Genetics and Pathway Analysis of Normative Cognitive Variation in the Philadelphia Neurodevelopmental Cohort

Authors:
Shraddha Pai[1], Shirley Hui[1], Philipp Weber[2], Owen Whitley[1,3], Peipei Li[4,5], Viviane Labrie[4,5], Jan Baumbach[2,6], Gary D Bader[1,3,7,8]

Affiliations:

1. The Donnelly Centre, University of Toronto, Toronto, Canada
2. Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark
3. Department of Molecular Genetics, University of Toronto, Toronto, Canada
4. Center for Neurodegenerative Science, Van Andel Research Institute, Grand Rapids, MI, USA
5. Division of Psychiatry and Behavioral Medicine, College of Human Medicine, Michigan State University, Grand Rapids, MI, USA
6. TUM School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany.
7. Department of Computer Science, University of Toronto, Toronto, Canada
8. The Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Canada

* gary.bader@utoronto.ca

Abstract

Identifying genes and cellular pathways associated with normative brain physiology and behavior could help discover molecular therapies that target specific psychiatric symptoms with minimal side effects. Linking genotype-phenotype associations from population-scale datasets to brain function is challenging because of the multi-level, heterogeneous nature of brain organization. To address this challenge, we developed a novel brain-focused gene and pathway prioritization workflow, which maps variants to genes based on knowledge of brain genome regulation, and subsequently to pathways, cells, diseases and drugs (21 resources). We applied this workflow to nine cognitive tasks from the Philadelphia Neurodevelopmental Cohort (subset of 3,319 individuals aged 8-21 years). We report genome-wide significance of variants associated with nonverbal reasoning within the 3' end of the *FBLN1* gene ($p=4.6x10^{-8}$), itself linked to fetal neurodevelopment and psychotic disorders. These findings suggest that nonverbal reasoning and *FBLN1* variation warrant further investigation in studies of psychosis. Multiple cognitive tasks demonstrated significant enrichment of variants in cellular pathways and brain-related gene sets, such as organ development, cell proliferation and nervous system dysfunction. Top-ranking genes in working memory associated pathways are genetically associated with multiple diseases with working memory deficits, including schizophrenia and Parkinson's disease, and with multiple drugs, suggesting that choice of therapy for memory deficits should consider disease context. Given the large amount of additional biological insight derived from our pathway analysis, versus a standard gene-based approach, we propose that "genes to behaviour" frameworks for modeling brain-related phenotypes, like RDoC, should include pathway information to create a "genes to pathways to behaviour" approach. Our workflow is broadly useful to put genotype-phenotype associations of brain-related phenotypes into the context of brain organization, function, disease and known molecular therapies.

Introduction

A major drive in the field of psychiatry is the reconceptualization of mental illnesses, diseases traditionally classified on the basis of clinical descriptions, as brain disorders treatable by neurobiologically-grounded therapies. The U.S. National Institute of Mental Health developed Research Domain Criteria (or RDoC), a framework to build a "genes to behaviour" model of the human brain that deconstructs behaviour into multiple domains mediated by different neuroanatomical regions, local cellular circuits and molecules (https://www.nimh.nih.gov/research-priorities/rdoc/index.shtml ; [1]). The ambition of the RDoC framework is to develop neurobiologically-grounded taxonomy, biomarkers and treatments for mental illness to support use of precision medicine in psychiatry[2]. The neurobehavioural framework has been embraced by psychiatric researchers at multiple levels of brain research, including cross-disorder genome-wide association studies and genetic risk prediction[3,4], and identification of genetic contributors to neuroimaging-based measures of brain activity and structure associated with disease[5]. Population-scale datasets that measure genotype and cognition-related phenotypes, such as the Philadelphia Neurodevelopmental Cohort[6,7], the Adolescent Brain Cognitive Development (ABCD) dataset[8] and UK Biobank[9], provide an attractive resource to build a molecules-to-behaviour model for brain-specific phenotypes. Moreover, maps of brain-specific genome regulation, such as those generated by the GTEx[10], NIH Roadmap Epigenomics[11] and PsychENCODE[12] projects, now enable the effect of genetic variants to be interpreted in brain-relevant neuroanatomical and developmental contexts. However, integrating genotype-phenotype associations with these data resources to methodically infer variant impact on various levels of brain organization represents a major challenge, due to the large number of complex data sets that need to be integrated.

In this work, we develop a novel brain-focused computational analysis workflow to identify genes, pathways and cellular functions, as well as gene-related brain functions, diseases and drugs. We apply this workflow to identify genes and functions associated with normative variation in nine cognitive phenotypes from the Philadelphia Neurodevelopmental Cohort (PNC). To our knowledge, little is known about the molecular basis of different cognitive phenotypes in humans, and the extent to which molecular and cellular players overlap across these. With extensive neurobehavioural and genotyping data available on 8,000 community youths aged 8-21 years, the PNC represents the largest publicly-available dataset of its kind for genotype-phenotype analysis of cognition[6,7]. All participants have computerized neurocognitive test battery (CNB) scores which measures speed and accuracy in multiple cognitive domains (e.g. emotion processing, executive function), and which has neurobehavioural validity (i.e. tasks known to activate specific brain regions), SNP-based heritability[13], and disease relevance[3,14]. The CNB has also been characterized for demographic effects[15] and neuropsychological validation[16], altogether providing a well-characterized set of phenotypes to study the genetic basis of specific cognitive abilities. While a number of CNB phenotypes demonstrate significant SNP-based heritability[13], and reduced test scores have been genetically associated with psychiatric disease risk[3], there has not been a methodical examination of the molecular players involved in individual phenotypes. We reasoned that identifying the genes, pathways, cellular and developmental context associated with these phenotypes could pinpoint genetic crosstalk between individual cognitive tasks and psychiatric and neurological diseases, and provide hypotheses for molecular therapy of corresponding cognitive impairments in disease.

Methods

Genetic imputation

The workflow for genomic imputation is shown in Supplementary Figure 1. Genotypes for four microarray genotyping platforms were downloaded from dbGaP (phs000607.v1). We performed genetic imputation for the Illumina Human610-Quad BeadChip, the Illumina HumanHap550 Genotyping BeadChip v1.1, Illumina HumanHap550 Genotyping BeadChip v3, and the Affymetrix AxiomExpress platform (Supplementary Table 1, total of 6,502 samples before imputation), using the protocol recommended by the EMERGE consortium[17]. Imputation was performed as follows:

Step 1: Platform-specific plink Quality Control: Quality control was first performed for each microarray platform. Single nucleotide polymorphisms (SNPs) were limited to those on chr1-22. SNPs in linkage disequilibium (LD) were excluded (--indep-pairwise 50 5 0.2), and alleles were recoded from numeric to letter (ACGT) coding. Samples were excluded if they demonstrated heterozygosity > 3 standard deviations (SD) from the mean, or if they were missing >=5% genotypes. Where samples had pairwise Identity by Descent (IBD) > 0.185, one of the pair was excluded. Variants with minor allele frequency (MAF) < 0.05 were excluded, as were those failing Hardy-Weinberg equilibrium with p < 1e-6 and those missing in >=5% samples.

