservation for modules CTX-M3 and CEREB-M1.

BW-M4 represents a non-specific neuronal gene set, identified independently throughout the telencephalon and subcortical regions, all sharing GO terms related to membrane organization or ion transport. Examination of significant genes within BW-M4 (defined as a Z-score from MAGMA > 3.0, **methods**) demonstrates enriched terms for both the Psychiatric Genomics Consortium (PGC2014) and ClozUK (Pardinas2018) SCZ GWAS studies are related to the synapse and synaptic transmission. This suggests a convergence of risk genes onto synaptic signaling pathways, consistent with a recent comprehensive pathway analysis (Schijven2018). Using meta-GSEA (**methods**) to rank ontologies across GWAS studies and brain regions, both ASD and SCZ appear to share highly ranked terms related to synapse assembly and plasticity. Interestingly, the term synaptic transmission shows very strong evidence only from SCZ association statistics, whereas the strongest terms with evidence in ASD alone are learning and social behavior. (**figure 3.4)**

BW-M1 contains genes and pathways corresponding to neurogenesis, differentiation, and migration (**figure 3.5a**), as well as components for RNA splicing, structural components of cell division, and stem cell population maintenance (**figure 3.5b**). Genes within BW-M1 are strongly loss-of-function intolerant, and the module enriches strongly for PPI interactions. The genes in this module, which are up-regulated in ASD cortex, enrich for the TGF-beta signaling pathway (FDR=0.0047, STRiNG; Szklarczyk2016), which is known to regulate neurogenesis (Battista2006), and consist mainly of the BMP/SMAD pathway (*BMPR1A*, *BMP2K*, *SMAD4*, *SMAD5*, *SMAD9*) which is critical for orchestrating proliferation/differentiation balance (Jovanovic2018). NPC proliferation/differentiation balance is another major theme of the module, as it contains key *REST* co-repressors *CTDSPL* and *RCOR1*, the down-regulation of

which promote proliferation over differentiation (Monaghan2017), as well as the differentiation repressors *ADH5*, *TLR3*, *SOX5*, *SOX6*, and *PROS1* (Wu2014, Okun2010, Lathia2008, Martinez-Morales2010, Lee2014, Zelentsova2016) and the differentiation/proliferation regulator *SPRED1* (Phoenix2010). The *de novo* LoF and GWAS enrichments suggest that NPC proliferation and differentiation – both in the prenatal and adult brain – are disrupted in ASD and SCZ. De la Torre Ubieta *et al.* recently observed that fetal-specific open chromatin regions in the cortical plate enrich for SCZ heritability, and reflect regulatory elements for genes involved in neurogenesis (De la Torre-Ubieta2016). Analysis of module trajectories in the developing brain (Sunkin2012) shows very strong prenatal upregulation, with continuing post-natal activity into early adulthood. Disruption of this module may be responsible for the observed ASD expression signature – downregulation of neuronal modules and upregulation of astrocyte modules – implicating brain-wide changes in neuronal proliferation/differentiation balance beginning in early development and persisting into adolescence.

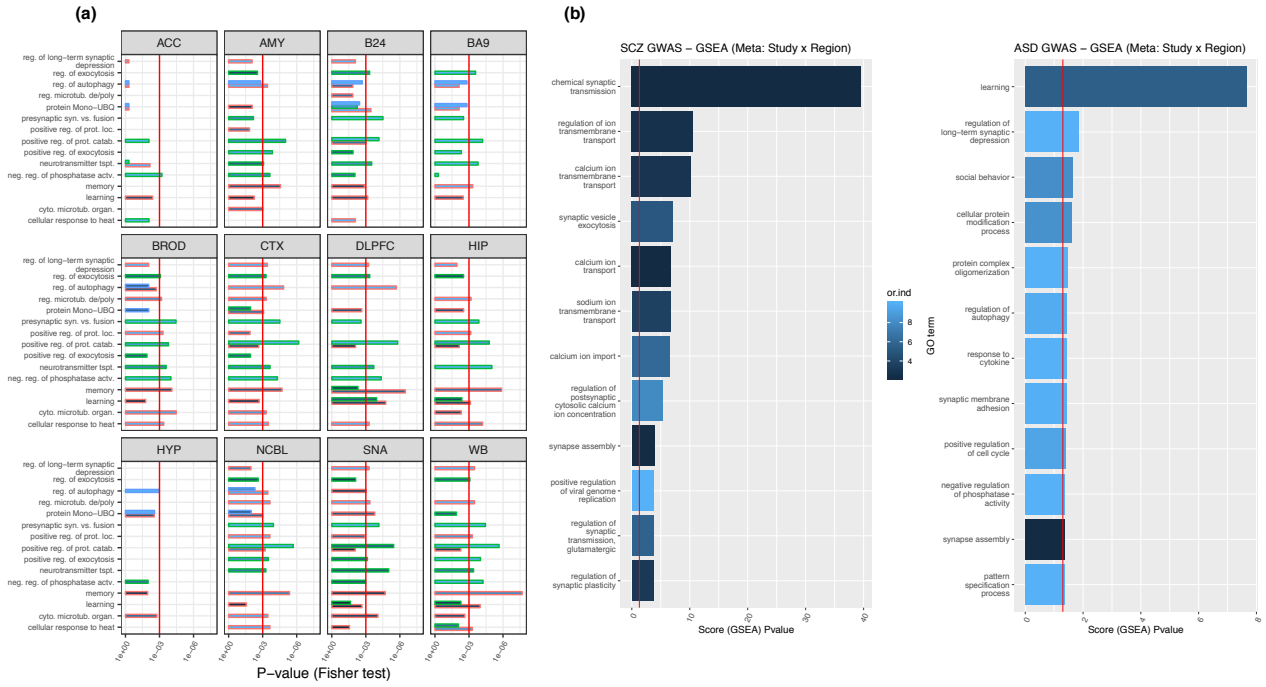

Figure 3.4 Meta GSEA of nominally significant genes

Meta GSEA of nominally significant genes (z > 3 SCZ, z > 2.5 ASD; MAGMA) in BW-M4 across three studies: iPsych (ASD), PGC (SCZ), and ClozUK (SCZ). **(a)** Significantly enriched GO terms within these significant BW-M4 gene sets. Length of the bar reflects significance value, the color the odds ratio of enrichment (scale shared with (b)), and outline color the particular study: PGC (green), ClozUK (red), and iPsych (blue). **(b)** Combined enrichment scores and odds ratios across ontology terms for SCZ and ASD. The P-values rely on an assumption of independence which is violated, and should be interpreted as a ranking or score as opposed to the result of a statistical test.

3.3c Neuropsychiatric disease risk enriches in cortical and cerebellar modules which are differentially co-expressed in ASD brains

The two regional modules that show convergent evidence of disruption in neuropsychiatric disorders, CTX-M3 and CEREB-M1, show an enrichment for *de novo* LoF variants linked to ASD, an enrichment for SCZ GWAS summary statistics, and are disrupted in ASD (**figure 3.3**). Although, they have distinct components, both CTX-M3 and CEREB-M1 show significant overlap in their genes, and modest evidence of preservation outside of their respective regions, with preservation AUPR scores < 0.5. Both modules enrich for PPI (CEREB-

M1 p<7e-15, CTX-M3 p<0.0023), as well as LoF-intolerant genes, providing validation that they contain coherent and essential biological pathways.

To validate the cortical-specificity of CTX-M3, I used normalized RNA expression values from the Allen Human Brain Atlas (Shen2012) to contrast the expression trajectories of hub genes in cortical versus non-cortical regions. I observed that the relative levels of these genes across all cortical regions (frontal, parietal, occipital, and temporal lobes plus cingulate gyrus) are tightly coupled in contrast to their highly variable expression across non-cortical regions (hippocampus, hypothalamus, striatum, and cerebellum), evidence of a preserved co-expression signature across the cortex (**figure 3.5**).

Notably, CTX-M3 contains both the syndromic ASD gene, *FMR1*, as well as its direct interactor, the protein *NUFIP1*, which (like *FMR1*) has been implicated in the regulation of activity-dependent translation and local synaptic translation (Bardoni2003) and additionally in ribophagy (Wyant2018). It also contains the intellectual disability (ID) gene *ATRX*, which forms a complex with the protein product of *DAXX* to regulate H3.3 loading onto and maintenance within heterochromatin. H3.3 is itself associated with activity-dependent transcription in neurons (Maze2015), suggesting that dysfunction or dysregulation of *ATRX* could alter the availability of this activity-related histone.

To further whether this module's disruption was related to changes in activity dependent processes in ASD, I examined genes previously identified as up-regulated following activity induction of rat hippocampal neurons (Schanzenbacher2018), and found that 10% of these genes fall into CTX-M3 (p = 0.0472, Fisher Exact). The observed enrichment is driven largely by components of protein phosphatase 1, which has both nuclear and synaptic roles in synaptic plasticity and long-term memory (Hu2007, Koshibu2009). However, several components of the

mitochondrial ribosome (*MRPL27*, *MRPL45*, *MRPS26*) are observed concomitant with activity-dependent upregulation, and CTX-M3 is highly enriched for the mitochondrial ribosome, containing 21 genes within this functional pathway ($p < 1.7e-10$, Fisher Exact). These observations indicate that activity-dependent up-regulated genes form one component of this ASD- and SCZ-associated module, CTX-M3. Other components of this module include poly-A binding, alternative polyadenylation and alternative splicing (*NGDN, MBNL1, MBNL2, CSTF3, SPSF3, CPSF6*), multiple endocytosis regulating genes (*RALA, VAMP4, VAMP7, TSG101, VPS25, RAB18, RAB3GAP2, CHMP2B*, and sorting nexins *SNX2, SNX3, SNX13, SNX14*), consistent with their role in supporting neuronal activity-dependent processes that are disrupted in ASD (Tsai2012, Quesnel-Vallieres2016, Chen2017, Ip2018, Sears2018, Stilling2018).

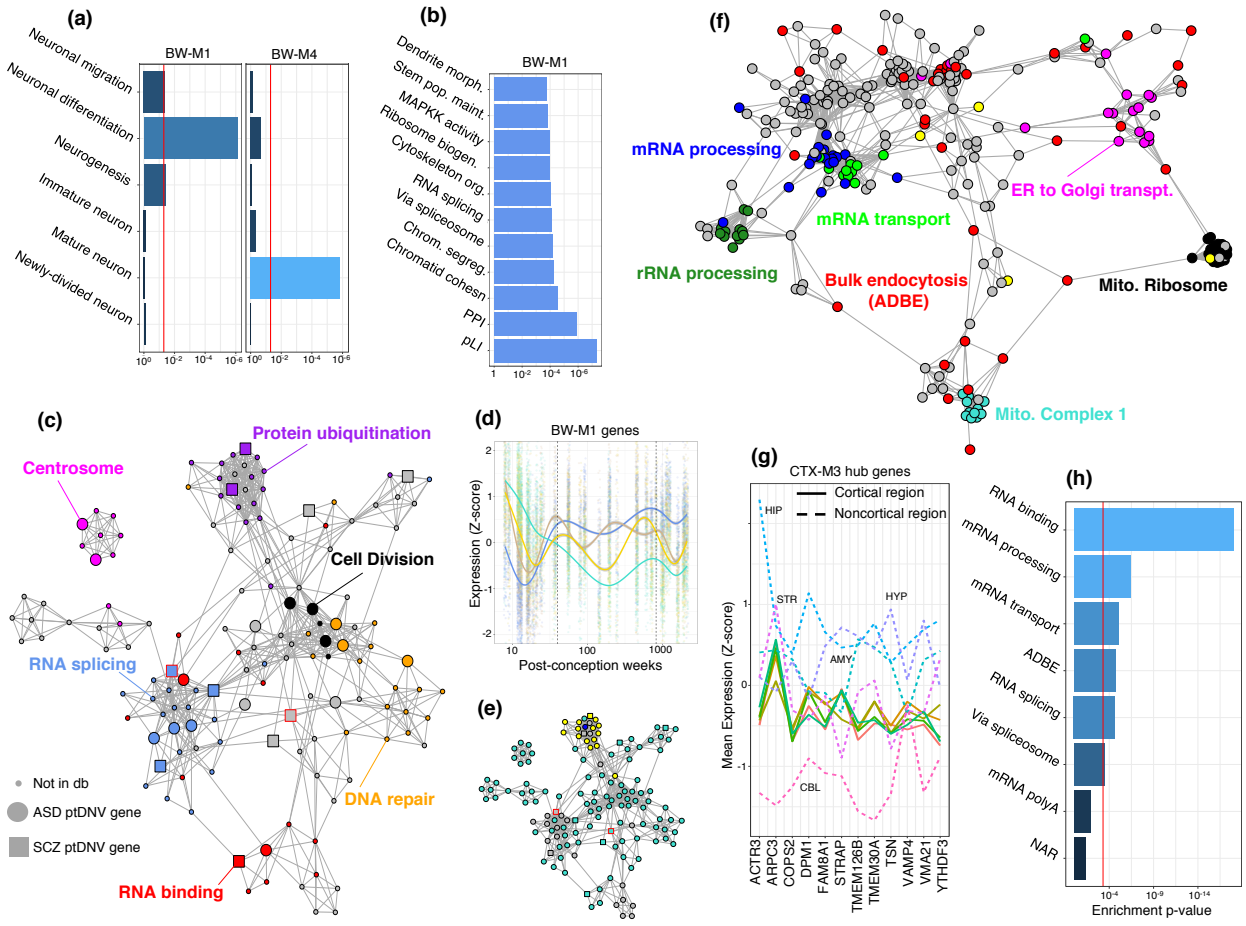

Figure 3.5 Ontologies, PPI networks, and expression profiles of ASD-associated modules.

(a) Enrichment p-values (Fisher exact test) for neuron-related ontologies in whole-brain modules. **(b)** Combined (geometric mean) enrichment p-values of ontologies for all modules in module set BW-M1 that showed enrichment for ASD-implicated de novo loss of function mutations. **(c)** Coexpression-PPI network of BW-M1, highlighting denovo loss of function mutations (large nodes) and ontologies (colors). **(d)** Expression of BW-M1 across developmental timepoints, sub-clustered into four modules using WGCNA. **(e)** Assignment of network nodes in (c) to the subclusters in (d) via label propagation. **(f)** Coexpression-PPI network for CTX-M3, colored by enriched gene ontology sets. **(g)** Expression profile of CTX-M3 hub genes across brain regions, demonstrating tight co-regulation in cortical regions (solid lines) by virtue of small variance, and variable expression across non-cortical regions (dashed lines). **(h)** Enrichment p-values (Fisher exact test) of the CTX-M3 module for gene ontologies, including bulk endocytosis genes.

(b)

3.4 Discussion

The findings that ASD-linked dnLoF mutations as well as SCZ GWAS signal enrich in brain-wide neuronal and neurogenesis modules underscore previous findings linking both common and *de novo* variation to synaptic genes (Pers2016, Alonso-Gonzales2018), neuronal genes (Skene2018), developmentally-expressed genes (Wang2019), and neurogenesis pathways (Yuen2016). One prevailing synthesis is that the transcriptomic signatures in ASD reflect progressive response to a primary insult to neuronal maturation, synapse formation, and early childhood synapse stabilization and pruning (Parikshak2016), which should give wide to brain-wide disruptions. Even though cortical regions contain a much higher proportion of neurons than other brain regions, the pattern of risk variant enrichment is not cortex-specific, and enrichments in the regional BW-M1 and BW-M4 modules are significant across the brain, implying widespread effects of these genetic risk variants on brain function.

The only region-specific modules with convergent evidence across disease and modality were CTX-M3 and CEREB-M1, which appear to reflect activity-dependent transcriptional profiles. Indeed, *VAMP4* – present in CTX-M3 – is an essential molecule for activity-dependent bulk endocytosis (ADBE) (Nicholson-Fish2015), and several module proteins (including RAB GTPases *RAB7a* and *RAB18*) overlap with the ADBE proteome (Kokotos2018). A parsimonious explanation is that this module concerns the maintenance of organelles and proteins required for long-term neuronal activity, (i.e., mitostasis and ADBE proteostasis), through activity-dependent mRNA transcription and neuropil targeting (Caleb2014). This represents an axis of neuropsychiatric genetic architecture that pertains to neurotransmitter release and reuptake. Notably, genetic insults to ADBE should have different impacts on excitatory and interneurons because GABAergic synaptic terminals are more prone to ADBE-reuptake than non-GABAergic

terminals (Wenzel2012). Thus, impairments to CTX-M3, including bulk endocytosis, at inhibitory synapses may play a role in the excitatory/inhibitory imbalance that is observed in neuropsychiatric disease (Selten2018).

3.5 Methods

Module Overlaps

Network modules from previous publications were obtained from main tables or supplementary information, converted to ENSEMBL gene IDs using the 'biomaRt' R API to the grch37 ensembl server.

To address ascertainment bias, enrichment was calculated using only genes overlapping between any two studies. This overlap set is used to compute the contingency table for two modules (in neither, in both, in only this study module, in only the comparison study module), and Fisher's exact test is used to obtain a p-value and odds ratio.

Enrichment tests for ADBE, neuronal migration, neuronal differentiation, and neurogenesis were performed in the same way. However, for ADBE and neurogenesis, the original publications did not publish the full set of ascertained genes, so the background was the entirety of our data. However, as these are all neurological gene lists, it is not likely that the ascertained set of genes were substantially different from the set of brain-expressed genes ascertained here.

De-novo variant enrichment

Denovo-DB (Tychele2016) was used to extract lists of genes harboring *de novo* variation linked to ASD and Schizophrenia. The v1.5 of the database was obtained on 02-17-2018, and we filter for "PrimaryPhenotype=autism" (or, separately, "PrimaryPhenotype=schizophrenia") and

"FunctionClass" as one of "frameshift", "frameshift-near-splice", "splice-acceptor", "splice-donor", "start-lost", "stop-gained", "stop-gained-near-splice", or "stop-lost."

Module enrichments are calculated via Fisher's Exact Test, using the contingency table formed by cross-tabulating module presence/absence with presence/absence on the denovo-db gene list.

As the denovo-db is a broad collection of *de novo* mutations in affected individuals and does not curate these variant lists on the basis of total evidence, we consider two additional data sources for alternative enrichment scores. First, there is the curated list of SFARI genes of rank S, 1, 2, or 3; and perform enrichment on the resulting list. Second, recent work from our lab (Ruzzo2019) computed transmission and de-novo association Bayes Factors for 18,472 genes. As an additional confirmation, the log Bayes Factor was regressed against module presence/absence and the coefficient tested against 0, with a positive alternative.

GWAS variant enrichment

Enrichment for GWAS signal was performed through the use of MAGMA (deLeeuw2015) gene set analysis. Briefly, variants were mapped to genes on the basis of genomic distance, while taking chromatin contact maps from adult brain Hi-C (Won2016) into account. MAGMA was used to generate gene scores and LD-based covariance. Subsequently, MAGMA's gene set analysis was used to compare the distribution of gene scores between modules and the background set of 'grey' genes.

5 GWAS studies were considered in this analysis: The iPsych (Anttila2018) and PGC (PGC2013) cross-disorder GWAS studies (accounting for ASD, SCZ, and cross-disorder), the

IGAP (Lambert2013) consortium's Alzheimer's disease study, the KKNMS (Andlauer2016) multiple sclerosis GWAS, and educational attainment from the SSGAC (Okbay2016).

Differential preservation analysis

Modules defined in the GTEx tissue samples were assessed for their preservation in ASD case samples and (separately) in normal samples according to WGCNA's module preservation statistics (Langfelder2011). These produced a pair of preservation Z-scores per module. Differential preservation of a module is a case where the control Z-score is preserved (>3) while the ASD Z-score is not preserved (<3).