Step 2: Convert coordinates to hg19. LiftOver[18] was used to convert SNPs from hg18 to hg19; Hap550K v1 data was in hg17 and was converted from this build to hg19.

Step 3: Strand-match check and prephasing: ShapeIt v2.r790[19] was used to confirm that the allelic strand in the input data matched that in the reference panel; where it did not, allele strands were flipped (shapeit "–check" flag). ShapeIt was used to prephase the variants using the genetic_b37 reference panel (downloaded from the Shapeit website, http://www.shapeit.fr/files/genetic_map_b37.tar.gz)

Step 4: Imputation: Genotypes were imputed using Impute2 v2.3.2[20] and a reference panel from the 1,000 Genomes (phase 1, prephased with Shapeit2, no singletons, 16 June 2014 release, downloaded from https://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated_SHAPEIT2_16-06-14.html) was used for imputation, using the parameter settings "–use_prephased_g –Ne 20000 –seed 367946". Average concordance for all chromosomes was ~95%, indicating successful imputation (Supplementary Figure 2). Imputed genotypes were merged across all platforms using software from the Ritchie lab[17] (impute2-group-join.py, from https://ritchielab.org/software/imputation-download) and converted to plink format. Following previous PNC genotype analysis[13], only SNPs with info score > 0.6 were retained, and deletions/insertions were excluded (plink "-snps-only just-acgt" flags). As preliminary quality control, when merging across chromosomes, samples with missingness exceeding 99% were excluded, as were SNPs with MAF < 1% and with missingness exceeding 99%. This step resulted in 10,845,339 SNPs and 6,327 individuals.

Step 5: Post-imputation Quality Control: The HapMap3 panel was used to assign genetic ancestry for samples, using steps from [21] (Supplementary Figure 3). Individuals within 5 SD of the centroid of the HapMap3 CEU (Utah residents with Northern or Western European ancestry) or TSI (Tuscans in Italy) cluster were assigned to belong to the respective groups, and were classified as being of European descent; 3,441 individuals pass this filter. Individuals with >5% missing data were excluded, as was one of each pair of individuals with IBS > 0.185 (47 individuals); 3,394 individuals passed this filter. Variants that were symmetric or in regions of high LD (Supplementary Table 2) were excluded (9,631,316 SNPs passed). Variants with >5% missingness were excluded (1,569,407 SNPs excluded). Finally, SNPs with MAF < 0.01 (3,168,339 SNPs) and failing Hardy-Weinberg equilibrium (HWE) with p value < 1e-6 (373 SNPs) were excluded, resulting in 4,893,197 SNPs. Unlike Verma et al, quality control steps were performed once, rather than repeated after samples were excluded. In sum, the

imputation process resulted in 3,394 individuals and 4,893,197 SNPs available for downstream analysis.

Phenotype processing

Phenotype data was downloaded from dbGaP for 8,719 individuals. 637 individuals with severe medical conditions (Medical rating=4) were excluded to avoid confounding the symptoms of their conditions with performance on the cognitive tests[13]. Linear regression was used to regress out the effect of age at test time (variable name: "age at cnb") and sex from sample-level phenotype scores, and the residualized phenotype was used for downstream analysis.

The nine phenotypes selected for pathway analysis were measures of overall performance accuracy in Penn Computerized Neurocognitive Test Battery (CNB; Supplementary Table 3) and represented major cognitive domains. Following regression, none of the phenotypes were significantly correlated with age after Bonferroni correction, indicating that the age effect had been reduced (Supplementary Table 4). Following guidelines from previous analyses on these data[3], individuals with scores more than four standard deviations from the mean for a particular test were excluded from the analysis of the corresponding phenotype. For a given phenotype, only samples with a code indicating a valid test score (codes "V" or "V2") were included; e.g. for pfmt_tp (Penn Face Memory Test), only samples with pfmt_valid = "V" or "V2" were retained; the rest had scores set to NA. Finally, each phenotype was dichotomized so that samples in the bottom 33rd percentile were relabeled as "poor" performers and those in the top 33rd were set to be "good" performers; for a given phenotype, this process resulted in ~1,000 samples in each group (Supplementary Table 3). Where an individual had good or poor performance in multiple phenotypes, they were included in the corresponding group for each of those phenotypes.

Genetic association analysis

For each of 9 CNB phenotypes, marginal SNP-level association was calculated using a mixed-effects linear model (MLMA), using the leave-one-chromosome-out (LOCO) method of estimating polygenic contribution (GCTA v1.97.7beta software[22]). In this strategy, a mixed-effect model is fit for each SNP:

$$y = a + bx + g\text{-} + e$$

In this model, y is the binarized label (good/poor performers on a particular task), x measures the effect of genotype (indicator variable encoded as 0, 1 or 2), and g- represents the polygenic contribution of all the SNPs in the genome (here, the ~4.89M imputed SNPs). In the LOCO variation, g- is calculated using a chromosome-specific genetic relatedness matrix, one that excludes the chromosome on which the candidate SNP is located[22]. SNPs and associated genes were annotated as described in Supplementary Notes 1-4.

Hi-C Data Processing

We generated Hi-C data from the human prefrontal cortex[23] (Illumina HiSeq 2000 paired-end raw sequence reads; n=1 sample; 746 Million reads; accession: GSM2322542 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2322542]). Hi-C analysis involved Trim Galore[24] (v0.4.3) for adapter trimming, HICUP[25] (v0.5.9) for mapping and performing quality control, and GOTHIC [26] for identifying significant interactions (Bonferroni p <0.05), with a 40 kb resolution. Hi-C gene annotation involved identifying interactions with gene promoters, defined as ± 2 kb of a gene TSS. This analysis identified 303,464 interactions used for our study.

SNP to gene mapping for annotation and enrichment analyses

SNPs were mapped to genes using a combination of positional information, brain-specific expression Quantitative Trait Locus (eQTL) and higher-order chromatin interaction (hi-C) information. For eQTL-based mapping, we limited the search to significant eQTLs in brain tissue (GTEx v7 brain anterior cingulate cortex BA24, brain cortex, brain frontal cortex BA9, and hippocampus; downloaded from https://www.gtexportal.org; Supplementary Note 1[10]); of these, only SNPs located in open chromatin regions of brain-related samples were included (Roadmap Epigenomics 15-core chromatin state <=7)[11]. These included maps derived from neurospheres, angular gyrus, anterior caudate, germinal matrix, hippocampus, inferior temporal lobe, dorsolateral prefrontal cortex, substantia nigra, and fetal brain of both sexes (samples E053, E054, E067, E068, E069, E070, E071, E072, E073, E074, E081, E082, and E125), downloaded from http://www.roadmapepigenomics.org/. For 3D chromatin interaction mapping, SNPs were mapped to genes if these were located within a region where higher-order interaction was ascertained in the dorsolateral prefrontal cortex[27]; this region was constrained to be 250bp upstream and 500bp downstream of the gene's transcription start site; of these SNPs, only those overlapping brain enhancers were included[11]. These included enhancers in angular gyrus, hippocampus, inferior temporal lobe, and dorsolateral prefrontal cortex (samples E067, E071, E072, and E073; chromatin state "Enh" or "EnhG"). Finally, SNPs were positionally mapped to the nearest gene if the shortest distance to either transcription start site or end site was 60kb. This cutoff was selected because it maps the majority (90%) of SNPs to their nearest gene.

The order of mapping was as follows: SNPs that mapped to a gene via brain eQTL or hi-C interactions were not also positionally mapped to a gene. A SNP was allowed to map to genes using both eQTL and hi-C, and where SNPs mapped to multiple genes all associations were retained. SNPs without eQTL or hi-C mappings were positionally mapped to a gene. Where a SNP mapped to multiple genes, all associations were retained. These SNP-gene mappings were used for the gene set enrichment analysis described below, as well as to annotate SNPs from the GWAS analysis.

Gene set enrichment analysis

For each of the nine CNB phenotypes, gene set enrichment analysis was performed using an implementation of GSEA for genetic variants[28,29]. GSEA was selected as it computes pathway enrichment scores using all available SNP information, which improves sensitivity, rather than using a hypergeometric model limited to SNPs passing a specific GWAS p-value cutoff. All SNPs were mapped to genes (as described in the "SNP-gene mapping for annotation and enrichment analyses" section) and each gene score is the best GWAS marginal p-value of all mapped SNPs. For each pathway, GSEA computes an enrichment score (ES) using the rank-sum of gene scores. The set of genes that appear in the ranked list before the rank-sum reaches its maximum deviation from zero, is called the "leading edge subset", and is interpreted as the core set of genes responsible for the pathway's enrichment signal. Following computation of the ES, a null distribution is created for each pathway by repeating genome-wide association tests with randomly label-permuted data and by computing ES from these permuted data; in this work, we use 100 permutations. Finally, the ES on the original data is normalized to the score computed for the same gene set for label-permuted data (Z-score of real ES relative to distribution of ES in label-permuted data), resulting in a Normalized Enrichment Score (NES) per pathway. The nominal p-value for the NES score is computed based on the null distribution and FDR correction is used to generate a q-value.

The first enrichment analysis used pathway information compiled from HumanCyc[30] (http://humancyc.org), NetPath (http://www.netpath.org)[31], Reactome (http://www.reactome.org)[32], NCI Curated Pathways[33], mSigDB[34] (http://software.broadinstitute.org/gsea/msigdb/), and Panther[35] (http://pantherdb.org/) and Gene Ontology[36]

(Human_GOBP_AllPathways_no_GO_iea_May_01_2018_symbol.gmt, downloaded from http://download.baderlab.org/EM_Genesets/May_01_2018/Human/symbol/Human_GOBP_AllPathways_no_GO_iea_May_01_2018_symbol.gmt); only pathways with 20-500 genes were used.

The second enrichment analysis used brain-related gene sets we compiled from various literature sources (see Supplementary Table 5 and Supplementary Note 5 for details). Gene sets included those identified through transcriptomic or proteomic assays in human brain tissue (i.e. direct measurement of expression), and genes associated with brain function by indirect inference (e.g. genetic association of nervous system disorders); both groups of gene sets were combined for this enrichment analysis. The transcriptomic/proteomic gene sets included: genes identified as markers for adult and fetal brain cell types through single-cell transcriptomic experiments[37-39], genes enriched for brain-specific expression (Human Protein Atlas project (https://www.proteinatlas.org[40]); genes co-expressed with markers of various stages of human brain development (BrainSpan[41]); and genes encoding proteins altered in the schizophrenia synaptosomal proteome[42]. Other gene sets included: genes associated with schizophrenia, bipolar disorder, autism spectrum disorder and major depressive disorder through large-scale genetic association studies by the Psychiatric Genomics Consortium (Supplementary Note 5); genes associated with nervous system disorders by the Human Phenotype Ontology[43]. Genes in the second group were filtered to only include genes with detectable expression (including long non-coding RNA genes) in the fetal[44] or adult human brain[40]. A total of 1,343 gene sets were collected. Only gene sets with 20-500 genes were included in the analysis; 421 gene sets met these criteria and were included in the enrichment analysis.

Enrichment map

An enrichment map was created to visualize the functional themes significant in enrichment analyses. We used the EnrichmentMap app v3.1.0[45] and Cytoscape v3.7.1[46] to create the map. Nodes in the map are pathways with FDR significance of FDR < 0.10 and edges in the map connect nodes with at least a gene set similarity of 0.375 (using Jaccard + Overlap similarity).

Leading edge gene interaction network

Genes contributing to pathway enrichment results (leading edge genes) were obtained as part of the implementation of GSEA for genetic variants[28]. The network was constructed from leading edge genes of pathways with q < 0.05. The online GeneMANIA service (v 3.6.0; https://genemania.org[47]) was used to obtain a gene-gene interaction network for leading edge genes (human database, default settings); the resulting network and edge attributes were downloaded. This network was imported into Cytoscape v3.7.1. Known drug associations were obtained from DGIdb[48] and GWAS associations with nervous system disorders were obtained from the NHGRI-EBI GWAS catalogue, via programmatic search using the TargetValidation.org API[49,50]. Cell type marker information was compiled from single cell RNA-seq datasets, including those for adult and fetal human brain[37-39].

Results

Figure 1a shows the workflow for the analysis performed in this work. Briefly, genotypes were imputed using a reference panel from the 1,000 Genomes Project[51], and samples were limited to those of European genetic ancestry (Supplementary Figure 1-3, Supplementary Table 1). 3,394 individuals and ~4.9M SNPs passed the quality control and imputation process. Following quality control of phenotype data, 3,116 European samples passed both genotype and phenotype filters and were included in downstream analyses. We selected nine phenotypes from the Penn Computerized Neurocognitive Test Battery (CNB) representing overall accuracy in four cognitive domains: complex

cognition, executive function, declarative memory, and social processing (Supplementary Table 3). Measures included performance for verbal reasoning, nonverbal reasoning, spatial reasoning, attention allocation, working memory, recall tests for faces, words and objects, and emotion identification[14]. In all instances, age and sex was regressed out of the phenotype (Supplementary Table 4) and samples were thereafter binarized into poor and good performers (bottom and top 33% percentile, respectively) resulting in ~1,000 samples per group for each phenotype (Supplementary Figure 4,5, Supplementary Table 3).