Chapter 4 Network genetic architecture and the omnigenic disease model

4.1 Abstract

Genes do not operate in isolation but function through complex interaction networks. The impact of a mutation may therefore depend on its genomic context, a property termed "epistasis." With the exception of twins, the genomic context of a mutation is different in each individual. This means that, in population studies such as GWAS, the effect size of a mutation reflects both a baseline (additive) effect, as well as the average epistatic effect. These two components cannot be separated, meaning that epistatic effects are necessarily absorbed into estimates of additive effects (Sackton2016). By capturing epistatic interactions (and therefore average epistatic effects), biological networks shape the additive effect size distribution.

In this chapter, I incorporate network structure into a model of genetic architecture, termed "network genetic architecture." I demonstrate the utility of this framework by distinguishing the hypotheses of omnigenic and polygenic architecture in terms of a key model parameter, γ_2, and also establish that co-expression networks significantly enrich for the heritability of schizophrenia (SCZ) and autism spectrum disorder (ASD).

4.2 Introduction

Complex genetic traits, characterized by causal mutations affecting several hundreds to thousands of genes, defy reductionist understanding in terms of a single gene or even a single pathway. Systems biology, by abstracting biological function to the coordinated action of many genes and other molecules, has proved to be a powerful organizing framework, leading to deeper

understanding of disease in terms of disrupted molecular relationships (Konopka2009, Willsey2018).

Gene networks in particular have demonstrated immense utility in identifying gene-gene relationships that contribute to disease etiology or manifestation. Recent work has shown that gene networks relate to genetic architecture, as many network modules enrich for rare and common disease-associated variants (Parikshak2015). Yet a complete and coherent interpretation of this relationship requires incorporating network architecture into models of heritability (Kim2019).

Two recent syntheses of the past decade of genetic discoveries in human traits – the omnigenic model of Boyle and Pritchard (Boyle2017), and an associated rebuttal by Wray and Visscher (Wray2018) – appeal to gene networks to structure functional genetic effects. In the case of the omnigenic model, gene networks are used to distinguish important "core" genes from effector "peripheral genes", while in the case of the polygenic model, gene networks capture the underlying complexity.

In this chapter, I present a model of network genetic architecture out of which omnigenic and polygenic architectures arise as special cases. From the omnigenic special case, I derive expectations for the network distribution of high-penetrant *de novo* variants, and show that a wide range of co-expression networks and gene regulatory networks are not consistent with this expectation, and therefore do not reflect an omnigenic structure. I generalize this model to arbitrary graphs defined on genes or variants, and provide a method for estimating heritability explained and genomic enrichment. Using this method, I demonstrate that co-expression network structures corresponding to receptor signaling, synaptic vesicle function, and pyramidal neurons, capture significantly more heritability for both ASD and SCZ than is expected by chance.

Finally, I derive a statistical test for network perturbation: whether a *trans*-QTL impacts a network so that genetic liability for a disease is increased. Using this model, I demonstrate that tens of modules – and potentially hundreds of genes – *directly* contribute to the etiology of neuropsychiatric and neurodegenerative disease, which is consistent with a genetic network architecture closer to the polygenic spectrum than the omnigenic spectrum.

4.3 Results

4.3a Likely high-penetrance ASD mutations do not exhibit omnigenic network enrichments

To evaluate the omnigenicity of neuropsychiatric disease, I reasoned that high-effect-size mutations should occur almost entirely within core genes. To assess this hypothesis, I simulated variants from a genetic architecture where the effect size is a function both of frequency and the network distance:

$$\beta_i | d_i(G), p_i \sim N(0, \sigma_g^2 \left(2p_i(1-p_i)\right)^{\gamma_1} \left(1 + \delta d_i(G)\right)^{\gamma_2})$$

I take the core distance, d_i, to be normalized to [0, 1]. I can then group variants into deciles, with D_1 reflecting those variants with d_i<0.1. Across a wide range of values for d_i, γ_1, and δ, I observed that whenever variants in D_1 explain >40% of heritability, there is a strong, decreasing trend between effect size and core distance, such that the vast majority of high-effect variants fall within D_1.

To quantify this observation, I defined a simple statistic, ϕ, which measures the fraction of the 1% highest-effect mutations that fall into D_1. For omnigenic values of γ_2 (≤-5), this value ranges from 45% to 100% implying that, for this model of network architecture, half to nearly all of high-effect mutations should occur within or very close to core genes. To validate the reasonableness of the simulation, I computed the effect size ratio of the 1% highest-effect variant

to the effect size of a typical GWAS variant (80% power to detect in a balanced sample size of 10,000), and found that this ratio ranged from 5x-80x, consistent with estimates of the relative risk ratio between *de novo* loss-of-function (dnLoF) variants and case-control-associated damaging variants of 3x-25x (Nguyen2017, Ballouz2017).

I reasoned that genes associated with a disease by an excess occurrence of dnLoF mutations are likely to be within the 1% highest effect sizes of all causal variants. As such, the proportion of dnLoF-implicated genes falling into D_1, termed $\hat{\phi}_{dnLoF}$, is an estimate for ϕ. Because ϕ is robustly large across a wide range of omnigenic architectures, a low value for $\hat{\phi}_{dnLoF}$ implies that the disease is not omnigenic *with respect to* the distance *d(G)*. I emphasize that $\hat{\phi}_{dnLoF}$ is specific to a network *G* and a distance function on that network *d*.

A set of 10,000 genes admits nearly 50 million unweighted networks, and as many distance functions, making network architecture – including omnigenics – non-falsifiable in practice. Yet *specific* instances of network architecture (e.g. co-expression graphs with module hubs as core genes, for which $\gamma_2<0$) almost surely exist. A high value of $\hat{\phi}_{dnLoF}$ for a network *G* would demonstrate one such instance, and further provide strong evidence that γ_2 is extreme ($\leq$ -5) with respect to *d(G)*. Biologically, this would indicate that the genes with the lowest value of *d*(*G*) are core genes: mutations in these genes have very large effect sizes, and thus the genes themselves play an important role in disease processes.

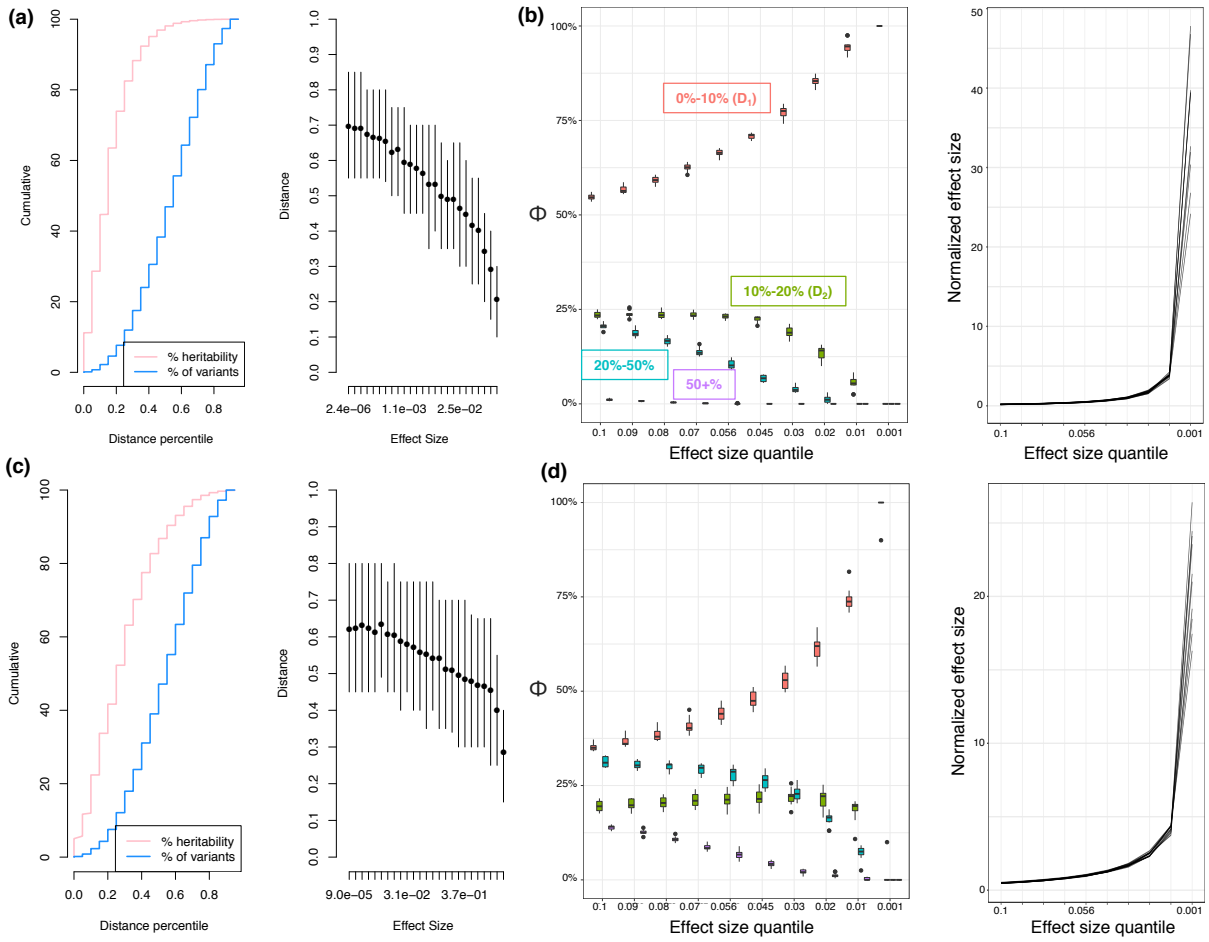

Figure 4.1 Simulation from omnigenic settings of network architecture

(a) *left.* Cumulative distributions of h^2 explained and total variant proportion by network distance for a network architecture with δ=1.8, γ_2=-10, γ_1=-0.4. *right.* Distribution of variant network distances by its effect size – higher effect variants falling closer to or within the core. **(b)** *left.* Distribution of ϕ across 10 simulations of the architecture in (a), across a range of thresholds including the final threshold 0.01. *right.* Relationship between the effect size quantile and the normalized effect size (effect size ratio versus a well-powered 5%-frequency GWAS SNP). **(c)** As (a), but for δ=2, γ_2=-5, γ_1=-0.4. **(d)** As (b), but for the architecture reflected in (c).

Recent work in our lab identified 69 ASD-associated (FDR<0.1) genes by integrating *de novo* and inherited LoF mutations (Ruzzo2019), and a previous integrative publication identified 64 ASD-associated genes on this basis (Nguyen2017). Using the Bayes factors from these studies, I computed $\hat{\phi}_{dnLoF}$ values for ASD across a large variety of gene networks: co-expression networks in DLPFC, whole cortex, developing cortex, fetal cortex, with whole blood as an outgroup, brain PPI networks, and transcription factor binding networks in whole cortex and NeuN+ cells. I found little difference when taking *d* to be shortest path distance, mean path distance, or negative module kME as distances; or when using a 10-nearest-neighbor network in place of the fully-connected weighted co-expression network. The core genes, to which distances are measured, are taken to be (i) the top 5% (minimum 5) module hub genes (ii) the top community exemplars (TF binding network); (iv) the syndromic ASD genes *FMRP, ANK2, SYNGAP1, CHD8, SHANK2, SHANK3,* and *SCN2A* (all networks); or the top 10 or 20 genes identified from the above studies (all networks).

Across all of these networks, the largest observed value of $\hat{\phi}_{dnLoF}$ = 54% from the whole-blood co-expression network; fetal, developing, and adult cortex co-expression, as well as transcription-factor-binding networks, all showed $\hat{\phi}_{dnLoF}$ values between 10% and 48%. I take this to reflect on the architecture of ASD with respect to these robust, relatively definitive, co-expression networks, suggesting that RNA co-expression does not reflect an omnigenic model. Within these networks it is likely characterized by a moderate value of γ_2. However, it could

also be that either the FDR values for the two association studies are in fact higher than 10%, or that the dnLoF-implicated genes do not reflect the top 1% (or even 5%) of variants by effect size. Though there may be networks other than co-expression for which $\gamma_2$takes on omnigenic values, the absence of large $\hat{\phi}_{dnLoF}$ values for TF bindng networks and PPI networks also do not reflect an omnigenic architecture.

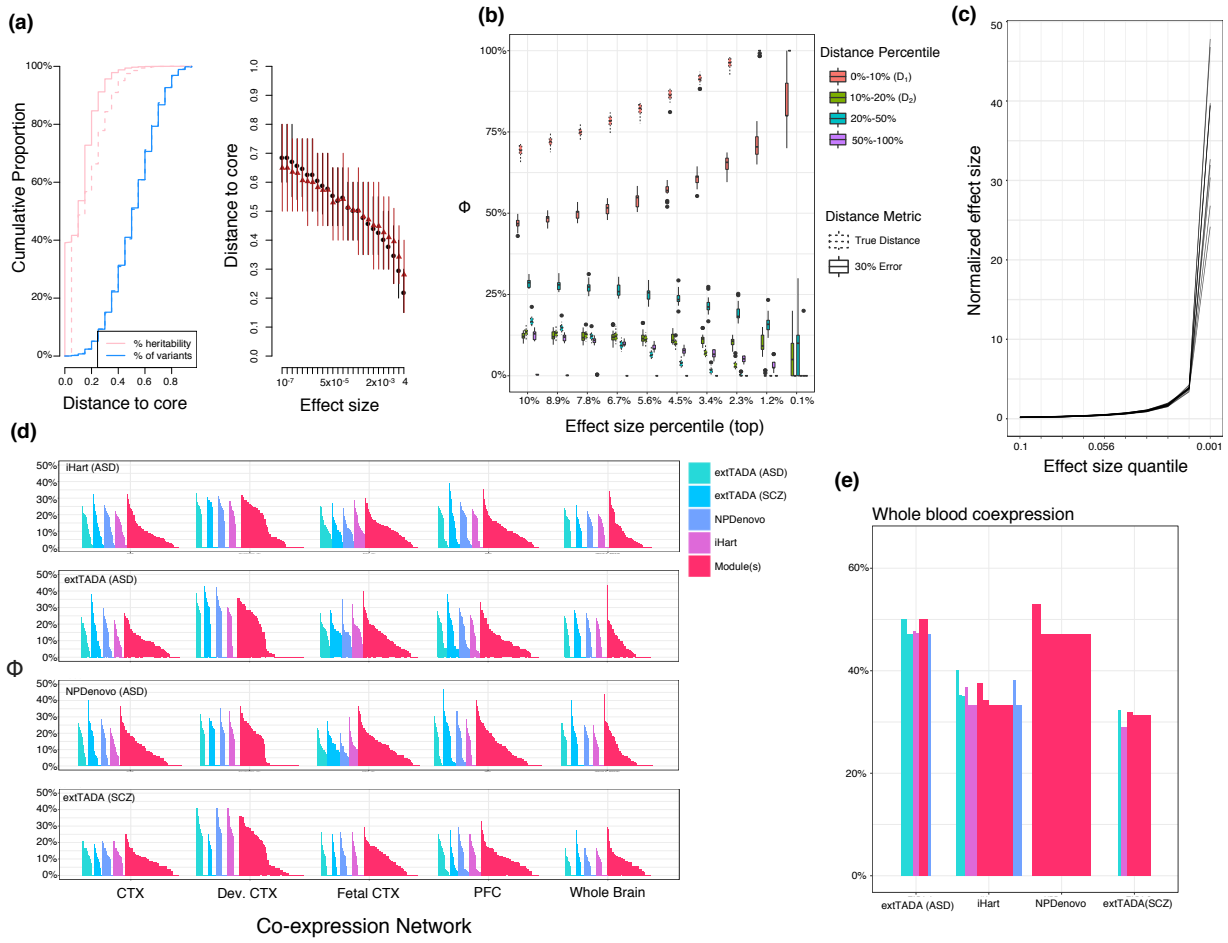

Figure 4.2 - Characterizing core-periphery structure of high-impact neuropsychiatric disease genes across multiple networks.

(a) Example simulation of network genetic architecture, where the variant effect size decays rapidly with distance to core. *Left*: Cumulative proportion of genes (blue) and heritability (pink) along the distance distribution. Dotted line shows the cumulative heritability when true distance is replaced by a corrupted (30% error) distance. *Right*: The relationship between core distance and effect size results in high-effect variants only appearing very close to core genes for both ground truth (black) distance as well as 30% corrupted (brown) distance **(b)** High-impact genes are defined by the effect-size percentile on the x-axis, and the % of genes falling into the core-distance decile is plotted on the y-axis. This plot encompasses 20 simulations. Dotted boxes represent the expected values for Φ when the distance is error-free, while solid boxes represents the case where distance is 30% corrupted by error. **(c)** Validation of the effect size distribution: the effect size of each quantile is normalized to the effect size for which a

balanced GWAS of 10,000 samples has 80% power; the highest-impact variants are only 20-50x stronger than empowered variants. **(d)** All values of Φ across distance metrics, core set size, module definitions, and brain co-expression networks, demonstrating that no value of Φ exceeds 50%. **(e)** top 10 Φ values (per core set) for the GTEx whole-blood co-expression network.

The expectation that $\hat{\phi}_{dnLoF}$ is large assumes that the core genes that define $d(G)$ are known exactly, with no missing core genes or extraneous peripheral genes. The extent of interpretation of the previous results depends on how sensitive $\hat{\phi}_{dnLoF}$ is to false-positives or false-negatives within the hypothesized gene set. To quantify the sensitivity of $\hat{\phi}_{dnLoF}$ to the distance function d I repeated the original simulations in two new scenarios: (i) using d_{true} to define the genetic architecture, and a separate but correlated d_{meas} to calculate ϕ and (ii) using a co-expression network to define d_{true} to a set of 50 core genes, and introducing false-positives and false-negatives into this gene set when computing d_{meas}.

Surprisingly, the larger γ_2 is in magnitude, the more robust ϕ is to errors – a result of the fact that for large enough values of γ_2, most high-impact variants fall within the lowest 1% of distance, so these variants remain in D_1 even in the face of high error. ϕ is also quite robust to core-gene false-positives, because this introduces excess low-effect variants which do not enter into the calculation of ϕ. The robustness of ϕ with respect to false-negatives is a function of how clustered the core genes are within the network. With 10 clusters – representing 10 distinct biological processes – the expectation of $\phi > 50\%$ can still withstand up to a 20% core set false-negative rate, while with 5 clusters this increases to 35%.

For an omnigenic architecture where core genes are clustered, ϕ is robust both to measurement errors in the distance, and false-positives in the core gene set. This implies that our conclusions about co-expression and regulatory networks can be made fairly strong: for ASD, γ_2 is unlikely to fall within the omnigenic range for these networks.