For each of the nine phenotypes, we first performed SNP-level genome-wide association analysis using a mixed-effects linear model that included genome-wide genetic ancestry as a covariate (GCTA[22]). Among the nine phenotypes, 661 SNPs had suggestive levels of significance at the genome-wide level ($p < 10^{-5}$; Figure 1b,c, Supplementary Figure 6,7, Supplementary Table 6). Over half of these SNPs are associated with tasks related to complex cognition (377 SNPs or 57%); 27% were associated with executive function (177 SNPs), 13% with declarative memory tasks (83 SNPs), and 4% with emotion identification (24 SNPs).

We mapped SNPs to genes using brain eQTL information, brain-specific higher-order chromatin interactions[10,27] and positional information. We integrated our findings with functional annotation maps of the brain to identify the neurodevelopmental and psychiatric significance of these genes (Figure 1d, Supplementary Table 7). The 661 suggestive peaks map to 106 genes. ~14% (15 genes) have been genetically associated with diseases of the nervous system, including schizophrenia (*SNAP91, CORO7*), bipolar disorder (*FBLN1*), multiple sclerosis (*THEMIS*, *CLECL16A*), alcohol dependence (*MREG, KCNJ6, FSTL5*), and Alzheimer's disease (*NRXN1*) (11 or 13% genes; Supplementary Table 7). Nearly one-third of these genes are markers of various cell-types in the fetal and newborn brain, including neuronal progenitor cells, neurons, radial glia, astrocytes, and endothelial cells (31 genes, 29%; [39]), and one gene is a marker of adult brain cells (*THEMIS*)[37]. Seven genes are known to interact with drugs; a notable interaction is between *CACNA2D3*, a voltage gated Calcium channel with suggestive association with working memory (top SNP p = 3.9e-6), and Gabapentin enacarbil, a drug used to treat epilepsy, neuralgia and restless legs syndrome[52]. One-sixth of suggestive peaks (112 SNPs or 17%) were predicted to have a functional consequence in brain tissue (Figure 1c, e), including nonsynonymous changes to protein sequence, presence in brain-specific promoters and enhancers, or association with changes in gene expression. In summary, genetic variants associated with typical variation in neurocognition map to genes implicated in human brain development, altered in psychiatric disease, and that are modulated by drugs used to treat neurological conditions.

Nonverbal reasoning was the only phenotype with SNPs passing the cutoff for genome-wide significance (rs77601382 and rs5765534, $p = 4.6 \times 10^{-8}$) (Figure 2). The peak is located in a ~33kb region (chr22:45,977,415-46,008,175) overlapping the 3' end of the Fibulin-1 (*FBLN1*) gene, including the last intron and exon (Figure 2b). To better understand the significance of this gene in brain function, we examined *FBLN1* expression in published fetal and adult transcriptomes, and single-cell data[10,39,41]. *FBLN1* transcription in the human brain is highest in the early stages of fetal brain development, with little to no expression in the adult (Figure 2c, Supplementary Figure 8); this is consistent with single-cell assays showing *FBLN1* to be a marker for dividing progenitor cells in the fetal brain (Figure 1d, [39]). *FBLN1* encodes a glycoprotein present in the extracellular matrix; this protein is a direct interactor of proteins involved in neuronal diseases, such as Amyloid Precursor Protein-1 (Supplementary Figure 9 [53]). *FBLN1* expression is upregulated in the brain in schizophrenia and has been previously associated with genetic risk for bipolar disorder (Figure 1d, [54,55]). Therefore, we conclude that *FBLN1*, associated with nonverbal reasoning test performance, shows characteristics

of a gene involved in neurodevelopment and the dysregulation of which could increase risk for psychotic disorders of neurodevelopmental origin.

We then performed pathway analysis for all nine selected CNB phenotypes using a rank-based pathway analysis strategy that includes all SNPs used in the association analysis (GSEA[28,34], 100 permutations; 4,102 pathways tested). SNPs were mapped to genes using brain-specific eQTL, chromatin interaction and positional information, using the same method as described above. Four out of nine phenotypes demonstrated significant enrichment of top-ranking genetic variants in pathways ($q < 0.1$; Figure 3a, Supplementary Tables 8-10). These included tasks in complex cognition (spatial reasoning), declarative memory (object and face memory), and executive function (working memory). The working memory phenotype showed significant enrichment of variants in pathways related to development, including neural development ($q<0.05$; Figure 3a, Supplementary Tables 8-10). To understand how genes contributing to pathway enrichment could be related to brain function, we annotated the corresponding leading edge genes with prior knowledge about associations with nervous system disorders, drug interactions and transcription in brain cell types[37-39,48,49]. Out of 355 leading edge genes, over half are known brain cell markers (228 genes or 64%), roughly one-third have known drug interactions (129 genes or 36%), and ~14% are associated with nervous system disease (51 genes) (pathway $q < 0.10$, Figure 3b, Supplementary Table 10). Among disease-associated genes were those associated with autism (*CSDE1*), multiple sclerosis (*CYP27B1*, *EOMES*), depression (*ROBO1*), glaucoma and wet macular degeneration (*LHCGR*). None of the SNPs associated with leading-edge genes (416 SNPs) overlapped suggestive or significant GWAS SNPs (661 SNPs).

To identify enrichment specific to brain-related processes and mental illness, we performed a second enrichment analysis using gene sets curated from the literature (Supplementary Note 5). These included gene sets derived from transcriptomic and proteomic profiles of the developing and adult healthy brain and brains affected by mental illness, genome-wide association studies and terms from phenotype ontology (421 gene sets tested, Supplementary Note 5, Supplementary Table 5, Supplementary Data 1). Six gene sets were significantly enriched ($q<0.10$), with five associated with working memory and the sixth with verbal reasoning (Figure 3c, Supplementary Table 11). A cluster of related gene sets related to autonomic nervous system dysfunction and a gene set related to locomotor dysfunction achieved significance at $q < 0.05$. Only one out of 157 SNPs associated with leading-edge genes overlaps with suggestive SNPs from GWAS analysis. Roughly 13% of the 134 leading edge genes are associated with nervous system disorders (18 genes), one-fifth have known drug targets (27 genes, 20%), and over half (81 genes or 60%) are markers of brain cell-types (Figure 3c,d; Supplementary Table 12, 13). Five genes have all three attributes: *SNCA*, *CAV1*, *LRRK2*, *ERBB4* and *MAPT* (Figure 3d, Supplementary Table 13). One example is Alpha-synuclein (*SNCA,* top SNP p=2.6e-4), which has been genetically associated with risk for developing Parkinson's disease[56], is a marker of excitatory neurons in the fetal brain[39], and is a drug target of BIIB504[48]. Another example is ERB-B2 receptor tyrosine kinase 4 (*ERBB4*), which has been genetically associated with mood disorders and unipolar depression[57], is a target of 24 drugs and is a marker of inhibitory neurons in the fetal brain. Other leading edge genes have been associated with schizophrenia, autism spectrum disorder, Parkinson's disease, Alzheimer's disease, depression and mood disorders (Figure 3d, Supplementary Table 13). In summary, genetic variants associated with normative variation in a range of neurocognitive phenotypes are enriched in pathways and gene sets related to cell proliferation, brain development, nervous system dysfunction and mental disorders.

Discussion

This study identifies molecular variants and cellular processes that contribute to normal human variation in specific cognitive domains. Consistent with heritability estimates, we find that the number of variant-level associations and enriched pathways varies considerably by phenotype (Figure 4). In particular, we find an enrichment of genetic variants associated with complex cognitive phenotypes (75-219 suggestive peaks), consistent with heritability estimates of up to 0.30-0.41 for these phenotypes[13]. A variety of cognitive phenotypes are enriched for variants in pathways. Moreover, the set of variants driving pathway enrichment has almost no overlap with suggestive variants from the GWAS analysis (no overlap for brain-related gene sets; a single SNP, rs9367669, overlaps for pathway sets). These results suggest that a molecules-to-behaviour research framework that includes genes and molecules, should also include pathways as a way to uncover new biological insights into existing genotype databases. Previous research in other polygenic psychiatric disorders, such as schizophrenia and major depression[58], has also shown an enrichment of disease-associated molecules in pathways. We suggest that the Research Domain Criteria (RDoC) matrix be updated to add a level for pathways, above that of genes and molecules and below cells. This modification will help associate additional genetic signal with brain related phenotypes, which otherwise would be missed if just considering SNPs and genes.

Variants, genes and pathways associated with typical variation in neurocognitive phenotypes, demonstrate evidence for a role in neurodevelopment, modulating gene expression in the fetal and adult brain and increasing risk for psychiatric disease (Figure 1, Supplementary Table 6, 7, 10, 13). Multiple lines of evidence suggest that *FBLN1*, the gene associated with genome-wide significant SNPs for nonverbal reasoning, is dysregulated in disease. In addition to the evidence provided in our results (Figure 1d, Figure 2c, Supplementary Figure 8,9), *FBLN1* has been associated with other rare genetic syndromes and protein levels of FBLN1 have been associated with altered risk for ischaemic stroke[59,60]. However, the mechanism by which *FBLN1* contributes to normal brain function is not known. We also do not exclude the possibility that suggestive peaks we identified within *FBLN1* may affect the function of neighbouring genes. One such gene is Ataxin-10 (*ATXN10*), in which a pentanucleotide repeat expansion causes spinocerebellar atrophy and ataxia[61].

An advantage of using a rank-based gene set enrichment analysis method, as compared to hypergeometric tests, is that the method ranks and prioritizes a subset of genes (leading edge genes) within a potentially large gene set (>100 genes), which are responsible for driving the enrichment statistic. In this work, we found five neurocognitive phenotypes with significant enrichment of high-ranking variants in pathways. We annotated leading edge genes to identify those that are jointly related to working memory, which demonstrated significant enrichment in both gene set analyses, and psychiatric disease (Figure 3). For instance, among the leading edge genes contributing to working memory were genes previously associated with Parkinson's disease, Alzheimer's disease, schizophrenia, autism, and depression, all of which have been associated with working memory impairments[62-67]. We note, however, that the individual genes connecting any given disease to working memory are different. For instance, among leading edge genes for working memory, *ERBB4* is associated with depression, whereas *SNCA* is associated with Parkinson's disease (Figure 3c, Supplementary Table 13). One implication of this partially overlapping gene network is that the therapeutic targets that may be relevant for working memory deficits may depend on what disease the patient has, as a different subset of the "working memory gene network" is affected by each condition.

This work contributes towards an understanding of the molecular underpinnings of human brain-related behaviour and could help to identify genetic contributors towards the heterogeneity in phenotypes associated with multiple brain-related disorders[68,69]. Our analysis is limited to univariate

genetic effects, and future work will explore the contribution of interactions between individual SNPs, possibly explaining lack of SNP-level or pathway-level signal in some of the phenotypes studied here[70]. Our findings also suggest that different cognitive phenotypes may be vulnerable to genetic alterations in different cellular pathways. Such exploration could identify disease-specific molecular targets that impinge on the same neurocognitive phenotype. Finally, we propose that research frameworks for linking genotype to phenotype for brain-related traits include cellular pathways as an organizational layer to support uncovering additional genetic signal from available genetic data.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank Shefali Verma for advice on genome imputation, and Sarah Gagliano for guidance on genetic analysis.

References

1 Insel, T. *et al.* Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. (2010).

2 Stein, M. B. & Smoller, J. W. Precision Psychiatry-Will Genomic Medicine Lead the Way? *JAMA psychiatry* **75**, 663-664, doi:10.1001/jamapsychiatry.2018.0375 (2018).

3 Germine, L. *et al.* Association between polygenic risk for schizophrenia, neurocognition and social cognition across development. *Translational psychiatry* **6**, e924, doi:10.1038/tp.2016.147 (2016).

4 Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet (London, England)* **381**, 1371-1379, doi:10.1016/s0140-6736(12)62129-1 (2013).

5 Thompson, P. M. *et al.* The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain imaging and behavior* **8**, 153-182, doi:10.1007/s11682-013-9269-5 (2014).

6 Calkins, M. E. *et al.* The Philadelphia Neurodevelopmental Cohort: constructing a deep phenotyping collaborative. *Journal of child psychology and psychiatry, and allied disciplines* **56**, 1356-1369, doi:10.1111/jcpp.12416 (2015).

7 Satterthwaite, T. D. *et al.* Neuroimaging of the Philadelphia neurodevelopmental cohort. *NeuroImage* **86**, 544-553, doi:10.1016/j.neuroimage.2013.07.064 (2014).

8 Jernigan, T. L. & Brown, S. A. Introduction. *Developmental cognitive neuroscience* **32**, 1-3, doi:10.1016/j.dcn.2018.02.002 (2018).

9 Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209, doi:10.1038/s41586-018-0579-z (2018).

10 Battle, A., Brown, C. D., Engelhardt, B. E. & Montgomery, S. B. Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213, doi:10.1038/nature24277 (2017).

11 Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

12 Akbarian, S. *et al.* The PsychENCODE project. *Nature neuroscience* **18**, 1707-1712, doi:10.1038/nn.4156 (2015).

13 Robinson, E. B. *et al.* The genetic architecture of pediatric cognitive abilities in the Philadelphia Neurodevelopmental Cohort. *Molecular psychiatry* **20**, 454-458, doi:10.1038/mp.2014.65 (2015).