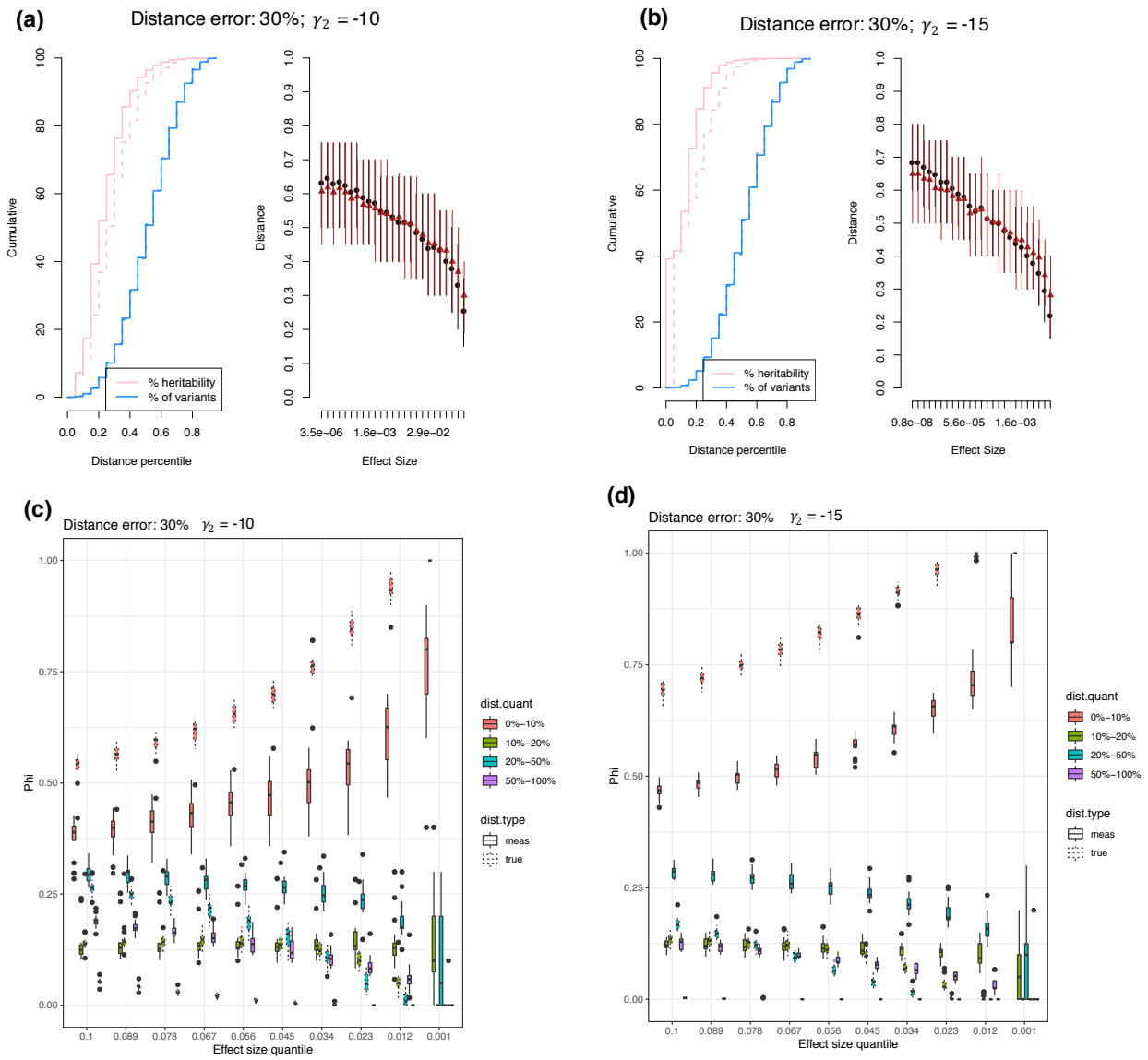

Figure 4.3 Tolerance of phi to distance error in omnigenic architectures

(a,b) Cumulative distributions of heritability explained and total variants, when d_{true} is used (solid) versus d_{meas} (dashed), and accompanying distance-effect distribution (black: d_{true}, brown: d_{meas}) for a 30% error rate in the distance, with network parameters of γ_2=-10 (a) and -15 (b). **(c,d)** Distributions of ϕ for the parameter settings in (a),(b) over 20 simulations, with solid boxes denoting the measured value of ϕ under noise, and dotted boxes showing the true value of ϕ.

4.3b Co-expression explains a significant fraction of genetic effects in neuropsychiatric disease

The intuition behind network genetic architecture is that genes with a similar relationship in the network should give rise to similar kinds of genetic effects. The model in the previous section, however, reduces the full network structure to the path distance alone. This motivates a full random effect model

$$\beta \sim N(0, 2\sigma_g^2 \Delta_p \mathrm{K}(G))$$

Where Δ_p is the diagonal frequency variance matrix $(2p(1\text{-}p))^{\gamma}$, and $\mathrm{K}(G)$ is a *graph kernel* on the network *G*. This model links the network structure to effect covariance, allowing risk-conferring mutations in related genes to co-occur far more frequently than in the previous model.

The more than 3,000,000 protein-altering genic variants (Hoehe2017), and the fact that β is not directly observed, makes the full model intractable in practice. While it may be possible to borrow ideas from Gaussian processes (Titsias2009) to perform approximate inference, a more accessible solution is to use a continuous version of LD score regression (Gazal2017):

$$E\left[z_j^2\right] = N \sum_C \tau_C \ell(j, C) + M$$

$$\ell(j, C) = \sum_{k \in C} a_C(k) \mathrm{cor}(x_j, x_k)^2$$

with a_C the C^{th} eigenvector of K(*G*). Because when K(*G*) is a valid kernel, $\mathrm{K}(G)^{-1}$ is also a valid kernel, then in practice the top 5 and bottom 5 (nonzero) eigenvectors by magnitude are used.

To test whether the co-expression networks capture the covariance of genetic effects, I applied this model to neuropsychiatric disease GWAS, using adult cortex, developing cortex, fetal cortex, and adult whole-blood co-expression networks, taking $\mathrm{K}(G)$ to be the squared

topological overlap. For ASD and SCZ, I find very strong enrichments for network effects within cortical co-expression networks. Table 4.3.1 shows a comparison between the network enrichment statistics for SCZ built from the whole network, and those from expression data and co-expression modules.

LDSR feature	**OR (p-value) Cortex**	**OR (p-value) Blood**
Brain-upregulated genes	1.5 (10^{-8})	N/A
CTX-upregulated genes	1.4 (10^{-4})	N/A
Co-expression modules	1.5 (10^{-5}) [BW-M4]	NS (min p: 0.08)
Module kME	1.4 (10^{-9}) [BW-M4] 1.4 (10^{-8}) [BW-M3]	NS (min p: 0.006)
Network features	5 (10^{-5}) [a_3] 1.5 (10^{-14}) [a_{10}]	NS (min p: 0.04)

Table 1 Network genetic architecture enrichments for SCZ (ClozUK) in adult co-expression networks. Brain-upregulated genes are the 1000 most highly-expressed genes in the GTEx brain samples compared to blood, skin, adipose, liver, and testes. CTX-upregualted genes are the 1000 most highly-expressed genes in the GTEx cortical samples as compared to all other brain regions. Co-expression modules and module kME refer to the the whole-brain co-expression modules described in Chapter 2. The network features are the top (and bottom) principal components of the whole-brain consensus TOM. In cases where multiple modules or features are tested, terms in brackets specify the feature or module that is significant.

Because association studies in ASD have lower sample sizes than SCZ, BP, and MDD, there is far less power to probe genetic architecture, and neither co-expression modules (Gandal2018a, Gandal2018b) nor brain-upregulated genes (Finucane2018) were shown to significantly enrich for excess LD scores. This is the case for network features as well, with two features that show enrichment for SCZ heritability (a_{10}, a_3) showing only nominal significance for ASD heritability ($p < 10^{-1}$). The genes with the highest loadings on these features (top 5%)

enrich for the ontologies receptor signaling, vesicle transport, and pyramidal neuron (a_{10}), as well as synaptic plasticity, synaptic vesicle maturation, and neuron (a_3), consistent with the findings presented in Chapter 3 (**figure 4.4**).

As enrichment ratios for continuous variables are not immediately commensurate with binary variables, I analyzed the 25 quantile bins for both a_{10} and a_3. In each case, the highest quantile bins showed significant enrichments ($p < 10^{-3}$ for bins 23, 24, and 25), and their enrichment values (1.8-2.3) were individually stronger than enrichment values for single modules. Because these features go beyond module co-membership and represent higher-order structures within the gene network, these suggest that restricting analysis to modules alone ignores relationships that are important to disease etiology. As such, methods to incorporate and interrogate networks should provide deeper insights into disease systems biology.

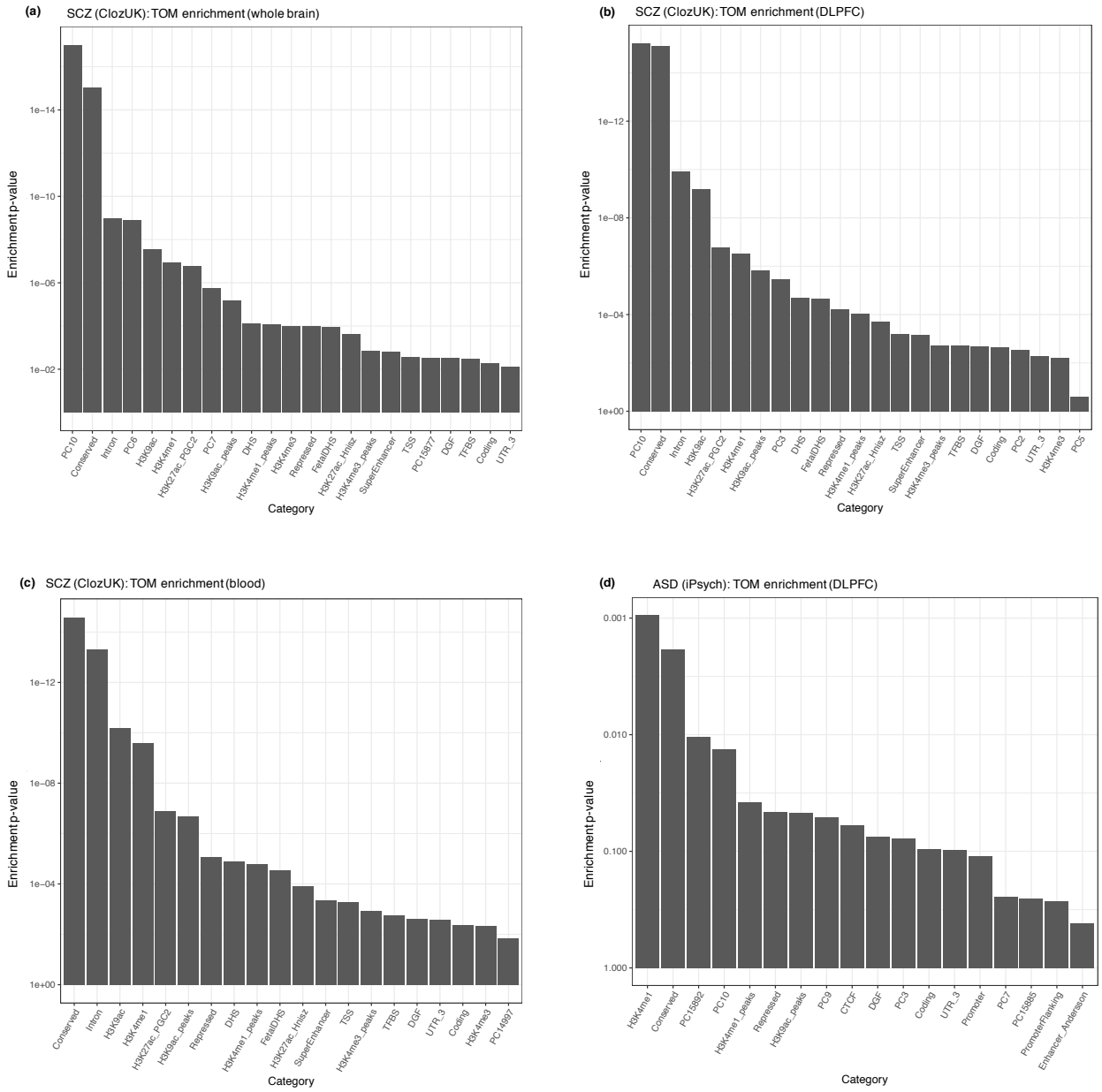

Figure 4.4 Network architecture enrichments for SCZ and ASD in brain and blood

(a) LD-score regression enrichment *p*-values for network features ("PC"s) with standard LD-score regression functional background annotations (all others, see Finucane2015 for details on this background), for the consensus whole-brain TOM built in chapter 2, showing 3 significant network features (PC10, PC6, PC7). **(b)** LDSC *p*-values for network features derived from the DLPFC TOM from chapter 2, showing two significant network features (PC10, PC2). **(c)** LDSC *p*-values for network features derived from a whole-blood TOM; showing no significant network features.

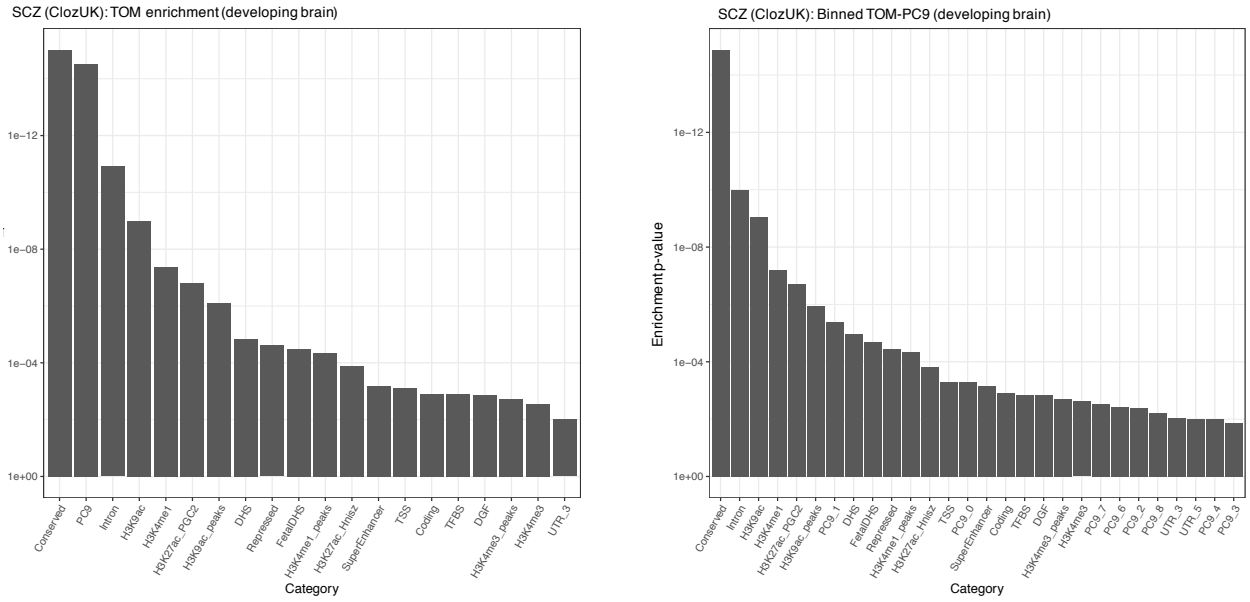

Figure 4.5 Network feature enrichment for SCZ in developing brain

Left: Using the topological overlap from developing brain (Parikshak2016), feature a_9 (PC9) is significantly enriched for SCZ heritability, above most of the background annotations. *Right*: Breaking a_9 into binned components shows that the enrichment comes only from genes with the most negative loadings (as opposed to genes at both extremes): bins 0-9 are all significantly enriched for SCZ heritability, with bins 0,1 significant even after multiple testing correction. Enrichment for all 10 bins is between 1.6 and 2, while enrichment for conserved sites is 1.9, for TSS is 3, and for active chromatin peaks is 6.1.

4.3c Neuropsychiatric peripheral master regulators support a polygenic architecture

In addition to defining a genetic architecture, the omnigenic model makes a distinction between direct and indirect effects. This is a separate mechanistic idea, much conflated with but separate from the genetic architecture aspect of omnigenics. Under the full omnigenic model, not only do large-effect mutations arise almost entirely in core genes, but these mutations are the only mutations to have a direct effect on the trait. Mechanistically, risk variants in peripheral genes act by altering the bioavailability of core genes. However, the majority of GWAS variants do not fall into core genes (Boyle2017), when the trivial expectation is that variants discovered by GWAS should be those with the strongest effects at their frequency, and that therefore they should be more likely to occur in core genes.

The peripheral master regulator (PMR) model is a transcriptional liability model, where risk liability acts through shifts in gene expression levels. It is a structured extension to the TWAS (Gamazon2015) model. TWAS treats liability as a function of local impacts on gene expression:

$$y^{(l)} = \alpha + \delta^T(e - \mu) + \varepsilon$$

$$e = \mu + Bx + \eta$$

Where e is an expression N-vector across all genes, x are an individual's frequency-normalized genotypes, and B (which is N x M) reflects the impact of each variant on each gene, while δ reflects the risk conferred by genetic up-regulation of each gene. ε and η are independent noise terms.

The PMR model changes how the liability and expression are modeled, by replacing e with e_c, the expression levels of core genes. Under this model:

$$y^{(l)} = \alpha + \delta_c{}^T(e_c - \mu) + \varepsilon$$

$$e_c = \mu + \Gamma_c B_c(x_c + \eta_c) + \Gamma_p B_p(x_p + \eta_p)$$

Here, x_c reflect variants local to core genes, and B_c their effects; while x_p and B_p are the variants and corresponding effects local to peripheral genes. Γ_c (N_c x N_c) and Γ_p (N_c x N_p) reflect causal co-expression relationships. Γ_p specifies how perturbations to peripheral genes alter the expression of core genes. A peripheral master regulator is a gene p_i for which $\delta_c{}^T \Gamma_{p_i} B_{p_i} x_{p_i}$ strongly deviates from 0.

Condensing the multiplication shows that the PMR model is a factored local-distal transcriptional liability model:

$$e_c = \mu + \underbrace{\Xi_{loc} x_{loc}}_{\text{TWAS}} + \underbrace{\Xi_{dis} x_{dis}}_{\text{TWAS-trans}} + \eta$$

This relates to network genetic architecture, as the marginal effects of x_{dis} on y will follow

$$\beta_{dis,y} \sim N(0, \Xi_{dis}^T \delta_c \delta_c^T \Xi_{dis})$$

and could be tested using techniques from section 4.2. However, identifying causal co-expression relationships (Γ) is notoriously difficult. By exploiting the fact that TWAS and PMR share the same marginal models for $y|e$, it is possible to test directly individual rows of Ξ_{dis}.

It can be seen that $(\Xi_{dis})_{ij}$ reflects the total change in the expression of gene i with respect to a one unit increase of the normalized dosage of variant j: in other words, $(\Xi_{dis})_{ij}$ is the total trans-QTL effect of variant j on gene i: $\beta_{ij}^{(\text{trQTL})}$. Further, it can be shown **(supplemental methods**) that the expected TWAS Z-score for gene i is equal to $\delta_i(\sqrt{n}\zeta_{LD}^{(i)})$, where $\sqrt{n}$ is the sample size of the GWAS used, and ζ_{LD} is a correction for the local LD structure. Therefore, for core genes c, the statistic

$$\rho_c = \mathrm{cor}_0\left(Z_c^{(\mathrm{TWAS})}, \hat{\beta}_{c,j}^{(\mathrm{trQTL})}\right) = \frac{\langle Z_c^{(\mathrm{TWAS})}, \hat{\beta}_{c,j}^{(\mathrm{trQTL})} \rangle}{\left\|\hat{\beta}_{c,j}^{(\mathrm{trQTL})}\right\| \left\|Z_c^{(\mathrm{TWAS})}\right\|}$$

is an estimator for $\frac{{\delta_c}^T \Gamma_j B_j}{\|\delta_c\| \cdot \|\Gamma_j B_j\|}$. In other words, ρ_c tests whether a distal variant, for a set of hypothesized core genes, up-regulates the risk genes and down-regulates the protective genes.