14 Gur, R. C. *et al.* A cognitive neuroscience-based computerized battery for efficient measurement of individual differences: standardization and initial construct validation. *Journal of neuroscience methods* **187**, 254-262, doi:10.1016/j.jneumeth.2009.11.017 (2010).

15 Gur, R. C. *et al.* Age group and sex differences in performance on a computerized neurocognitive battery in children age 8-21. *Neuropsychology* **26**, 251-265, doi:10.1037/a0026712 (2012).

16 Moore, T. M., Reise, S. P., Gur, R. E., Hakonarson, H. & Gur, R. C. Psychometric properties of the Penn Computerized Neurocognitive Battery. *Neuropsychology* **29**, 235-246, doi:10.1037/neu0000093 (2015).

17 Verma, S. S. *et al.* Imputation and quality control steps for combining multiple genome-wide datasets. *Frontiers in Genetics* **5**, 370 (2014).

18 Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic acids research* **34**, D590-598, doi:10.1093/nar/gkj144 (2006).

19 Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods* **10**, 5-6, doi:10.1038/nmeth.2307 (2013).

20 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5**, e1000529, doi:10.1371/journal.pgen.1000529 (2009).

21 Anderson, C. A. *et al.* Data quality control in genetic case-control association studies. *Nature protocols* **5**, 1564-1573, doi:10.1038/nprot.2010.116 (2010).

22 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics* **88**, 76-82, doi:10.1016/j.ajhg.2010.11.011 (2011).

23 Schmidt, A. *et al.* Acute effects of heroin on negative emotional processing: relation of amygdala activity and stress-related responses. *Biological psychiatry* **76**, 289-296, doi:10.1016/j.biopsych.2013.10.019 (2014).

24 Krueger F. *Trim Galore!*, <http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/> (

25 Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4**, 1310, doi:10.12688/f1000research.7334.1 (2015).

26 Mifsud, B. *et al.* GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. *PloS one* **12**, e0174744, doi:10.1371/journal.pone.0174744 (2017).

27 Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell reports* **17**, 2042-2059, doi:10.1016/j.celrep.2016.10.061 (2016).

28 Wang, K., Li, M. & Bucan, M. Pathway-based approaches for analysis of genomewide association studies. *American journal of human genetics* **81**, 1278-1283, doi:10.1086/522374 (2007).

29 Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. *Nature reviews. Genetics* **11**, 843-854, doi:10.1038/nrg2884 (2010).

30 Romero, P. *et al.* Computational prediction of human metabolic pathways from the complete human genome. *Genome biology* **6**, R2, doi:10.1186/gb-2004-6-1-r2 (2005).

31 Kandasamy, K. *et al.* NetPath: a public resource of curated signal transduction pathways. *Genome biology* **11**, R3, doi:10.1186/gb-2010-11-1-r3 (2010).

32 Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nuc Acids Res* **44**, D481-487, doi:10.1093/nar/gkv1351 (2016).

33 Schaefer, C. F. *et al.* PID: the Pathway Interaction Database. *Nucleic acids research* **37**, D674-679, doi:10.1093/nar/gkn653 (2009).

34 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

35 Mi, H. *et al.* The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic acids research* **33**, D284-288, doi:10.1093/nar/gki078 (2005).

36 Consortium, T. G. O. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic acids research* **47**, D330-d338, doi:10.1093/nar/gky1055 (2019).

37 Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 7285-7290, doi:10.1073/pnas.1507125112 (2015).

38 Lake, B. B. *et al.* Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science (New York, N.Y.)* **352**, 1586-1590, doi:10.1126/science.aaf1204 (2016).

39 Nowakowski, T. J. *et al.* Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science (New York, N.Y.)* **358**, 1318-1323, doi:10.1126/science.aap8809 (2017).

40 Yu, N. Y. *et al.* Complementing tissue characterization by integrating transcriptome profiling from the Human Protein Atlas and from the FANTOM5 consortium. *Nucleic acids research* **43**, 6787-6798, doi:10.1093/nar/gkv608 (2015).

41 Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483-489, doi:10.1038/nature10523 (2011).

42 Velasquez, E. *et al.* Synaptosomal Proteome of the Orbitofrontal Cortex from Schizophrenia Patients Using Quantitative Label-Free and iTRAQ-Based Shotgun Proteomics. *Journal of proteome research* **16**, 4481-4494, doi:10.1021/acs.jproteome.7b00422 (2017).

43 Kohler, S. *et al.* Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic acids research* **47**, D1018-d1027, doi:10.1093/nar/gky1105 (2019).

44 Zhong, S. *et al.* A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524-528, doi:10.1038/nature25980 (2018).

45 Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PloS one* **5**, e13984, doi:10.1371/journal.pone.0013984 (2010).

46 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).

47 Franz, M. *et al.* GeneMANIA update 2018. *Nucleic acids research* **46**, W60-w64, doi:10.1093/nar/gky311 (2018).

48 Cotto, K. C. *et al.* DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. *Nucleic acids research* **46**, D1068-d1073, doi:10.1093/nar/gkx1143 (2018).

49 Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research* **47**, D1005-d1012, doi:10.1093/nar/gky1120 (2019).

50 Carvalho-Silva, D. *et al.* Open Targets Platform: new developments and updates two years on. *Nucleic acids research* **47**, D1056-d1065, doi:10.1093/nar/gky1133 (2019).

51 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

52 Yaltho, T. C. & Ondo, W. G. The use of gabapentin enacarbil in the treatment of restless legs syndrome. *Therapeutic advances in neurological disorders* **3**, 269-275, doi:10.1177/1756285610378059 (2010).

53 Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic acids research* **34**, D535-539, doi:10.1093/nar/gkj109 (2006).

54 Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science (New York, N.Y.)* **362**, doi:10.1126/science.aat8464 (2018).

55 Greenwood, T. A., Akiskal, H. S., Akiskal, K. K. & Kelsoe, J. R. Genome-wide association study of temperament in bipolar disorder reveals significant associations with three novel Loci. *Biological psychiatry* **72**, 303-310, doi:10.1016/j.biopsych.2012.01.018 (2012).

56 Simon-Sanchez, J. *et al.* Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nature genetics* **41**, 1308-1312, doi:10.1038/ng.487 (2009).

57 Nagel, M. *et al.* Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nature genetics* **50**, 920-927, doi:10.1038/s41588-018-0151-7 (2018).

58 Consortium, N. a. P. A. S. o. P. G. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature neuroscience* **18**, 199-209, doi:10.1038/nn.3922 (2015).