There are two components of the PMR-based hypothesis for omnigenics: (i) PMRs exist, and (ii) for an omnigenic trait, all the PMRs converge onto the same core set of genes *c*. Because *trans*-QTL hotspots are well known, (i) is likely to be true *a priori* for all sufficiently complex traits. On the other hand, for a polygenic trait (ii) is likely to fail, by virtue of the existence of PMRs for many disjoint sets of genes. With this in mind, I extracted a subset of the data used in chapter 2 of 288 samples from the telencephalon, representing 101 genotyped individuals, and computed ρ_c for 8 publicly-available neuropsychiatric and neurodegenerative TWAS datasets (Mancuso2017) across whole-brain, whole-cortex, and DLPFC co-expression modules.

Because the mechanistic component of the omnigenic hypothesis specifically refers to expression-altering genes as peripheral, I excluded genes that (i) do not generate proteins, (ii) encode known transcription factors, (iii) encode DNA binding proteins, or (iv) encode RNA binding proteins. This left a total of 18,726 potential core genes, of which 15,435 are expressed in the brain. As these genes are also likely candidates for PMRs, I include local variants for these genes in the PMR test. For analysis, I group these variants by categories (i-iv).

I find several modules which appear to be controlled by peripheral master regulators for SCZ, demonstrating that significant PMRs alone implicate many hundreds to thousands of causal protein-coding genes in neuropsychiatric and neurodegenerative disease. Figure 4.5 shows an example of cross-TWAS PMR: the core set of genes is taken to be the top 100 genes from Ruderfer18, while the Z-scores for these genes are taken from the independent PGC GWAS.

Variants near the transcription factor *IRF6*, and variants near the polymerase-associated protein *RPAP1*, show a statistically-significant relationship between *trans*-QTL effect sizes and core gene Z-scores, implicating them as peripheral master regulators. Disruption of *IF6* was recently shown to contribute to neural tube defects (Kousa2019), and *RPAP1* is required to establish and maintain cell identity (Lynch2018). Yet, the PGC data shows significant PMR statistics for other groups of hypothesized core genes, including the module PFC-M1 (*NSUN6, RBM6*), and the neuronal module BW-M4 (*FBXO21, GSX2*). *GSX2* is known to specify neuronal fate (Pei2011) and to play a role in adult neural progenitor activation (Lopez-Juares2013), while *NSUN6* is a member of a family of RNA methyltransferases that play a role in dendritic transcription of mRNA (Majumder2017), and loss of a related family member, *NSUN2,* leads to intellectual disability (Abbasi-Moheb2012). A parsimonious explanation of these data is that several large, disjoint co-expression pathways play a role in schizophrenia etiology and that PMRs play a significant role in shaping their expression levels. This implies that there are many hundreds to thousands of "core" genes, which "in turn may [be] indistinguishable from a model of no core genes" (Wray2018).

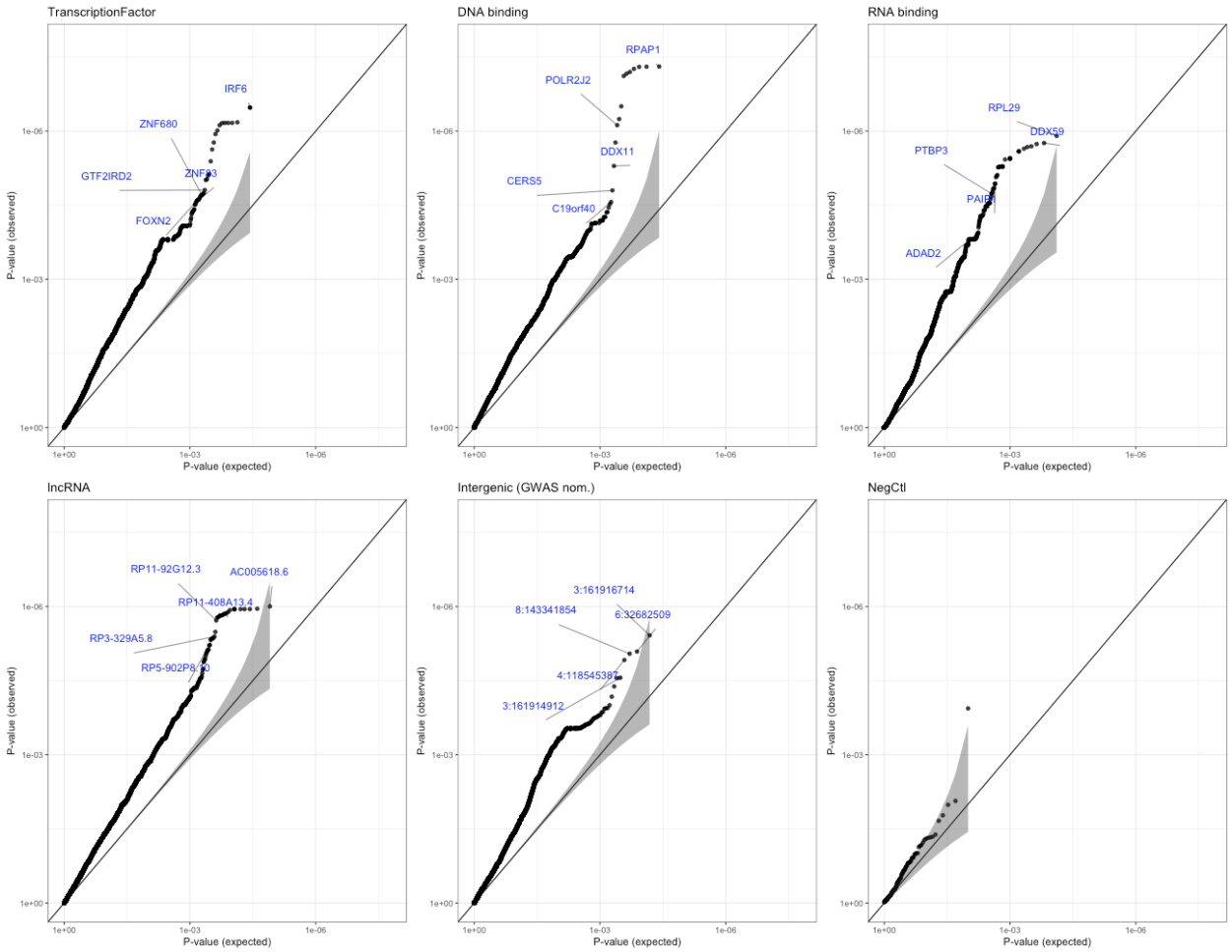

Figure 4.6 PMR significance for empirically-defined core genes in SCZ

Association *p*-values, by regulatory gene type, of the coordinated regulation of a hypothesized core gene set for SCZ. TWAS Z-scores built from PGC GWAS of Schizophrenia, and the hypothesized core set is the top 100 genes from a separate SCZ TWAS built from Ruderfer 2018 ("Ruderfer18$_{100}$"). This test identifies IRF6 and RPAP1 as potential trans-regulators of Ruderfer18$_{100}$. Only known eQTLs are tested, and nominally-significant GWAS intergenic SNPs, so inflation is to be expected. The NegCtl set are a set of 500 randomly-selected SNPs which were permuted within individual to form an empirical null distribution.

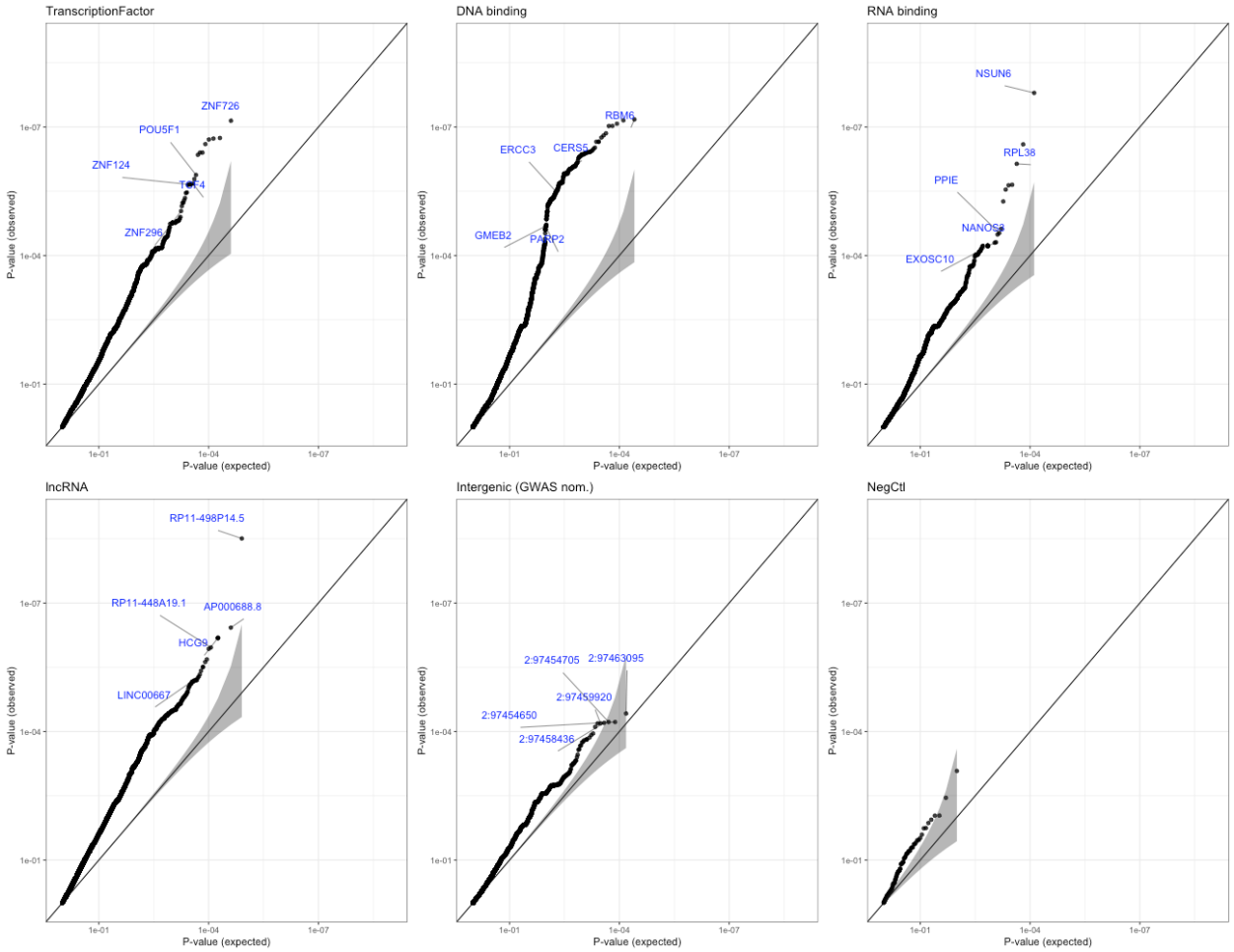

Figure 4.7 PMR significance for PFC-M1 in SCZ

Association *p*-values, by regulatory gene type, of the coordinated regulation of a hypothesized core gene set for SCZ. TWAS Z-scores built from PGC GWAS of Schizophrenia, and the hypothesized core set are those genes present in module PFC-M1. This test identifies NSUN6 and RBM6 as potential trans-regulators of PFC-M1. Only known eQTLs are tested, and nominally-significant GWAS intergenic SNPs, so inflation is to be expected. The NegCtl set are a set of 500 randomly-selected SNPs which were permuted within individual to form an empirical null distribution.

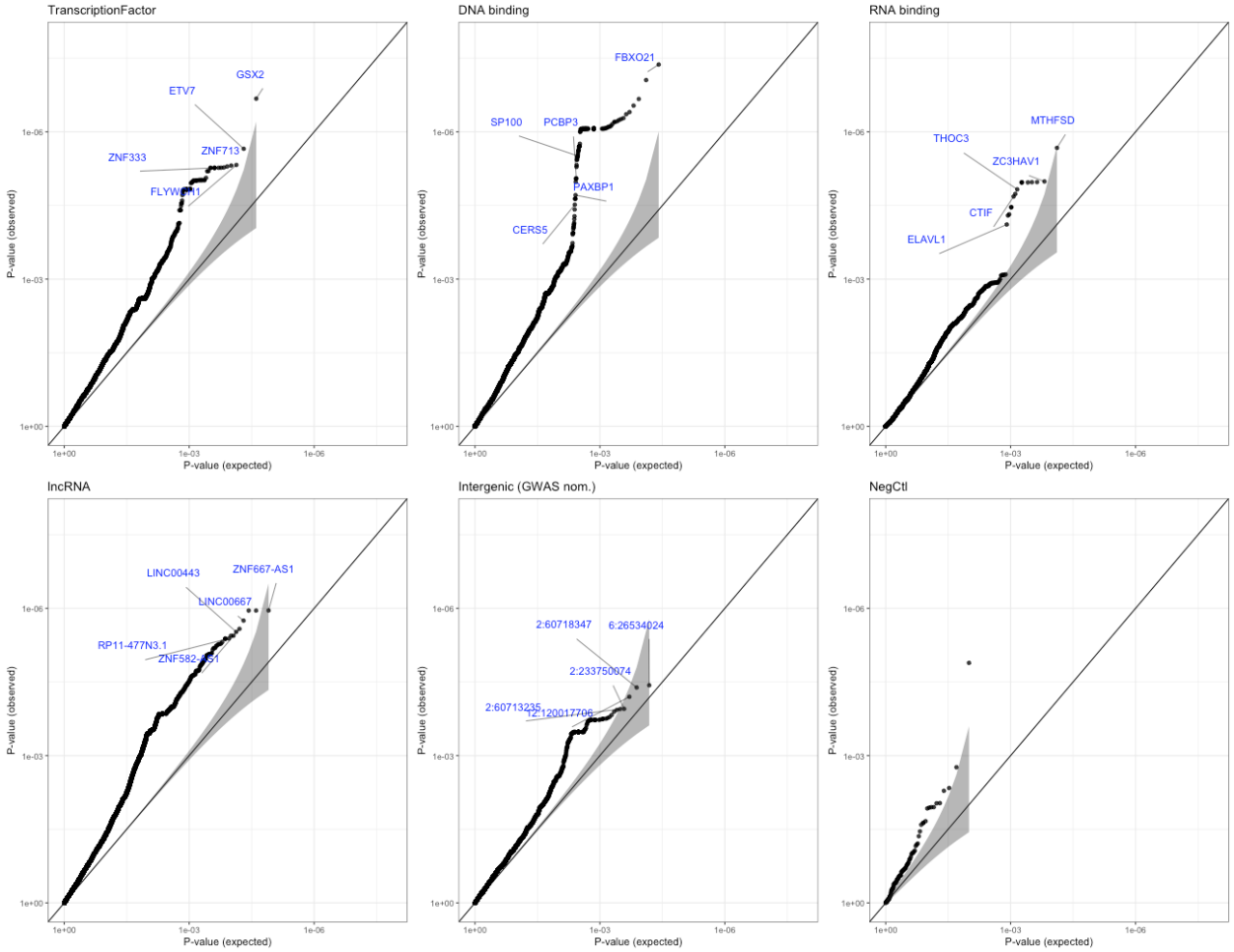

Figure 4.8 PMR significance for BW-M4 in SCZ

Association *p*-values, by regulatory gene type, of the coordinated regulation of a hypothesized core gene set for SCZ. TWAS Z-scores built from PGC GWAS of Schizophrenia, and the hypothesized core set are those genes present in module BW-M4. This test identifies FBXO21 and GSX2 as potential trans-regulators of BW-M4. Only known eQTLs are tested, and nominally-significant GWAS intergenic SNPs, so inflation is to be expected. The NegCtl set are a set of 500 randomly-selected SNPs which were permuted within individual to form an empirical null distribution.

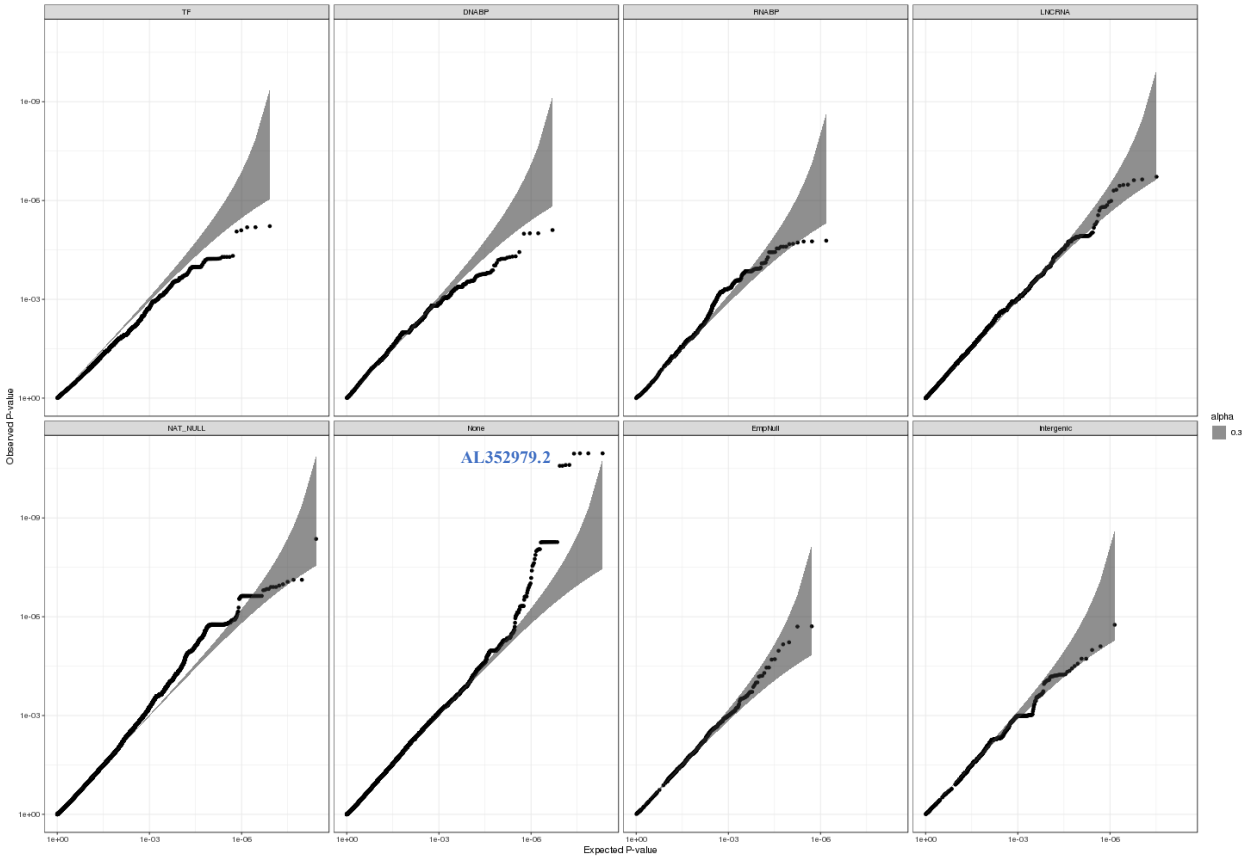

Figure 4.9 PMR significance for BD+SCZ in PFC-M1

Association p-values, by regulatory gene type, of the coordinated regulation of a hypothesized core gene set for SCZ. TWAS Z-scores built from GWAS of Schizophrenia and Bipolar Disorder (Ruderfer2018), and the hypothesized core set are those genes present in module PFC-M1. This test identifies the lncRNA AL352979.2 as a potential regulator of PFC-M1. This lncRNA is nominally significant (p=0.0013) for SCZ in figure 4.6. Only known eQTLs are tested, and nominally-significant GWAS intergenic SNPs, so inflation is to be expected. Here, 5000 negative control SNPs were introduced and used for genomic control (**methods**).