59 Palumbo, P. *et al.* Clinical and molecular characterization of an emerging chromosome 22q13.31 microdeletion syndrome. *American journal of medical genetics. Part A* **176**, 391-398, doi:10.1002/ajmg.a.38559 (2018).

60 Vadgama, N., Lamont, D., Hardy, J., Nasir, J. & Lovering, R. C. Distinct proteomic profiles in monozygotic twins discordant for ischaemic stroke. *Molecular and cellular biochemistry*, doi:10.1007/s11010-019-03501-2 (2019).

61 Matsuura, T. *et al.* Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nature genetics* **26**, 191-194, doi:10.1038/79911 (2000).

62 Forbes, N. F., Carrick, L. A., McIntosh, A. M. & Lawrie, S. M. Working memory in schizophrenia: a meta-analysis. *Psychological medicine* **39**, 889-905, doi:10.1017/s0033291708004558 (2009).

63 Huntley, J. D. & Howard, R. J. Working memory in early Alzheimer's disease: a neuropsychological review. *International journal of geriatric psychiatry* **25**, 121-132, doi:10.1002/gps.2314 (2010).

64 Kehagia, A. A., Barker, R. A. & Robbins, T. W. Neuropsychological and clinical heterogeneity of cognitive impairment and dementia in patients with Parkinson's disease. *The Lancet. Neurology* **9**, 1200-1213, doi:10.1016/s1474-4422(10)70212-x (2010).

65 Rose, E. J. & Ebmeier, K. P. Pattern of impaired working memory during major depression. *Journal of affective disorders* **90**, 149-161, doi:10.1016/j.jad.2005.11.003 (2006).

66 Stopford, C. L., Thompson, J. C., Neary, D., Richardson, A. M. & Snowden, J. S. Working memory, attention, and executive function in Alzheimer's disease and frontotemporal dementia. *Cortex; a journal devoted to the study of the nervous system and behavior* **48**, 429-446, doi:10.1016/j.cortex.2010.12.002 (2012).

67 Wang, Y. *et al.* A Meta-Analysis of Working Memory Impairments in Autism Spectrum Disorders. *Neuropsychology review* **27**, 46-61, doi:10.1007/s11065-016-9336-y (2017).

68 Clementz, B. A. *et al.* Identification of Distinct Psychosis Biotypes Using Brain-Based Biomarkers. *The American journal of psychiatry* **173**, 373-384, doi:10.1176/appi.ajp.2015.14091200 (2016).

69 Jeste, S. S. & Geschwind, D. H. Disentangling the heterogeneity of autism spectrum disorder through genetic findings. *Nature reviews. Neurology* **10**, 74-81, doi:10.1038/nrneurol.2013.278 (2014).

70 Wang, W. *et al.* Pathway-based discovery of genetic interactions in breast cancer. *PLoS genetics* **13**, e1006973, doi:10.1371/journal.pgen.1006973 (2017).

71 Smedley, D. *et al.* The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic acids research* **43**, W589-598, doi:10.1093/nar/gkv350 (2015).

72 Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nature communications* **8**, 1826, doi:10.1038/s41467-017-01261-5 (2017).

73 Robinson, J. T. *et al.* Integrative genomics viewer. *Nature biotechnology* **29**, 24-26, doi:10.1038/nbt.1754 (2011).

510 74 Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-
511 performance genomics data visualization and exploration. *Briefings in bioinformatics* **14**, 178-
512 192, doi:10.1093/bib/bbs017 (2013).
513 75 Gold, J. M. Cognitive deficits as treatment targets in schizophrenia. *Schizophrenia research* **72**,
514 21-28, doi:10.1016/j.schres.2004.09.008 (2004).
515
516

Figures

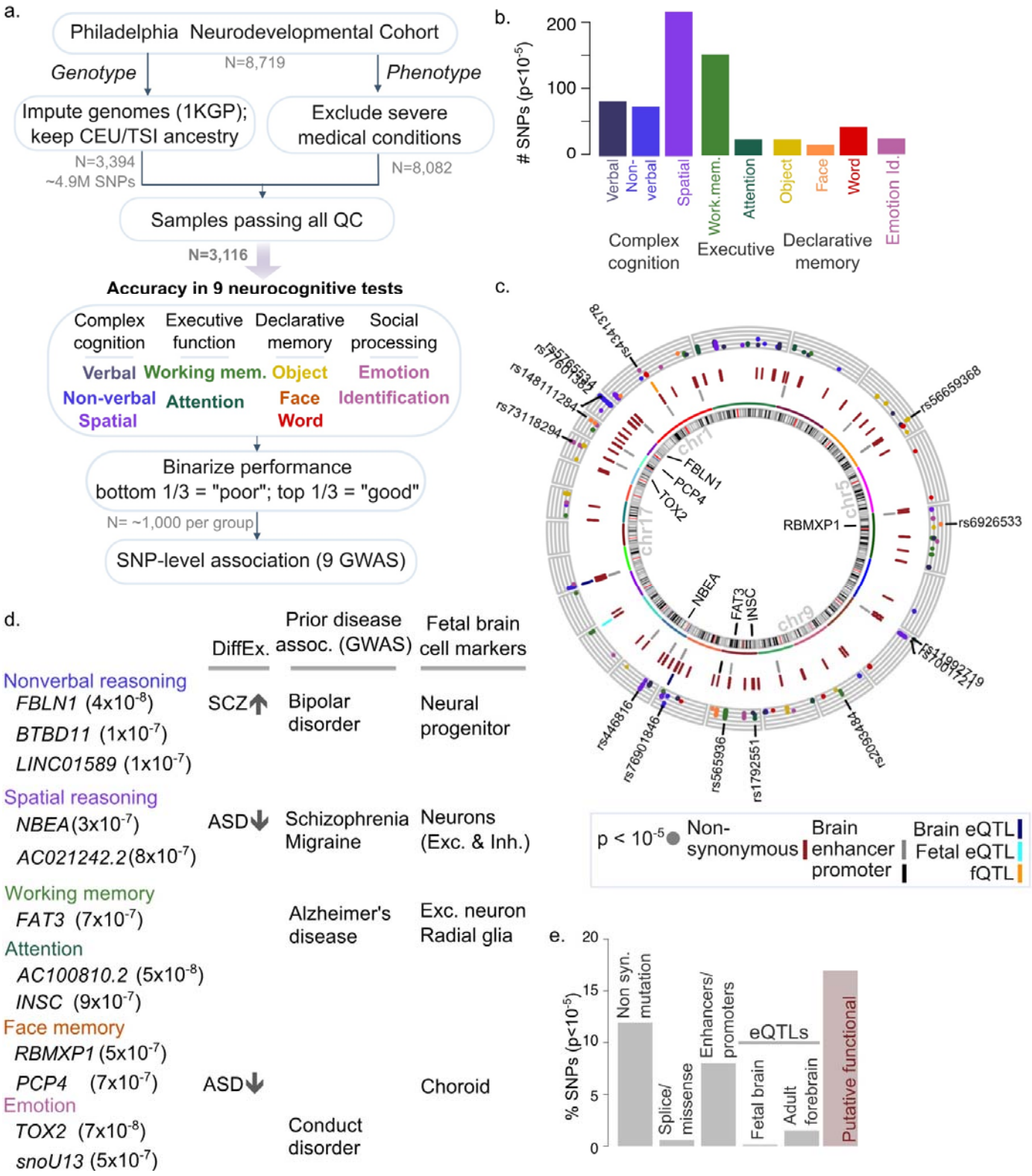


Figure 1. Genome-wide association analysis for neurocognitive phenotypes from the Philadelphia Neurodevelopmental Cohort.