4.4 Discussion

Incorporating gene networks into models of genetic architecture remains a major challenge, with application both to predictive disease models and to translational genetics. This chapter's approach to investigating the omnigenic model comes from a unifying view: that there is a relationship between mutational effect size and genetic network distance – with omnigenic and polygenic architectures representing the strong and weak extremes of that relationship. From this point of view, quantifying a trait's network effect in terms of the decay parameter is of a higher concern than labeling it as strictly omnigenic or polygenic, as there are likely traits at both ends of the spectrum. For instance, secondary phenotypes of Mendelian disorders (such as age-at-onset or disease severity) likely show a strong omnigenic-like relationship within the relevant network. My results show, at least from a co-expression network perspective, that ASD does not have such a strong relationship, as high-impact mutations reside too far from core genes in these networks – and from each other – to represent a strong relationship. These results suggest that, if neuropsychiatric disorders are omnigenic, then gene expression networks, transcription factor binding networks, or PPI networks are not sufficient to explain the relationships that drive the disease state.

Improved models of genetic architecture lead to more precise insights into trait biology. By developing a new model of network genetic architecture – one that is fully compatible with prior approaches – I was able to partition a significant fraction of SCZ and ASD heritability onto gene relationships pertaining to vesicle transport, pyramidal neurons, and receptor signaling, highlighting the genetic correlation between these traits. Interestingly, the estimate of effect size

for a_{10} (which loads strongly onto genes related to synaptic plasticity) is about double in ASD compared to SCZ. Because the ASD transcriptomic signature is correlated to the SCZ transcriptional signature, but shows five-fold larger differences versus controls; and the sibling relative risk for ASD is several times higher than it is for SCZ, an intriguing possibility is that ASD may result from more extreme genetic insults to the same set of underlying pathways. In fact, any two mutually-exclusive traits that show i) high genetic correlation, ii) different heritabilities, and iii) network architecture will be such that the more heritable trait looks like a more genetically extreme version of the less-heritable trait. Seen in this light, the results of (Gandal2018a) are indirect evidence of shared polygenic *network* architecture in neuropsychiatric disease.

There is much research work to be done within the field of network genetic architecture. There are basic questions about which network kernels to choose and the proper way of decomposing them into features for LD-score regression; how to build network genetic architectures from epigenetic (non-genic) data; and how to combine multiple networks into a joint architecture.

Trans-QTLs account for a substantial proportion of gene expression heritability and, like *cis*-QTLs and *splice*-QTLs, will mediate disease liability. Yet identification of *trans*-QTLs, and linking *trans*-acting SNPs to disease, remains largely elusive: GTEx identified only 93 *trans*-QTL genes across 42 tissues. The insights of section 4.3c result in a practical model for simultaneously identifying *trans*-QTLs and associating them with disease. One potential drawback of this model, which is true for all joint analyses of expression (Brynedal2017), is that it implicates no specific downstream gene as causal, but rather an entire set of risk genes. On the other hand, it may be that risk *trans*-QTLs predominantly act on collections of downstream risk

genes, a substantial fraction of which are causal, rather than controlling any one specific gene. The results of section 4.3c identify several examples risk-conferring *trans*-QTLs across SCZ, MDD, AD, and BP, and provide potential roles for specific lncRNA as expression modulators of risk co-expression modules.

As sample sizes for gene expression and gene association studies increase, the analyses pioneered in this section will provide more precise insights into the genetics of neuropsychiatric disease. Higher fidelity estimates of variant effect sizes will enable direct estimates of the network architecture parameter, and provide stronger signals with which to aggregate heritability into genes or network relationships. Larger RNA-seq sample sizes will form a backbone for improved network polygenic models, as well as substantially increase power to detect *trans*-QTLs and PMRs. Together with new data, these tools will enable future scientists to map the genetic complexity of neuropsychiatric disease at a network level.

4.5 Methods

Simulation of network genetic architecture

Simulation: 10,000 causal variants are simulated with frequency parameters estimated from human populations (Ionita-Laza2009), and distances drawn from a binned Beta distribution:

$$p_i \sim \text{Beta}(0.14, 0.7)$$

$$d_i \sim \frac{\lfloor k_d \text{Beta}(a_d, b_d) \rfloor}{k_d}$$

$$\beta_i | d_i, p_i \sim N(0, \sigma_g^2 \left(2p_i(1-p_i)\right)^{\gamma_1} (1+\delta d_i)^{\gamma_2})$$

σ_g^2 is arbitrary and set to 1; k_d is arbitrary so long as it is greater than 10, and is set to k_d; $a_d, b_d, \gamma_1, \gamma_2$, and δ are model parameters. Recent results from the UK Biobank (Schoech2019) suggest that a value of $\gamma_1 = -0.4$ is reasonable for a polygenic trait (height=-0.45, education=-0.32, blood pressure = -0.39) and is fixed to this value. Architectures were simulated on a grid of a_d, b_d=1,1.5,…6; δ=1,1.2,…,2.6; γ_2=-15,-10,-7,-5,-2. Notably for any values of a_d, b_d, δ and γ_2 can be found such that D_1 explains >40% of the heritability. Errors-in-distance: Here the above simulation of distance is replaced by a normal copula (where 20% error corresponds to r=0.8 – this is a purposeful under-estimate, as r^2=0.64 so the latent error is more like 36%):

$$Z \sim N\left(0, \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}\right)$$

$$d_{true} = \frac{\left\lfloor k_d \Phi^{-1}_{\text{Beta}(a_d,b_d)}(\Phi_{N(0,1)}(Z_1)) \right\rfloor}{k_d}$$

$$d_{meas} = \frac{\left\lfloor k_d \Phi^{-1}_{\text{Beta}(a_d,b_d)}(\Phi_{N(0,1)}(Z_2)) \right\rfloor}{k_d}$$

When simulated from a network, first a set of K=1, …, 10 hub genes are simulated with the constraint that no pair can be directly connected by an edge. These form initial communities of size 1. For the remaining 40 core genes, a community is selected at random, a community member is selected at random, and a neighbor is selected at random and added to the community and to the set of core genes. These form the basis of d_{true}, which is taken as the minimal path distance to any core gene. For d_{meas} the communities are distorted by removing M=1,..,10 core genes at random; or by adding K=5,10,…,25 non-core genes at random.

Normalized effect sizes: Identifying the effect size of an empowered 5% frequency GWAS variant happens through three steps: (i) Estimating the liability distribution; (ii) Mapping case/control frequency differences to effect sizes (iii) Estimating power.

(i) Liability Distribution: A 5000x10,000 genotype matrix X is sampled independently, with frequencies given by the previously-simulated vector f, and 5,000 genetic liabilities are generated by $l_g=X\beta$. These liabilities are used to estimate parameters for a T-distribution using 'fitdist' from the R package 'MASS'; the degrees of freedom are reduced by 25% to account partially for rare variants not sampled in this population of 5,000; and these parameters used to generate 400,000 genetic liability scores. These are converted to total liability scores by adding noise $l = l_g + \mathrm{N}(0, \sigma_e)$; with σ_e chosen so that the heritability is 0.85.

(ii) Frequency-ratio-to-effect: The goal is to estimate the ratio p_{aff}/p_{unaff} for a variant with a frequency p_i and effect β_i. The genetic liabilities $l_{new} = l+x\beta_i$ with $x \sim$ binomial(2, p_i) are computed for 400,000 simulated individuals. As 10,000 variants contribute to l, the addition of $x\beta_i$ is assumed to have a minimal effect on heritability. Case/control labels are defined by $l_{new} \geq$ quantile($l_{new}, 0.95$) so that the disease prevalence is 5%, and the empirical frequency

mean(x_{aff})/mean(x_{unaff}) is taken as an estimate of the ratio $p_{aff}/p_{unaff.}$ Fixing p_i=0.05 and varying β_i produces an empirical and invertible map from variant effect to frequency ratio.

(iii) Estimating power: Given an effect size β_i, the case and control frequencies for a $p = 0.05$ variant are obtained from (ii). 5000 case and 5000 control genotypes are sampled according to the corresponding frequencies, and a two-sided T-test performed by 't.test' in R. 1,000 simulations are performed, and the number of times the T-test p-value achieved a Bonferroni-corrected p-value of 0.1/10,000 (the number of causal variants) was tabulated.

Network construction and computation of d(G)

Co-expression network and modules were constructed as in Chapter 2. In addition, an sparse ε=2.5%+1-NN graph is calculated as follows: the cosine distance graph is subset to only the 2.5% smallest edges, and any singleton genes are connected to their closest neighbor. This graph is treated as unweighted, and not necessarily connected. Cross-component distances are treated as 1 + the maximum observed within-component distance. This is referred to as "sparse distance."

Module hub genes are defined as the 2.5% of module genes with largest kWithin values (minimum 5). Distances between a gene and a module is computed as (i) 1 – kME; (ii) mean cosine distance to a module hub; (iii) minimum cosine distance to a module hub; (iv) mean sparse distance to a module hub; (v) minimum sparse distance to a module hub. When using arbitrary gene sets as core genes, (ii)-(iv) are be computed with respect to the gene set in place of module hubs.

Bipartite transcription factor binding graphs were obtained from regulatorycircuits.org, and converted to a similarity network as in Marbach2016. Briefly, the probability weights are taken as edge weights, and the random-walk kernel $K=(I+W)^4$ with W the symmetrically-normalized Laplacian $D^{-1/2}AD^{-1/2}$ of the adjacency matrix; and converted to a dissimilarity via $D_K = 1 - \frac{(K-\min(K))}{(\max(K)-\min(K))}$. A natural set of "core" genes on this network are the most highly-connected genes of K; of which the top 25 are taken. Distances are either the mean or minimum path distance under D_K.

InWeb (Li2016) was used for the protein-protein interaction network. The refined brain-PPI network was obtained from the resource, and a confidence of 0.05 required for an edge to be defined; and the interactions were converted into a binary matrix. Distances were defined as either the minimum or mean path distance in this network.

LD score regression

The LD score regression package was obtained from https://github.com/bulik/ldsc in 5/2019; and run following the best practices for continuous annotations. Network annotations were extracted by extracting the top 5 and bottom 5 (nonzero) principal components of the topological overlap matrix using 'eigsh' in SciPy. Variants were assigned to genes on the basis of being within 150kb of the gene body, forming a (n_{snp} x n_{gene}) binary matrix, onto which the gene loadings are projected. I found that the results were qualitatively unchanged using 50kb and 100kb windows.

Because co-expression analysis routinely identifies modules with low separability (correlated eigengenes), kME values are pruned prior to being used as scores in LD score

regression. The kME-kME correlation r^2 matrix is computed, the module with the highest average r^2 is selected, and all modules with $r^2 > 0.4$ to the selected module are removed. This process is repeated until no modules remain. Thus, for DLPFC, the kME values were selected for BW-M4, PFC-M2, BW-M6, CTX-M4, and PFC-M4.

Peripheral Master Regulator Test

The Statistic for Peripheral Master Regulators (SPMR) test takes in an expression matrix and corresponding genotype dosages, individual IDs (for repeated samples), a TWAS summary statistics .dat file, a gene set (i.e. putative core genes) to test, and any expression covariates. Following best practices for trans-eQTLs, the covariates from chapter 2 were arranged into a (n_{sample} x n_{cov}) and used as prior information to extract 15 HCP factors; which were then used as covariates for SPMR. For efficiency, the expression is pre-corrected for covariates, as opposed to including the covariates as part of a linear model. Expression values and genotype dosages are scaled and centered across samples. To ensure only direct effects are driving the signal, core genes are subset to protein-coding genes that are not known transcription factors, (Lambert2018) and are not present in GO categories 0003677 (DNA binding) or 0003723 (RNA binding). To account for LD, the test gene set is then randomly subset to only genes that do not fall within 100KB of each other. The TWAS Z-scores are subset to the test gene set, and the 2-norm calculated $d_Z = ||_{Ztest}||$.

For each variant, the scaled effect sizes on core gene expression are computed as the correlation between genotype dosage and expression level across samples. This produces a vector of values, θ, one for each core gene. The test statistic, ρ_c, is given by

$$\rho_c = \frac{\theta^T Z}{d_Z}$$

Note that the norm of θ has been dropped from the denominator for efficiency, as the statistic is tested via bootstrap.

Because there are multiple samples per brain in the GTEx telencephalon data, an individual bootstrap is used to approximate the null distribution of ρ_c. For 5,000 replicates, the individual IDs are sampled, with replacement, from the pool of individual IDs. When an individual is included in the bootstrap set, all tissue samples corresponding to that individual are added. For each bootstrapped dosage and expression matrix, ρ_c is calculated, and the resulting distribution is summarized by its mean, standard deviation, skew, excess kurtosis, and quantiles 0.001, 0.1, 0.1, 0.9, 0.99, and 0.999. Approximate *Z*-scores and p-values are given by $(\rho_c^{(obs)} - \hat{\mu}_{\rho_c}^{(boot)})/\hat{\sigma}_{\rho_c}^{(boot)}$. Normality of the bootstrap is assessed by skew and excess kurtosis.

The SPMR has a natural null set of genes: protein-coding genes not likely to have any impact on expression. I identified a set of such genes by filtering all genes by the list of restrictions applied to core genes, plus the restriction that the gene not belong to any GO category containing the terms ‘signaling’, ‘cascade’, ‘channel’, ‘transduction’, ‘transport’, or ‘translation.’

To reduce the multiple testing burden, I test only 727,224 variants, of which 717,572 are significant cis-QTLs in GTEx for a brain-expressed gene, and the remaining are nominally-significant GWAS hits ($p < 10^{-6}$) in either the iPsych ASD GWAS (407 variants), or the PGC SCZ+BIP vs control GWAS (18,816 variants), some of which overlap cis-QTLs. In addition, I treat variants which are cis-QTLs to a gene that could be considered a core candidate (protein coding, does not bind nucleic acids) as background, and do not correct for these tests.

I also apply SPMR more widely than testing TWAS Z-scores. For studies that produce Bayes factors or p-values for each gene, the log-Bayes factor or inverse normal CDF of the p-value is taken in place of the TWAS Z-score. Similarly, module QTLs are identified by taking a +1 score for a gene in the module, and a -1 score for a gene not in the module.

Inflation control

Due to the use of different numbers of samples from the same individual brain, the PMR statistic can show inflation, even using an individual bootstrap, as it is sensitive to expression outliers. By testing *cis*-QTLs instead of all SNPs, it is difficult to distinguish between inflation due to true signal, and inflation due to ancestry differences or assumption violations. To control for inflation, I include a set of 5,000 permuted (within individual) genotypes to assess the genomic inflation for each run of the PMR test. A simple variant of Genomic Control is applied: the slope of the log10-expected and log10-observed p-values from these 5,000 permuted genotypes is obtained by fitting a robust linear model (M-estimation); and (if the slope is > 1) the expected p-values and confidence intervals are adjusted via $p_{adj} = 10^{\beta \log_{10}(p)}$.

Chapter 5 Conclusions and future directions

This thesis presents both an annotated atlas of human brain co-expression networks as well as methodology for incorporating gene networks into models of genetic architecture. In all analyses, the network is means of extracting modules, membership probabilities, and SNP-level features, but is not itself an object of primary study. There are two primary reasons for this strategy. Firstly, the noise inherent in RNA-seq means that any *particular* edge between a pair of genes is highly suspect, while a highly-connected community is far more robust.[4] Secondly, the mathematical descriptors of networks (e.g., chromatic number, conductance, diameter, genus, toughness, thickness) do not readily correspond to biological analogues. Even the very simple network property of vertex connectivity does not capture any disease enrichment signal that can not also be explained by a baseline model of functional annotation such as coding, intronic, promoter, TSS, enhancer, and species conservation (Kim2019) – in contrast with the more general features explored in section 4.3, which are significant over and above the same baseline model. Thus, in all aspects of the work presented, the use of a co-expression network is sufficient but not strictly necessary.

Tensor Expression Analysis, presented in section 2.3c, provides an elegant and powerful alternative to co-expression network analysis within the setting of multiple tissues or multiple cell types. While this approach was evidently first proposed by Hore *et al.* (Hore2016), it has largely been abandoned. Yet its power in identifying global compositional effects that are shared across multiple tissues, regions, or cell types renders it a promising tool. By decomposing an

[4] This is a simple consequence of the fact that, for a multivariate Gaussian $N(0, \Sigma)$, the largest eigenvalues and eigenvectors converge far more rapidly than the rest and are the least impacted by noise – in other words the signal to noise ratio is much better for the top principal components than for any one gene

(individual, tissue, gene) tensor, this approach provides inputs to eQTL screens or heritability partitioning (loadings on individuals), tissue or cell-specific expression (loadings on tissues), and gene set enrichment analysis (loadings on genes). As multi-regional datasets become more common, tensor expression analysis will become more common. However, future work is needed to adapt this method to identify region-specific effects: the objective is to explain *total* variance, so region-specific co-expression patterns can be washed out.

An additional and important extension to this method would be to single-cell sequencing. For example, the single-cell sequencing in Velmeshev *et al.* (Velmeshev2019) can, once cell type clusters have been identified, be collapsed to an (individual, region, cell type, gene) tensor, to which the decomposition approach could be applied. Future work is needed to establish the proper and unbiased method for collapsing from single-cells to cell-type.

Splicing variation across cell types remains largely uncharacterized. The approach taken here relies only on bulk tissue expression, and thus module-specific splicing can be identified without reference to any single-cell expression. The reverse may also be true: It should be possible to identify cell-specific isoforms without reference to co-expression modules. Future work in the area of cell-specific splicing could consider building a "kME-like" statistic on the basis of co-expression between isoform and high-relative-expression genes, enabling the identification of splicing variation within rare cell types that only weakly contribute to bulk expression variance. This same reasoning applies to lncRNA, and was used in Liu *et al.* (Liu2016) to assign functional annotations to the few lncRNA identified in single-cell sequencing, though not to expand the initial set of cell-specific lncRNA to the > 9,000 that were identified in bulk tissue but not in single cells. Thus, future work could attempt to link lncRNA,

isoforms, and even miRNA from bulk tissue to cell types directly, without recourse to network construction or module detection.

The network genetic architecture model introduced in section 4.3b is a starting point for a rich research program. One path of future research is to compare different types of networks (e.g., co-expression, PPI, pathway, etc.) to see which network has a structure that captures most of the heritability, and how much can be explained by combining these networks. Another path of computational statistics research is to provide computational estimators for the variance parameter. Third, the network under consideration need not be restricted to genes alone, but can include functional genetic elements as well, enabling co-regulated epigenetic marks (and thus intergenic variants) to be incorporated into network models of genetic architecture.

Finally, biological validation of network-based genetic findings presents a significant challenge. The observation that adult brain co-expression captures a statistically significant proportion of schizophrenia heritability needs to be translated into what precise biological impact mutational load within the network may have. Because the network structure implicitly aggregates genetic risk across thousands of genes, standard gene-knockout or overexpression approaches are not likely to be fruitful. One potential way forward is to construct a network polygenic risk score and screen populations for extreme individuals in order to extensively phenotype cells derived from those individuals as a natural experiment. Similar approaches have been suggested for the study of complex disease (Hoekstra2017) – in this context, focusing on a network polygenic risk as opposed to total polygenic risk can be thought of (I hypothesize) as examining the phenotypic impact of epistasis specifically.

As a concluding thought: the role of disease association studies has turned an important corner. Identifying candidate disease genes is no longer the sole purpose of such studies – these

data are now used for the secondary purposes of assaying genetic architecture and partitioning heritability among functional annotations. This marks a shift in focus from single-variants and single-genes to systems biology. As demonstrated in this thesis, gene networks (including gene co-expression networks), by grouping functional regions of the genome into coherent downstream units, will help to shape our understanding of human disease.

REFERENCES

Adhya, D., Swarup, V., Nagy, R., Shum, C., Nowosiad, P., Jozwik, K. M., … Baron-Cohen, S. (2018). Atypical neurogenesis and excitatory-inhibitory progenitor generation in induced pluripotent stem cell (iPSC) from autistic individuals. Cold Spring Harbor Laboratory. https://doi.org/10.1101/349415

Akbarian, S., Liu, C., Knowles, J. A., Vaccarino, F. M., Farnham, P. J., Crawford, G. E., … Sestan, N. (2015). The PsychENCODE project. Nature Neuroscience, 18(12), 1707–1712. https://doi.org/10.1038/nn.4156

Allen, J. D., Xie, Y., Chen, M., Girard, L., & Xiao, G. (2012). Comparing Statistical Methods for Constructing Large Scale Gene Networks. PLoS ONE, 7(1), e29348. https://doi.org/10.1371/journal.pone.0029348

Alonso-Gonzalez, A., Rodriguez-Fontenla, C., & Carracedo, A. (2018). De novo Mutations (DNMs) in Autism Spectrum Disorder (ASD): Pathway and Network Analysis. Frontiers in Genetics, 9. https://doi.org/10.3389/fgene.2018.00406

Andlauer, T. F. M., Buck, D., Antony, G., Bayas, A., Bechmann, L., Berthele, A., … Müller-Myhsok, B. (2016). Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation. Science Advances, 2(6), e1501678. https://doi.org/10.1126/sciadv.1501678

Anttila, V., Bulik-Sullivan, B., Finucane, H. K., Walters, R. K., Bras, J., … Neale, B. M. (2018). Analysis of shared heritability in common disorders of the brain. Science, 360(6395), eaap8757. https://doi.org/10.1126/science.aap8757

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics, 25(1), 25–29. https://doi.org/10.1038/75556

Avery, A. R., & Duncan, G. E. (2019). Heritability of Type 2 Diabetes in the Washington State Twin Registry. Twin Research and Human Genetics, 22(2), 95–98. https://doi.org/10.1017/thg.2019.11

Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E. L., Leite, R. E. P., … Herculano-Houzel, S. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. The Journal of Comparative Neurology, 513(5), 532–541. https://doi.org/10.1002/cne.21974

Bakken, T. E., Miller, J. A., Ding, S.-L., Sunkin, S. M., Smith, K. A., Ng, L., … Lein, E. S. (2016). A comprehensive transcriptional map of primate brain development. Nature, 535(7612), 367–375. https://doi.org/10.1038/nature18637

Ballouz, S., Verleyen, W., & Gillis, J. (2015). Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics, 31(13), 2123–2130. https://doi.org/10.1093/bioinformatics/btv118

Ballouz, S., & Gillis, J. (2017). Strength of functional signature correlates with effect size in autism. Genome Medicine, 9(1). https://doi.org/10.1186/s13073-017-0455-8

Barbeira, A. N., Dickinson, S. P., Bonazzola, R., Zheng, J., Wheeler, H. E., … Im, H. K. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nature Communications, 9(1). https://doi.org/10.1038/s41467-018-03621-1

Bardoni, B. (2003). NUFIP1 (nuclear FMRP interacting protein 1) is a nucleocytoplasmic shuttling protein associated with active synaptoneurosomes. Experimental Cell Research, 289(1), 95–107. https://doi.org/10.1016/s0014-4827(03)00222-2

Basu, S. N., Kollu, R., & Banerjee-Basu, S. (2008). AutDB: a gene reference resource for autism research. Nucleic Acids Research, 37(suppl_1), D832–D836. https://doi.org/10.1093/nar/gkn835

Battista, D., Ferrari, C. C., Gage, F. H., & Pitossi, F. J. (2006). Neurogenic niche modulation by activated microglia: transforming growth factor β increases neurogenesis in the adult dentate gyrus. European Journal of Neuroscience, 23(1), 83–93. https://doi.org/10.1111/j.1460-9568.2005.04539.x

Ben-Dor, A., Shamir, R., & Yakhini, Z. (1999). Clustering Gene Expression Patterns. Journal of Computational Biology, 6(3–4), 281–297. https://doi.org/10.1089/106652799318274

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell, 169(7), 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

Bray, N. J., & O'Donovan, M. C. (2018). The genetics of neuropsychiatric disorders. Brain and Neuroscience Advances, 2. https://doi.org/10.1177/2398212818799271

Brynedal, B., Choi, J., Raj, T., Bjornson, R., Stranger, B. E., Neale, B. M., … Cotsapas, C. (2017). Large-Scale trans -eQTLs Affect Hundreds of Transcripts and Mediate Patterns of Transcriptional Co-regulation. The American Journal of Human Genetics, 100(4), 581–591. https://doi.org/10.1016/j.ajhg.2017.02.004

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., … Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics, 47(3), 291–295. https://doi.org/10.1038/ng.3211

Caceres, M., Lachuer, J., Zapala, M. A., Redmond, J. C., Kudo, L., Geschwind, D. H., … Barlow, C. (2003). Elevated gene expression levels distinguish human from non-human primate brains. Proceedings of the National Academy of Sciences, 100(22), 13030–13035. https://doi.org/10.1073/pnas.2135499100

Camp, J. G., Badsha, F., Florio, M., Kanton, S., Gerber, T., Wilsch-Bräuninger, M., … Treutlein, B. (2015). Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. Proceedings of the National Academy of Sciences, 201520760. https://doi.org/10.1073/pnas.1520760112

Carlson, M., Zhang, B., Fang, Z., Mischel, P., Horvath, S., & Nelson, S. (2006). BMC Genomics, 7(1), 40. https://doi.org/10.1186/1471-2164-7-40

Cembrowski, M. S., Bachman, J. L., Wang, L., Sugino, K., Shields, B. C., & Spruston, N. (2016). Spatial Gene-Expression Gradients Underlie Prominent Heterogeneity of CA1 Pyramidal Neurons. Neuron, 89(2), 351–368. https://doi.org/10.1016/j.neuron.2015.12.013

Chen, T., & Guestrin, C. (2016). XGBoost. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. ACM Press. https://doi.org/10.1145/2939672.2939785

Chen, L.-F., Zhou, A. S., & West, A. E. (2017). Transcribing the connectome: roles for transcription factors and chromatin regulators in activity-dependent synapse development. Journal of Neurophysiology, 118(2), 755–770. https://doi.org/10.1152/jn.00067.2017

Chen, R., Xu, X., Huang, L., Zhong, W., & Cui, L. (2019). The Regulatory Role of Long Noncoding RNAs in Different Brain Cell Types Involved in Ischemic Stroke. Frontiers in Molecular Neuroscience, 12. https://doi.org/10.3389/fnmol.2019.00061

Chi, E. C., Allen, G. I., & Baraniuk, R. G. (2016). Convex biclustering. Biometrics, 73(1), 10–19. https://doi.org/10.1111/biom.12540

Clark, B. S., & Blackshaw, S. (2017). Understanding the Role of lncRNAs in Nervous System Development. In Advances in Experimental Medicine and Biology (pp. 253–282). Springer Singapore. https://doi.org/10.1007/978-981-10-5203-3_9

Cogill, S. B., Srivastava, A. K., Yang, M. Q., & Wang, L. (2018). Co-expression of long non-coding RNAs and autism risk genes in the developing human brain. BMC Systems Biology, 12(S7). https://doi.org/10.1186/s12918-018-0639-x

Collado-Torres, L., Burke, E. E., Peterson, A., Shin, J., Straub, R. E., Rajpurohit, A., … Jaffe, A. E. (2019). Regional Heterogeneity in Gene Expression, Regulation, and Coherence in the Frontal Cortex and Hippocampus across Development and Schizophrenia. Neuron, 103(2), 203–216.e8. https://doi.org/10.1016/j.neuron.2019.05.013

Cope, L., Naiman, D. Q., & Parmigiani, G. (2014). Integrative correlation: Properties and relation to canonical correlations. Journal of Multivariate Analysis, 123, 270–280. https://doi.org/10.1016/j.jmva.2013.09.011

Crow, M., Paul, A., Ballouz, S., Huang, Z. J., & Gillis, J. (2018). Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. Nature Communications, 9(1). https://doi.org/10.1038/s41467-018-03282-0

Csardi G., Tamas N. (2006). The igraph software package for complex network research. InterJournal, Complex Systems, 1695.

Davies, M. N., Volta, M., Pidsley, R., Lunnon, K., Dixit, A., Lovestone, S., … Mill, J. (2012). Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. Genome Biology, 13(6), R43. https://doi.org/10.1186/gb-2012-13-6-r43

de la Torre-Ubieta, L., Won, H., Stein, J. L., & Geschwind, D. H. (2016). Advancing the understanding of autism disease mechanisms through genetics. Nature Medicine, 22(4), 345–361. https://doi.org/10.1038/nm.4071

de la Torre-Ubieta, L., Stein, J. L., Won, H., Opland, C. K., Liang, D., Lu, D., & Geschwind, D. H. (2018). The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. Cell, 172(1–2), 289–304.e18. https://doi.org/10.1016/j.cell.2017.12.014

de Leeuw, C. A., Mooij, J. M., Heskes, T., & Posthuma, D. (2015). MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLOS Computational Biology, 11(4), e1004219. https://doi.org/10.1371/journal.pcbi.1004219

DeRosa, B. A., El Hokayem, J., Artimovich, E., Garcia-Serje, C., Phillips, A. W., Van Booven, D., … Dykxhoorn, D. M. (2018). Convergent Pathways in Idiopathic Autism Revealed by Time Course Transcriptomic Analysis of Patient-Derived Neurons. Scientific Reports, 8(1). https://doi.org/10.1038/s41598-018-26495-1

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., … Gingeras, T. R. (2012). Landscape of transcription in human cells. Nature, 489(7414), 101–108. https://doi.org/10.1038/nature11233
Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2012). STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dobrin, R., Zhu, J., Molony, C., Argman, C., Parrish, M. L., Carlson, S., … Schadt, E. E. (2009). Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. Genome Biology, 10(5), R55. https://doi.org/10.1186/gb-2009-10-5-r55

Doll, C. A., & Broadie, K. (2014). Impaired activity-dependent neural circuit assembly and refinement in autism spectrum disorder genetic models. Frontiers in Cellular Neuroscience, 8. https://doi.org/10.3389/fncel.2014.00030

Dörrbaum, A. R., Kochen, L., Langer, J. D., & Schuman, E. M. (2018). Local and global influences on protein turnover in neurons and glia. eLife, 7. https://doi.org/10.7554/elife.34202

Félix, M.-A., & Barkoulas, M. (2015). Pervasive robustness in biological systems. Nature Reviews Genetics, 16(8), 483–496. https://doi.org/10.1038/nrg3949

Finucane, H. K., Reshef, Y. A., Anttila, V., Slowikowski, K., Gusev, A., … Price, A. L. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nature Genetics, 50(4), 621–629. https://doi.org/10.1038/s41588-018-0081-4

Freytag, S., Gagnon-Bartsch, J., Speed, T. P., & Bahlo, M. (2015). Systematic noise degrades gene co-expression signals but can be corrected. BMC Bioinformatics, 16(1). https://doi.org/10.1186/s12859-015-0745-3

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. The Annals of Statistics, 19(1), 1–67. https://doi.org/10.1214/aos/1176347963

Fromer, M., Roussos, P., Sieberts, S. K., Johnson, J. S., Kavanagh, D. H., Perumal, T. M., … Sklar, P. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nature Neuroscience, 19(11), 1442–1453. https://doi.org/10.1038/nn.4399

Fukuda, T., Itoh, M., Ichikawa, T., Washiyama, K., & Goto, Y. (2005). Delayed Maturation of Neuronal Architecture and Synaptogenesis in Cerebral Cortex ofMecp2-Deficient Mice. Journal of Neuropathology & Experimental Neurology, 64(6), 537–544. https://doi.org/10.1093/jnen/64.6.537

Fullard, J. F., Hauberg, M. E., Bendl, J., Egervari, G., Cirnaru, M.-D., Reach, S. M., … Roussos, P. (2018). An atlas of chromatin accessibility in the adult human brain. Genome Research, 28(8), 1243–1252. https://doi.org/10.1101/gr.232488.117

Galatro, T. F., Holtman, I. R., Lerario, A. M., Vainchtein, I. D., Brouwer, N., Sola, P. R., … Eggen, B. J. L. (2017). Transcriptomic analysis of purified human cortical microglia reveals age-associated changes. Nature Neuroscience, 20(8), 1162–1171. https://doi.org/10.1038/nn.4597

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., … Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. Nature Genetics, 47(9), 1091–1098. https://doi.org/10.1038/ng.3367

Gamazon, E. R., Zwinderman, A. H., Cox, N. J., Denys, D., & Derks, E. M. (2019). Multi-tissue transcriptome analyses identify genetic mechanisms underlying neuropsychiatric traits. Nature Genetics, 51(6), 933–940. https://doi.org/10.1038/s41588-019-0409-8

Gandal, M. J., Leppa, V., Won, H., Parikshak, N. N., & Geschwind, D. H. (2016). The road to precision psychiatry: translating genetics into disease mechanisms. Nature Neuroscience, 19(11), 1397–1407. https://doi.org/10.1038/nn.4409

Gandal, M. J., Haney, J. R., Parikshak, N. N., Leppa, V., Ramaswami, G., … Hartl, C. (2018). Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. Science, 359(6376), 693–697. https://doi.org/10.1126/science.aad6469

Gandal, M. J., Zhang, P., Hadjimichael, E., Walker, R. L., Chen, C., … Liu, S. (2018). Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science, 362(6420), eaat8127. https://doi.org/10.1126/science.aat8127

Garbett, K., Ebert, P. J., Mitchell, A., Lintas, C., Manzi, B., Mirnics, K., & Persico, A. M. (2008). Immune transcriptome alterations in the temporal cortex of subjects with autism. Neurobiology of Disease, 30(3), 303–311. https://doi.org/10.1016/j.nbd.2008.01.012

Gaugler, T., Klei, L., Sanders, S. J., Bodea, C. A., Goldberg, A. P., Lee, A. B., … Buxbaum, J. D. (2014). Most genetic risk for autism resides with common variation. Nature Genetics, 46(8), 881–885. https://doi.org/10.1038/ng.3039

Gazal, S., Finucane, H. K., Furlotte, N. A., Loh, P.-R., Palamara, P. F., Liu, X., … Price, A. L. (2017). Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nature Genetics, 49(10), 1421–1427. https://doi.org/10.1038/ng.3954

Gerstner, J. R., Koberstein, J. N., Watson, A. J., Zapero, N., Risso, D., Speed, T. P., … Peixoto, L. (2016). Removal of unwanted variation reveals novel patterns of gene expression linked to sleep homeostasis in murine cortex. BMC Genomics, 17(S8). https://doi.org/10.1186/s12864-016-3065-8

Geschwind, D. H. (2000). Mice, microarrays, and the genetic diversity of the brain. Proceedings of the National Academy of Sciences, 97(20), 10676–10678. https://doi.org/10.1073/pnas.97.20.10676

Geschwind, D. H., & Flint, J. (2015). Genetics and genomics of psychiatric disease. Science, 349(6255), 1489–1494. https://doi.org/10.1126/science.aaa8954

Gong, S., Zheng, C., Doughty, M. L., Losos, K., Didkovsky, N., Schambra, U. B., … Heintz, N. (2003). A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. Nature, 425(6961), 917–925. https://doi.org/10.1038/nature02033

Gosselin, D., Skola, D., Coufal, N. G., Holtman, I. R., Schlachetzki, J. C. M., Sajti, E., … Glass, C. K. (2017). An environment-dependent transcriptional network specifies human microglia identity. Science, 356(6344), eaal3222. https://doi.org/10.1126/science.aal3222

Goyal, M. S., Hawrylycz, M., Miller, J. A., Snyder, A. Z., & Raichle, M. E. (2014). Aerobic Glycolysis in the Human Brain Is Associated with Development and Neotenous Gene Expression. Cell Metabolism, 19(1), 49–57. https://doi.org/10.1016/j.cmet.2013.11.020

Graham, T. G. W., Carney, S. M., Walter, J. C., & Loparo, J. J. (2018). A single XLF dimer bridges DNA ends during nonhomologous end joining. Nature Structural & Molecular Biology, 25(9), 877–884. https://doi.org/10.1038/s41594-018-0120-y

Gratten, J., Wray, N. R., Keller, M. C., & Visscher, P. M. (2014). Large-scale genomics unveils the genetic architecture of psychiatric disorders. Nature Neuroscience, 17(6), 782–790. https://doi.org/10.1038/nn.3708

Grove, J., Ripke, S., Als, T. D., Mattheisen, M., … Walters, R. K. (2019). Identification of common genetic risk variants for autism spectrum disorder. Nature Genetics, 51(3), 431–444. https://doi.org/10.1038/s41588-019-0344-8

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. Nature, 550(7675), 204–213. https://doi.org/10.1038/nature24277

Guan, J., Cai, J. J., Ji, G., & Sham, P. C. (2019). Commonality in dysregulated expression of gene sets in cortical brains of individuals with autism, schizophrenia, and bipolar disorder. Translational Psychiatry, 9(1). https://doi.org/10.1038/s41398-019-0488-4

Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., … Regev, A. (2017). Massively parallel single-nucleus RNA-seq with DroNc-seq. Nature Methods, 14(10), 955–958. https://doi.org/10.1038/nmeth.4407

Hawrylycz, M. J., Lein, E. S., Guillozet-Bongaarts, A. L., Shen, E. H., Ng, L., Miller, J. A., … Jones, A. R. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. Nature, 489(7416), 391–399. https://doi.org/10.1038/nature11405

Hawrylycz, M., Miller, J. A., Menon, V., Feng, D., Dolbeare, T., Guillozet-Bongaarts, A. L., … Lein, E. (2015). Canonical genetic signatures of the adult human brain. Nature Neuroscience, 18(12), 1832–1844. https://doi.org/10.1038/nn.4171

Heintz, N. (2004). Gene Expression Nervous System Atlas (GENSAT). Nature Neuroscience, 7(5), 483–483. https://doi.org/10.1038/nn0504-483

Hernandez, D. G., Nalls, M. A., Moore, M., Chong, S., Dillman, A., Trabzuni, D., … Cookson, M. R. (2012). Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. Neurobiology of Disease, 47(1), 20–28. https://doi.org/10.1016/j.nbd.2012.03.020

Hoehe, M. R., Herwig, R., Mao, Q., Peters, B. A., Drmanac, R., Church, G. M., & Huebsch, T. (2017). Significant abundance of cis configurations of mutations in diploid human genomes. Cold Spring Harbor Laboratory. https://doi.org/10.1101/221085

Hoekstra, S. D., Stringer, S., Heine, V. M., & Posthuma, D. (2017). Genetically-Informed Patient Selection for iPSC Studies of Complex Diseases May Aid in Reducing Cellular Heterogeneity. Frontiers in Cellular Neuroscience, 11. https://doi.org/10.3389/fncel.2017.00164

Hoffman, G. E., & Schadt, E. E. (2016). variancePartition: interpreting drivers of variation in complex gene expression studies. BMC Bioinformatics, 17(1). https://doi.org/10.1186/s12859-016-1323-z

Hormozdiari, F., Penn, O., Borenstein, E., & Eichler, E. E. (2014). The discovery of integrated gene networks for autism and related disorders. Genome Research, 25(1), 142–154. https://doi.org/10.1101/gr.178855.114

Hormozdiari, F., Gazal, S., van de Geijn, B., Finucane, H. K., Ju, C. J.-T., Loh, P.-R., … Price, A. L. (2018). Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. Nature Genetics, 50(7), 1041–1047. https://doi.org/10.1038/s41588-018-0148-2

Hornik, K., & Grün, B. (2014). movMF: AnRPackage for Fitting Mixtures of von Mises-Fisher Distributions. Journal of Statistical Software, 58(10). https://doi.org/10.18637/jss.v058.i10

Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K., & Marchini, J. (2016). Tensor decomposition for multiple-tissue gene expression experiments. Nature Genetics, 48(9), 1094–1100. https://doi.org/10.1038/ng.3624

Hu, X. -d., Huang, Q., Yang, X., & Xia, H. (2007). Differential Regulation of AMPA Receptor Trafficking by Neurabin-Targeted Synaptic Protein Phosphatase-1 in Synaptic Transmission and Long-Term Depression in Hippocampus. Journal of Neuroscience, 27(17), 4674–4686. https://doi.org/10.1523/jneurosci.5365-06.2007

Huguet, G., Ey, E., & Bourgeron, T. (2013). The Genetic Landscapes of Autism Spectrum Disorders. Annual Review of Genomics and Human Genetics, 14(1), 191–213. https://doi.org/10.1146/annurev-genom-091212-153431

Ionita-Laza, I., Lange, C., & M. Laird, N. (2009). Estimating the number of unseen variants in the human genome. Proceedings of the National Academy of Sciences, 106(13), 5008–5013. https://doi.org/10.1073/pnas.0807815106

Ip, J. P. K., Nagakura, I., Petravicz, J., Li, K., Wiemer, E. A. C., & Sur, M. (2018). Major Vault Protein, a Candidate Gene in 16p11.2 Microdeletion Syndrome, Is Required for the Homeostatic Regulation of Visual Cortical Plasticity. The Journal of Neuroscience, 38(16), 3890–3900. https://doi.org/10.1523/jneurosci.2034-17.2018

Jay, J. J., Eblen, J. D., Zhang, Y., Benson, M., Perkins, A. D., Saxton, A. M., … Langston, M. A. (2012). A systematic comparison of genome-scale clustering algorithms. BMC Bioinformatics, 13(Suppl 10), S7. https://doi.org/10.1186/1471-2105-13-s10-s7

Jha, M., Guzzi, P. H., & Roy, S. (2019). Qualitative assessment of functional module detectors on microarray and RNASeq data. Network Modeling Analysis in Health Informatics and Bioinformatics, 8(1). https://doi.org/10.1007/s13721-018-0180-2

Johnson, M. B., Kawasawa, Y. I., Mason, C. E., Krsnik, Ž., Coppola, G., Bogdanović, D., … Šestan, N. (2009). Functional and Evolutionary Insights into Human Brain Development through Global Transcriptome Analysis. Neuron, 62(4), 494–509. https://doi.org/10.1016/j.neuron.2009.03.027

Johnson, R. C., Nelson, G. W., Troyer, J. L., Lautenberger, J. A., Kessing, B. D., Winkler, C. A., & O'Brien, S. J. (2010). Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics, 11(1), 724. https://doi.org/10.1186/1471-2164-11-724

Johnson, M. R., Shkura, K., Langley, S. R., Delahaye-Duriez, A., Srivastava, P., Hill, W. D., … Petretto, E. (2015). Systems genetics identifies a convergent gene network for cognition and neurodevelopmental disease. Nature Neuroscience, 19(2), 223–232. https://doi.org/10.1038/nn.4205

Jovanovic, V. M., Salti, A., Tilleman, H., Zega, K., Jukic, M. M., Zou, H., … Brodski, C. (2018). BMP/SMAD Pathway Promotes Neurogenesis of Midbrain Dopaminergic Neurons In Vivo and in Human Induced Pluripotent and Neural Stem Cells. The Journal of Neuroscience, 38(7), 1662–1676. https://doi.org/10.1523/jneurosci.1540-17.2018

Kadarmideen, H. N., & Watson-haigh, N. S. (2012). Building gene co-expression networks using transcriptomics data for systems biology investigations: Comparison of methods using microarray data. Bioinformation, 8(18), 855–861. https://doi.org/10.6026/97320630008855

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., … Šestan, N. (2011). Spatio-temporal transcriptome of the human brain. Nature, 478(7370), 483–489. https://doi.org/10.1038/nature10523

Kelley, K. W., Nakao-Inoue, H., Molofsky, A. V., & Oldham, M. C. (2018). Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes. Nature Neuroscience, 21(9), 1171–1184. https://doi.org/10.1038/s41593-018-0216-z

Kim, S. S., Dai, C., Hormozdiari, F., van de Geijn, B., Gazal, S., Park, Y., … Price, A. L. (2019). Genes with High Network Connectivity Are Enriched for Disease Heritability. The American Journal of Human Genetics, 104(5), 896–913. https://doi.org/10.1016/j.ajhg.2019.03.020

Kokotos, A. C., Peltier, J., Davenport, E. C., Trost, M., & Cousin, M. A. (2018). Activity-dependent bulk endocytosis proteome reveals a key presynaptic role for the monomeric GTPase Rab11. Proceedings of the National Academy of Sciences, 115(43), E10177–E10186. https://doi.org/10.1073/pnas.1809189115

Konopka, G., Friedrich, T., Davis-Turak, J., Winden, K., Oldham, M. C., Gao, F., … Geschwind, D. H. (2012). Human-Specific Transcriptional Networks in the Brain. Neuron, 75(4), 601–617. https://doi.org/10.1016/j.neuron.2012.05.034

Koshibu, K., Graff, J., Beullens, M., Heitz, F. D., Berchtold, D., Russig, H., … Mansuy, I. M. (2009). Protein Phosphatase 1 Regulates the Histone Code for Long-Term Memory. Journal of Neuroscience, 29(41), 13079–13089. https://doi.org/10.1523/jneurosci.3610-09.2009

Kousa, Y. A., Zhu, H., Fakhouri, W. D., Lei, Y., Kinoshita, A., Roushangar, R. R., … Schutte, B. C. (2019). The TFAP2A–IRF6–GRHL3 genetic pathway is conserved in neurulation. Human Molecular Genetics, 28(10), 1726–1737. https://doi.org/10.1093/hmg/ddz010

Lachmann, A., Giorgi, F. M., Lopez, G., & Califano, A. (2016). ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. Bioinformatics, 32(14), 2233–2235. https://doi.org/10.1093/bioinformatics/btw216

Lake, B. B., Ai, R., Kaeser, G. E., Salathia, N. S., Yung, Y. C., Liu, R., … Zhang, K. (2016). Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science, 352(6293), 1586–1590. https://doi.org/10.1126/science.aaf1204

Lake, B. B., Chen, S., Sos, B. C., Fan, J., Kaeser, G. E., Yung, Y. C., … Zhang, K. (2017). Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nature Biotechnology, 36(1), 70–80. https://doi.org/10.1038/nbt.4038

Lambert, J.-C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., … Sims, R. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nature Genetics, 45(12), 1452–1458. https://doi.org/10.1038/ng.2802

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., … Weirauch, M. T. (2018). The Human Transcription Factors. Cell, 172(4), 650–665. https://doi.org/10.1016/j.cell.2018.01.029

Langfelder, P., Luo, R., Oldham, M. C., & Horvath, S. (2011). Is My Network Module Preserved and Reproducible? PLoS Computational Biology, 7(1), e1001057. https://doi.org/10.1371/journal.pcbi.1001057

Lathia, J. D., Okun, E., Tang, S.-C., Griffioen, K., Cheng, A., Mughal, M. R., … Mattson, M. P. (2008). Toll-Like Receptor 3 Is a Negative Regulator of Embryonic Neural Progenitor Cell Proliferation. Journal of Neuroscience, 28(51), 13978–13984. https://doi.org/10.1523/jneurosci.2140-08.2008

Leduc, M. S., Blair, R. H., Verdugo, R. A., Tsaih, S.-W., Walsh, K., Churchill, G. A., & Paigen, B. (2012). Using bioinformatics and systems genetics to dissect HDL-cholesterol genetics in an MRL/MpJ × SM/J intercross. Journal of Lipid Research, 53(6), 1163–1175. https://doi.org/10.1194/jlr.m025833

Lee, H. K. (2004). Coexpression Analysis of Human Genes Across Many Microarray Data Sets. Genome Research, 14(6), 1085–1094. https://doi.org/10.1101/gr.1910904

Lee, K. E., Seo, J., Shin, J., Ji, E. H., Roh, J., Kim, J. Y., … Kim, J. (2014). Positive feedback loop between Sox2 and Sox6 inhibits neuronal differentiation in the developing central nervous system. Proceedings of the National Academy of Sciences, 111(7), 2794–2799. https://doi.org/10.1073/pnas.1308758111

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics, 28(6), 882–883. https://doi.org/10.1093/bioinformatics/bts034

Lein, E. S., Hawrylycz, M. J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., … Jones, A. R. (2006). Genome-wide atlas of gene expression in the adult mouse brain. Nature, 445(7124), 168–176. https://doi.org/10.1038/nature05453

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., … MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature, 536(7616), 285–291. https://doi.org/10.1038/nature19057

Lewis, E. M. A., Meganathan, K., Baldridge, D., Gontarz, P., Zhang, B., Bonni, A., … Kroll, K. L. (2019). Cellular and molecular characterization of multiplex autism in human induced pluripotent stem cell-derived neurons. Cold Spring Harbor Laboratory. https://doi.org/10.1101/620807

Li, G., Ma, Q., Tang, H., Paterson, A. H., & Xu, Y. (2009). QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. Nucleic Acids Research, 37(15), e101–e101. https://doi.org/10.1093/nar/gkp491

Li, T., Wernersson, R., Hansen, R. B., Horn, H., Mercer, J., Slodkowicz, G., … Lage, K. (2016). A scored human protein–protein interaction network to catalyze genomic interpretation. Nature Methods, 14(1), 61–64. https://doi.org/10.1038/nmeth.4083

Li, M., Santpere, G., Imamura Kawasawa, Y., Evgrafov, O. V., Gulden, F. O., … Pochareddy, S. (2018). Integrative functional genomic analysis of human brain development and neuropsychiatric risks. Science, 362(6420), eaat7615. https://doi.org/10.1126/science.aat7615

Liddelow, S. A., Guttenplan, K. A., Clarke, L. E., Bennett, F. C., Bohlen, C. J., Schirmer, L., … Barres, B. A. (2017). Neurotoxic reactive astrocytes are induced by activated microglia. Nature, 541(7638), 481–487. https://doi.org/10.1038/nature21029

Liu H., Roeder, K. & Wasserman, L. (2010). Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. Proceedings of the 23rd International Conference on Neural Information Processing Systems, 2, 1432-1440

Liu, S. J., Nowakowski, T. J., Pollen, A. A., Lui, J. H., Horlbeck, M. A., Attenello, F. J., … Lim, D. A. (2016). Single-cell analysis of long non-coding RNAs in the developing human neocortex. Genome Biology, 17(1). https://doi.org/10.1186/s13059-016-0932-1

Long, Q., Argmann, C., Houten, S. M., Huang, T., Peng, S., … Zhu, J. (2016). Inter-tissue coexpression network analysis reveals DPP4 as an important gene in heart to blood communication. Genome Medicine, 8(1). https://doi.org/10.1186/s13073-016-0268-1

Luo, F., Yang, Y., Zhong, J., Gao, H., Khan, L., Thompson, D. K., & Zhou, J. (2007). Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. BMC Bioinformatics, 8(1), 299. https://doi.org/10.1186/1471-2105-8-299

Luo, C., Keown, C. L., Kurihara, L., Zhou, J., He, Y., Li, J., … Ecker, J. R. (2017). Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science, 357(6351), 600–604. https://doi.org/10.1126/science.aan3351

Madabhushi, R., Gao, F., Pfenning, A. R., Pan, L., Yamakawa, S., Seo, J., … Tsai, L.-H. (2015). Activity-Induced DNA Breaks Govern the Expression of Neuronal Early-Response Genes. Cell, 161(7), 1592–1605. https://doi.org/10.1016/j.cell.2015.05.032

Mahfouz, A., Ziats, M. N., Rennert, O. M., Lelieveldt, B. P. F., & Reinders, M. J. T. (2015). Shared Pathways Among Autism Candidate Genes Determined by Co-expression Network Analysis of the Developing Human Brain Transcriptome. Journal of Molecular Neuroscience, 57(4), 580–594. https://doi.org/10.1007/s12031-015-0641-3

Mahfouz, A., Huisman, S. M. H., Lelieveldt, B. P. F., & Reinders, M. J. T. (2016). Brain transcriptome atlases: a computational perspective. Brain Structure and Function, 222(4), 1557–1580. https://doi.org/10.1007/s00429-016-1338-2

Malhotra, D., & Sebat, J. (2012). CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics. Cell, 148(6), 1223–1241. https://doi.org/10.1016/j.cell.2012.02.039

Mancarci, B. O., Toker, L., Tripathy, S. J., Li, B., Rocco, B., Sibille, E., & Pavlidis, P. (2017). Cross-Laboratory Analysis of Brain Cell Type Transcriptomes with Applications to Interpretation of Bulk Tissue Data. Eneuro, 4(6), ENEURO.0212-17.2017. https://doi.org/10.1523/eneuro.0212-17.2017

Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., & Pasaniuc, B. (2017). Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. The American Journal of Human Genetics, 100(3), 473–487. https://doi.org/10.1016/j.ajhg.2017.01.031

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., & Califano, A. (2006). ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics, 7(S1). https://doi.org/10.1186/1471-2105-7-s1-s7

Martinez-Morales, P. L., Quiroga, A. C., Barbas, J. A., & Morales, A. V. (2010). SOX5 controls cell cycle progression in neural progenitors by interfering with the WNT–β-catenin pathway. EMBO Reports, 11(6), 466–472. https://doi.org/10.1038/embor.2010.61

Maycox, P. R., Kelly, F., Taylor, A., Bates, S., Reid, J., Logendra, R., … de Belleroche, J. (2009). Analysis of gene expression in two large schizophrenia cohorts identifies multiple changes associated with nerve terminal function. Molecular Psychiatry, 14(12), 1083–1094. https://doi.org/10.1038/mp.2009.18

Maze, I., Wenderski, W., Noh, K.-M., Bagot, R. C., Tzavaras, N., Purushothaman, I., … Allis, C. D. (2015). Critical Role of Histone Turnover in Neuronal Transcription and Plasticity. Neuron, 87(1), 77–94. https://doi.org/10.1016/j.neuron.2015.06.014

McKenzie, A. T., Wang, M., Hauberg, M. E., Fullard, J. F., Kozlenkov, A., Keenan, A., … Zhang, B. (2018). Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. Scientific Reports, 8(1). https://doi.org/10.1038/s41598-018-27293-5

McNicholas, P. D., & Murphy, T. B. (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models. Bioinformatics, 26(21), 2705–2712. https://doi.org/10.1093/bioinformatics/btq498

Milborrow, S., Derived from mda:mars by T. Hastie and R. Tibshirani. earth: Multivariate Adaptive Regression Splines (2011). R package.

Miller, J. A., Oldham, M. C., & Geschwind, D. H. (2008). A Systems Level Analysis of Transcriptional Changes in Alzheimer's Disease and Normal Aging. Journal of Neuroscience, 28(6), 1410–1420. https://doi.org/10.1523/jneurosci.4098-07.2008

Miller, J. A., Horvath, S., & Geschwind, D. H. (2010). Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. Proceedings of the National Academy of Sciences, 107(28), 12698–12703. https://doi.org/10.1073/pnas.0914257107

Miller, J. A., Ding, S.-L., Sunkin, S. M., Smith, K. A., Ng, L., Szafer, A., … Lein, E. S. (2014). Transcriptional landscape of the prenatal human brain. Nature, 508(7495), 199–206. https://doi.org/10.1038/nature13185
Mo, A., Mukamel, E. A., Davis, F. P., Luo, C., Henry, G. L., Picard, S., … Nathans, J. (2015). Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. Neuron, 86(6), 1369–1384. https://doi.org/10.1016/j.neuron.2015.05.018

Monaghan, C. E., Nechiporuk, T., Jeng, S., McWeeney, S. K., Wang, J., Rosenfeld, M. G., & Mandel, G. (2017). REST corepressors RCOR1 and RCOR2 and the repressor INSM1 regulate the proliferation–differentiation balance in the developing brain. Proceedings of the National Academy of Sciences, 114(3), E406–E415. https://doi.org/10.1073/pnas.1620230114

Moslem, M., Olive, J., & Falk, A. (2019). Stem cell models of schizophrenia, what have we learned and what is the potential? Schizophrenia Research, 210, 3–12. https://doi.org/10.1016/j.schres.2018.12.023

Mostafavi, S., Battle, A., Zhu, X., Urban, A. E., Levinson, D., Montgomery, S. B., & Koller, D. (2013). Normalizing RNA-Sequencing Data by Modeling Hidden Covariates with Prior Knowledge. PLoS ONE, 8(7), e68141. https://doi.org/10.1371/journal.pone.0068141

Mostafavi, S., Gaiteri, C., Sullivan, S. E., White, C. C., Tasaki, S., Xu, J., … De Jager, P. L. (2018). A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. Nature Neuroscience, 21(6), 811–819. https://doi.org/10.1038/s41593-018-0154-9

Myers, C. L., Chiriac, C., & Troyanskaya, O. G. (2009). Discovering Biological Networks from Diverse Functional Genomic Data. In Methods in Molecular Biology (pp. 157–175). Humana Press. https://doi.org/10.1007/978-1-60761-175-2_9

Negi, S. K., & Guda, C. (2017). Global gene expression profiling of healthy human brain and its application in studying neurological disorders. Scientific Reports, 7(1). https://doi.org/10.1038/s41598-017-00952-9

Nehme, R., Zuccaro, E., Ghosh, S. D., Li, C., Sherwood, J. L., Pietilainen, O., … Eggan, K. (2018). Combining NGN2 Programming with Developmental Patterning Generates Human Excitatory Neurons with NMDAR-Mediated Synaptic Transmission. Cell Reports, 23(8), 2509–2523. https://doi.org/10.1016/j.celrep.2018.04.066

Nguyen, H. T., Bryois, J., Kim, A., Dobbyn, A., Huckins, L. M., Munoz-Manchado, A. B., … Stahl, E. A. (2017). Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. Genome Medicine, 9(1). https://doi.org/10.1186/s13073-017-0497-y

Nicholson-Fish, J. C., Kokotos, A. C., Gillingwater, T. H., Smillie, K. J., & Cousin, M. A. (2015). VAMP4 Is an Essential Cargo Molecule for Activity-Dependent Bulk Endocytosis. Neuron, 88(5), 973–984. https://doi.org/10.1016/j.neuron.2015.10.043

Nowakowski, T. J., Bhaduri, A., Pollen, A. A., Alvarado, B., Mostajo-Radji, M. A., Di Lullo, E., … Kriegstein, A. R. (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. Science, 358(6368), 1318–1323. https://doi.org/10.1126/science.aap8809

Okbay, A., Baselmans, B. M. L., De Neve, J.-E., Turley, P., Nivard, M. G., … Cesarini, D. (2016). Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. Nature Genetics, 48(6), 624–633. https://doi.org/10.1038/ng.3552

Okun, E., Griffioen, K., Barak, B., Roberts, N. J., Castro, K., Pita, M. A., … Mattson, M. P. (2010). Toll-like receptor 3 inhibits memory retention and constrains adult hippocampal neurogenesis. Proceedings of the National Academy of Sciences, 107(35), 15625–15630. https://doi.org/10.1073/pnas.1005807107

Oldham, M. C., Horvath, S., & Geschwind, D. H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proceedings of the National Academy of Sciences, 103(47), 17973–17978. https://doi.org/10.1073/pnas.0605938103

Oldham, M. C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., & Geschwind, D. H. (2008). Functional organization of the transcriptome in human brain. Nature Neuroscience, 11(11), 1271–1282. https://doi.org/10.1038/nn.2207

Oldham, M. C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., & Geschwind, D. H. (2008). Functional organization of the transcriptome in human brain. Nature Neuroscience, 11(11), 1271–1282. https://doi.org/10.1038/nn.2207

Pardiñas, A. F., Holmans, P., Pocklington, A. J., Escott-Price, V., … Ripke, S. (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nature Genetics, 50(3), 381–389. https://doi.org/10.1038/s41588-018-0059-2

Parikshak, N. N., Luo, R., Zhang, A., Won, H., Lowe, J. K., Chandran, V., … Geschwind, D. H. (2013). Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. Cell, 155(5), 1008–1021. https://doi.org/10.1016/j.cell.2013.10.031

Parikshak, N. N., Gandal, M. J., & Geschwind, D. H. (2015). Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. Nature Reviews Genetics, 16(8), 441–458. https://doi.org/10.1038/nrg3934

Parikshak, N. N., Swarup, V., Belgard, T. G., Irimia, M., Ramaswami, G., Gandal, M. J., … Geschwind, D. H. (2016). Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. Nature, 540(7633), 423–427. https://doi.org/10.1038/nature20612

Pedregosa, F., Varoquax, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Cournapeu, D. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825-2830

Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., & Goldstein, D. B. (2013). Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLoS Genetics, 9(8), e1003709. https://doi.org/10.1371/journal.pgen.1003709

Cross-disorder Working Group of the Psychiatric Genomics Consortium (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. The Lancet, 381(9875), 1371–1379. https://doi.org/10.1016/s0140-6736(12)62129-1

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature, 511(7510), 421–427. https://doi.org/10.1038/nature13595

Autism Working Group of the Psychiatric Genomics Consortium (2017). Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. Molecular Autism, 8(1). https://doi.org/10.1186/s13229-017-0137-9

Phoenix, T. N., & Temple, S. (2010). Spred1, a negative regulator of Ras-MAPK-ERK, is enriched in CNS germinal zones, dampens NSC proliferation, and maintains ventricular zone structure. Genes & Development, 24(1), 45–56. https://doi.org/10.1101/gad.1839510

Pierson, E., Koller, D., Battle, A., & Mostafavi, S. (2015). Sharing and Specificity of Co-expression Networks across 35 Human Tissues. PLOS Computational Biology, 11(5), e1004220. https://doi.org/10.1371/journal.pcbi.1004220

Polioudakis, D., de la Torre-Ubieta, L., Langerman, J., Elkins, A. G., Shi, X., Stein, J. L., … Geschwind, D. H. (2019). A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. Neuron. https://doi.org/10.1016/j.neuron.2019.06.011

Pollard, K. S., Salama, S. R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J. S., … Haussler, D. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. Nature, 443(7108), 167–172. https://doi.org/10.1038/nature05113

Ponomarev, I., Wang, S., Zhang, L., Harris, R. A., & Mayfield, R. D. (2012). Gene Coexpression Networks in Human Brain Identify Epigenetic Modifications in Alcohol Dependence. Journal of Neuroscience, 32(5), 1884–1897. https://doi.org/10.1523/jneurosci.3136-11.2012

Prudencio, M., Belzil, V. V., Batra, R., Ross, C. A., Gendron, T. F., Pregent, L. J., … Petrucelli, L. (2015). Distinct brain transcriptome profiles in C9orf72-associated and sporadic ALS. Nature Neuroscience, 18(8), 1175–1182. https://doi.org/10.1038/nn.4065

Quesnel-Vallières, M., Dargaei, Z., Irimia, M., Gonatopoulos-Pournatzis, T., Ip, J. Y., Wu, M., … Cordes, S. P. (2016). Misregulation of an Activity-Dependent Splicing Network as a Common Mechanism Underlying Autism Spectrum Disorders. Molecular Cell, 64(6), 1023–1034. https://doi.org/10.1016/j.molcel.2016.11.033

Radulescu, E., Jaffe, A. E., Straub, R. E., Chen, Q., Shin, J. H., Hyde, T. M., … Weinberger, D. R. (2018). Identification and prioritization of gene sets associated with schizophrenia risk by co-expression network analysis in human brain. Molecular Psychiatry. https://doi.org/10.1038/s41380-018-0304-1

Ramasamy, A., Trabzuni, D., Guelfi, S., Varghese, V., … Smith, C. (2014). Genetic variability in the regulation of gene expression in ten regions of the human brain. Nature Neuroscience, 17(10), 1418–1428. https://doi.org/10.1038/nn.3801

Romanov, R. A., Zeisel, A., Bakker, J., Girach, F., Hellysaz, A., Tomer, R., … Harkany, T. (2016). Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. Nature Neuroscience, 20(2), 176–188. https://doi.org/10.1038/nn.4462

Rossin, E. J., Lage, K., Raychaudhuri, S., Xavier, R. J., Tatar, D., … Benita, Y. (2011). Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. PLoS Genetics, 7(1), e1001273. https://doi.org/10.1371/journal.pgen.1001273

Ruan, J., Dean, A. K., & Zhang, W. (2010). A general co-expression network-based approach to gene expression analysis: comparison and applications. BMC Systems Biology, 4(1). https://doi.org/10.1186/1752-0509-4-8

Ruzzo, E. K., Pérez-Cano, L., Jung, J.-Y., Wang, L., Kashef-Haghighi, D., Hartl, C., … Wall, D. P. (2019). Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks. Cell, 178(4), 850–866.e26. https://doi.org/10.1016/j.cell.2019.07.015

Saelens, W., Cannoodt, R., & Saeys, Y. (2018). A comprehensive evaluation of module detection methods for gene expression data. Nature Communications, 9(1). https://doi.org/10.1038/s41467-018-03424-4

Saha, A., Kim, Y., Gewirtz, A. D. H., Jo, B., Gao, C., … McDowell, I. C. (2017). Co-expression networks reveal the tissue-specific regulation of transcription and splicing. Genome Research, 27(11), 1843–1858. https://doi.org/10.1101/gr.216721.116

Schafer, S. T., Paquola, A. C. M., Stern, S., Gosselin, D., Ku, M., Pena, M., … Gage, F. H. (2019). Pathological priming causes developmental gene network heterochronicity in autistic subject-derived neurons. Nature Neuroscience, 22(2), 243–255. https://doi.org/10.1038/s41593-018-0295-x

Schanzenbächer, C. T., Langer, J. D., & Schuman, E. M. (2018). Time- and polarity-dependent proteomic changes associated with homeostatic scaling at central synapses. eLife, 7. https://doi.org/10.7554/elife.33322

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science, 270(5235), 467–470. https://doi.org/10.1126/science.270.5235.467

Schijven, D., Kofink, D., Tragante, V., Verkerke, M., Pulit, S. L., Kahn, R. S., … Luykx, J. J. (2018). Comprehensive pathway analyses of schizophrenia risk loci point to dysfunctional postsynaptic signaling. Schizophrenia Research, 199, 195–202. https://doi.org/10.1016/j.schres.2018.03.032

Schoech, A. P., Jordan, D. M., Loh, P.-R., Gazal, S., O'Connor, L. J., Balick, D. J., … Price, A. L. (2019). Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. Nature Communications, 10(1). https://doi.org/10.1038/s41467-019-08424-6

Schreiber, A. W., Shirley, N. J., Burton, R. A., & Fincher, G. B. (2008). Combining transcriptional datasets using the generalized singular value decomposition. BMC Bioinformatics, 9(1). https://doi.org/10.1186/1471-2105-9-335

Schulze, T. & McMahon, F. (2018). Psychiatric genetics : a primer for clinical and basic scientists. New York, NY: Oxford University Press.

Scott-Boyer, M.-P., Haibe-Kains, B., & Deschepper, C. F. (2013). Network statistics of genetically-driven gene co-expression modules in mouse crosses. Frontiers in Genetics, 4. https://doi.org/10.3389/fgene.2013.00291

Sears, J. C., & Broadie, K. (2018). Fragile X Mental Retardation Protein Regulates Activity-Dependent Membrane Trafficking and Trans-Synaptic Signaling Mediating Synaptic Remodeling. Frontiers in Molecular Neuroscience, 10. https://doi.org/10.3389/fnmol.2017.00440

Segrè, A. V., Groop, L., Mootha, V. K., Daly, M. J., & Altshuler, D. (2010). Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits. PLoS Genetics, 6(8), e1001058. https://doi.org/10.1371/journal.pgen.1001058

Selten, M., van Bokhoven, H., & Nadif Kasri, N. (2018). Inhibitory control of the excitatory/inhibitory balance in psychiatric disorders. F1000Research, 7, 23. https://doi.org/10.12688/f1000research.12155.1

Shen, E. H., Overly, C. C., & Jones, A. R. (2012). The Allen Human Brain Atlas. Trends in Neurosciences, 35(12), 711–714. https://doi.org/10.1016/j.tins.2012.09.005

Si, Y., Liu, P., Li, P., & Brutnell, T. P. (2013). Model-based clustering for RNA-seq data. Bioinformatics, 30(2), 197–205. https://doi.org/10.1093/bioinformatics/btt632

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A Sparse-Group Lasso. Journal of Computational and Graphical Statistics, 22(2), 231–245. https://doi.org/10.1080/10618600.2012.681250

Skene, N. G., Bryois, J., Bakken, T. E., Breen, G., Crowley, J. J., … Hjerling-Leffler, J. (2018). Genetic identification of brain cell types underlying schizophrenia. Nature Genetics, 50(6), 825–833. https://doi.org/10.1038/s41588-018-0129-5

Somekh, J., Shen-Orr, S. S., & Kohane, I. S. (2019). Batch correction evaluation framework using a-priori gene-gene associations: applied to the GTEx dataset. BMC Bioinformatics, 20(1). https://doi.org/10.1186/s12859-019-2855-9

Song, W.-M., & Zhang, B. (2015). Multiscale Embedded Gene Co-expression Network Analysis. PLOS Computational Biology, 11(11), e1004574. https://doi.org/10.1371/journal.pcbi.1004574

Sousa, A. M. M., Zhu, Y., Raghanti, M. A., Kitchen, R. R., Onorati, M., Tebbenkamp, A. T. N., … Sestan, N. (2017). Molecular and cellular reorganization of neural circuits in the human lineage. Science, 358(6366), 1027–1032. https://doi.org/10.1126/science.aan3456

Stegle, O., Parts, L., Piipari, M., Winn, J., & Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nature Protocols, 7(3), 500–507. https://doi.org/10.1038/nprot.2011.457

Stilling, R. M., Moloney, G. M., Ryan, F. J., Hoban, A. E., Bastiaanssen, T. F., Shanahan, F., … Cryan, J. F. (2018). Social interaction-induced activation of RNA splicing in the amygdala of microbiome-deficient mice. eLife, 7. https://doi.org/10.7554/elife.33070

Su, Y., Shin, J., Zhong, C., Wang, S., Roychowdhury, P., Lim, J., … Song, H. (2017). Neuronal activity modifies the chromatin accessibility landscape in the adult brain. Nature Neuroscience, 20(3), 476–483. https://doi.org/10.1038/nn.4494

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

Sullivan, P. F., Daly, M. J., & O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nature Reviews Genetics, 13(8), 537–551. https://doi.org/10.1038/nrg3240

Sullivan, P. F., & Geschwind, D. H. (2019). Defining the Genetic, Genomic, Cellular, and Diagnostic Architectures of Psychiatric Disorders. Cell, 177(1), 162–183. https://doi.org/10.1016/j.cell.2019.01.015

Sunkin, S. M., Ng, L., Lau, C., Dolbeare, T., Gilbert, T. L., Thompson, C. L., … Dang, C. (2012). Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. Nucleic Acids Research, 41(D1), D996–D1008. https://doi.org/10.1093/nar/gks1042

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., … von Mering, C. (2016). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Research, 45(D1), D362–D368. https://doi.org/10.1093/nar/gkw937

Tasic, B., Menon, V., Nguyen, T. N., Kim, T. K., Jarsky, T., Yao, Z., … Zeng, H. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nature Neuroscience, 19(2), 335–346. https://doi.org/10.1038/nn.4216

Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes. Proceedings of Artificial Intelligence and Statistics. 567-574.

Torkamani, A., Dean, B., Schork, N. J., & Thomas, E. A. (2010). Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia. Genome Research, 20(4), 403–412. https://doi.org/10.1101/gr.101956.109

Tsai, N.-P., Wilkerson, J. R., Guo, W., Maksimova, M. A., DeMartino, G. N., Cowan, C. W., & Huber, K. M. (2012). Multiple Autism-Linked Genes Mediate Synapse Elimination via Proteasomal Degradation of a Synaptic Scaffold PSD-95. Cell, 151(7), 1581–1594. https://doi.org/10.1016/j.cell.2012.11.040

Turner, T. N., Yi, Q., Krumm, N., Huddleston, J., Hoekzema, K., F. Stessman, H. A., … Eichler, E. E. (2016). denovo-db: a compendium of humande novovariants. Nucleic Acids Research, 45(D1), D804–D811. https://doi.org/10.1093/nar/gkw865

Tyssowski, K. M., DeStefino, N. R., Cho, J.-H., Dunn, C. J., Poston, R. G., Carty, C. E., … Gray, J. M. (2018). Different Neuronal Activity Patterns Induce Different Gene Expression Programs. Neuron, 98(3), 530–546.e11. https://doi.org/10.1016/j.neuron.2018.04.001

van Dam, S., Võsa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2017). Gene co-expression analysis for functional classification and gene–disease predictions. Briefings in Bioinformatics, bbw139. https://doi.org/10.1093/bib/bbw139

Vergara, C., Parker, M. M., Franco, L., Cho, M. H., Valencia-Duarte, A. V., Beaty, T. H., & Duggal, P. (2018). Genotype imputation performance of three reference panels using African ancestry individuals. Human Genetics, 137(4), 281–292. https://doi.org/10.1007/s00439-018-1881-4

Voineagu, I., Wang, X., Johnston, P., Lowe, J. K., Tian, Y., Horvath, S., … Geschwind, D. H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature, 474(7351), 380–384. https://doi.org/10.1038/nature10110

Wang, Y. X. R., & Huang, H. (2014). Review on statistical methods for gene network reconstruction using expression data. Journal of Theoretical Biology, 362, 53–61. https://doi.org/10.1016/j.jtbi.2014.03.040

Wang, M., Roussos, P., McKenzie, A., Zhou, X., Kajiwara, Y., Brennand, K. J., … Zhang, B. (2016). Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. Genome Medicine, 8(1). https://doi.org/10.1186/s13073-016-0355-3

Wang, Q., Chen, R., Cheng, F., Wei, Q., Ji, Y., Yang, H., … Li, B. (2019). A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. Nature Neuroscience, 22(5), 691–699. https://doi.org/10.1038/s41593-019-0382-7

Wei, C.-W., Luo, T., Zou, S.-S., & Wu, A.-S. (2018). The Role of Long Noncoding RNAs in Central Nervous System and Neurodegenerative Diseases. Frontiers in Behavioral Neuroscience, 12. https://doi.org/10.3389/fnbeh.2018.00175

Weinstein, M.R., Histopathological changes in the brain in schizophrenia; a critical review. AMA Arch Neurol Psychiatry. 1954;71(5):539-53.

Wenzel, E. M., Morton, A., Ebert, K., Welzel, O., Kornhuber, J., Cousin, M. A., & Groemer, T. W. (2012). Key Physiological Parameters Dictate Triggering of Activity-Dependent Bulk Endocytosis in Hippocampal Synapses. PLoS ONE, 7(6), e38188. https://doi.org/10.1371/journal.pone.0038188

Willsey, A. J., Sanders, S. J., Li, M., Dong, S., Tebbenkamp, A. T., Muhle, R. A., … State, M. W. (2013). Coexpression Networks Implicate Human Midfetal Deep Cortical Projection Neurons in the Pathogenesis of Autism. Cell, 155(5), 997–1007. https://doi.org/10.1016/j.cell.2013.10.020

Willsey, A. J., Morris, M. T., Wang, S., Willsey, H. R., Sun, N., Teerikorpi, N., … Krogan, N. J. (2018). The Psychiatric Cell Map Initiative: A Convergent Systems Biological Approach to Illuminating Key Molecular Pathways in Neuropsychiatric Disorders. Cell, 174(3), 505–520. https://doi.org/10.1016/j.cell.2018.06.016

Won, H., de la Torre-Ubieta, L., Stein, J. L., Parikshak, N. N., Huang, J., Opland, C. K., … Geschwind, D. H. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. Nature, 538(7626), 523–527. https://doi.org/10.1038/nature19847

Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J., & Visscher, P. M. (2018). Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. Cell, 173(7), 1573–1580. https://doi.org/10.1016/j.cell.2018.05.051

Wu, K., Ren, R., Su, W., Wen, B., Zhang, Y., Yi, F., … Chen, C. (2014). A Novel Suppressive Effect of Alcohol Dehydrogenase 5 in Neuronal Differentiation. Journal of Biological Chemistry, 289(29), 20193–20199. https://doi.org/10.1074/jbc.c114.561860

Wu, H. C., Yamankurt, G., Luo, J., Subramaniam, J., Hashmi, S. S., Hu, H., & Cunha, S. R. (2015). Identification and characterization of two ankyrin-B isoforms in mammalian heart. Cardiovascular Research, 107(4), 466–477. https://doi.org/10.1093/cvr/cvv184

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A Tool for Genome-wide Complex Trait Analysis. The American Journal of Human Genetics, 88(1), 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Yoon, S., Nguyen, H. C. T., Yoo, Y. J., Kim, J., Baik, B., Kim, S., … Nam, D. (2018). Efficient pathway enrichment and network analysis of GWAS summary data using GSA-SNP2. Nucleic Acids Research, 46(10), e60–e60. https://doi.org/10.1093/nar/gky175

Yuen, R. K., Merico, D., Cao, H., Pellecchia, G., Alipanahi, B., Thiruvahindrapuram, B., … Scherer, S. W. (2016). Genome-wide characteristics of de novo mutations in autism. Npj Genomic Medicine, 1(1). https://doi.org/10.1038/npjgenmed.2016.27

Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., … Linnarsson, S. (2018). Molecular Architecture of the Mouse Nervous System. Cell, 174(4), 999–1014.e22. https://doi.org/10.1016/j.cell.2018.06.021

Zelentsova, K., Talmi, Z., Abboud-Jarrous, G., Sapir, T., Capucha, T., Nassar, M., & Burstyn-Cohen, T. (2016). Protein S Regulates Neural Stem Cell Quiescence and Neurogenesis. STEM CELLS, 35(3), 679–693. https://doi.org/10.1002/stem.2522

Zhang, B., & Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis. Statistical Applications in Genetics and Molecular Biology, 4(1). https://doi.org/10.2202/1544-6115.1128

Zhang, Y., Chen, K., Sloan, S. A., Bennett, M. L., Scholze, A. R., O'Keeffe, S., … Wu, J. Q. (2014). An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. Journal of Neuroscience, 34(36), 11929–11947. https://doi.org/10.1523/jneurosci.1860-14.2014

Zhang, Y., Sloan, S. A., Clarke, L. E., Caneda, C., Plaza, C. A., Blumenthal, P. D., … Barres, B. A. (2016). Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. Neuron, 89(1), 37–53. https://doi.org/10.1016/j.neuron.2015.11.013

Zuo, L., Tan, Y., Wang, Z., Wang, K.-S., Zhang, X., Chen, X., … Luo, X. (2016). Long noncoding RNAs in psychiatric disorders. Psychiatric Genetics, 26(3), 109–116. https://doi.org/10.1097/ypg.0000000000000129