a. Workflow. Genotypes were imputed (1KGP reference), and limited to European samples. Samples with severe medical conditions were removed and invalid test scores excluded. Nine neurocognitive test scores were binarized after age and sex had been regressed out. GWAS was performed for accuracy for each of these nine phenotypes.

b. Breakdown of SNPs achieving suggestive significance, by phenotype (top).

c. Suggestive and significant SNPs and associated genes. The outermost ring shows the location of suggestive peaks ($p < 10^{-5}$), coloured by phenotype (see b); y-axis shows –log10(SNP p), so that SNPs with stronger significance are higher. SNPs with $p<10^{-7}$ are labeled. The tracks with ticks indicate functional consequences of associated SNPs. The track closest to the middle indicate SNPs overlapping brain enhancers (light gray) or promoters (black). The dark red middle track indicates SNPs with nonsynonymous variation, including NMD transcript, missense or splice variants[71]. The outermost track indicates QTL associations, including eQTL in adult prefrontal cortex (dark blue), fetal brain (cyan), or neuronal cell proportions in the adult brain (fQTL; orange). Genes associated with top SNPs are indicated within the circle.

d. Genes associated with top SNPs ($p < 3x10^{-7}$) with prior knowledge about relevance to brain development and psychiatric disorders. Columns indicate differential expression in neurodevelopmental disorders[54] (SCZ = schizophrenia; ASD= autism), significant association with a nervous system disorder[49], or status as marker gene for specific cell types in fetal brain[39].

e. Breakdown of functional consequence of top SNPs and by functional consequence (bottom). Consequence shown is limited to effect on protein sequence[71], presence in enhancers or promoters in adult cortical regions[11], eQTL in fetal brain, or adult forebrain. Final bar shows cumulative proportion of putatively functional SNPs.

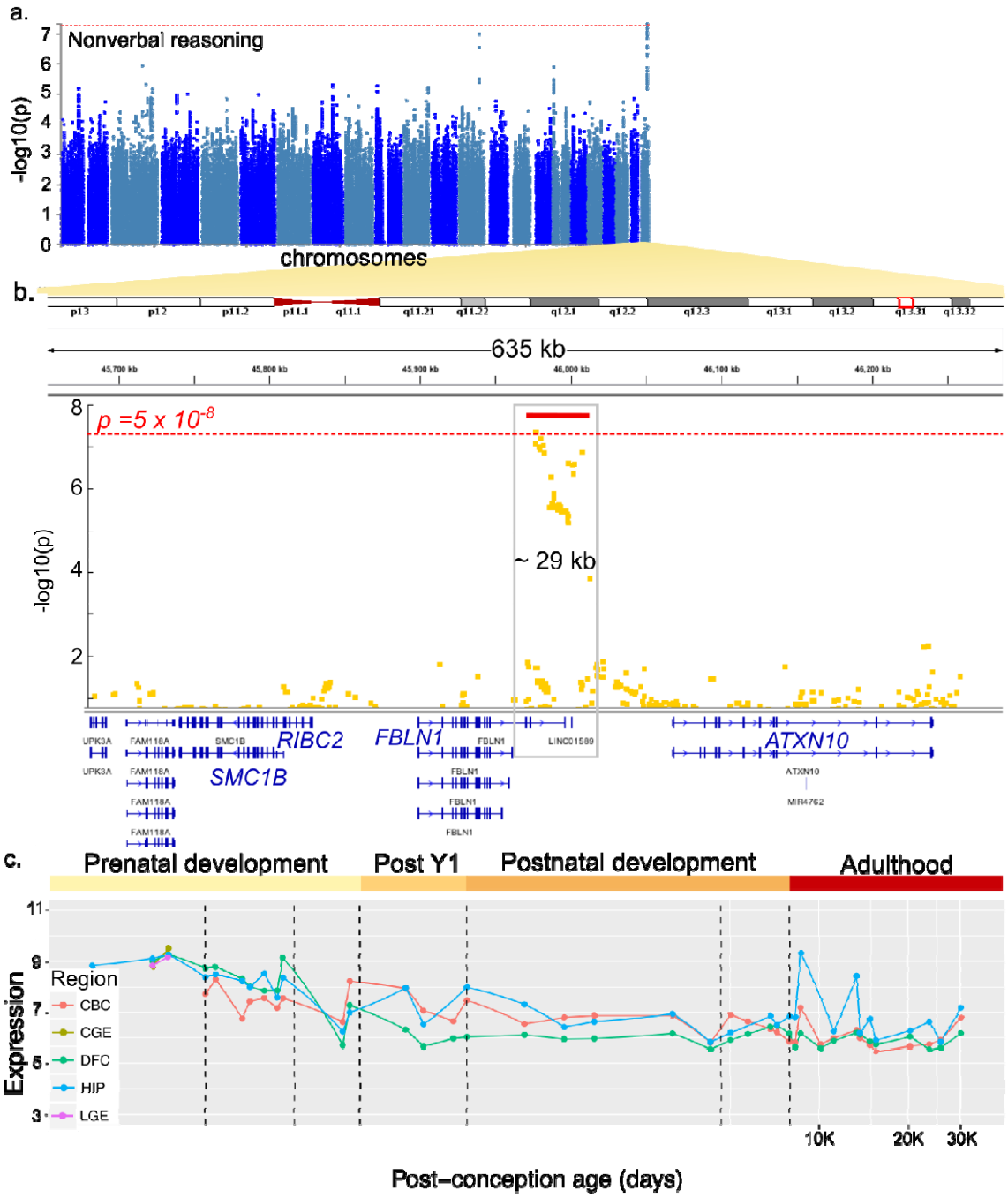


Figure 2. Genome-wide significance of *FBLN1* region for binarized performance in nonverbal reasoning

a. Manhattan plot of univariate SNP association with binarized performance in nonverbal reasoning (N=1,024 poor vs. 1,023 good performers; 4,893,197 SNPs). Plot generated using FUMA[72].

b. Detailed view of hit region at chr22q13. Two SNPs pass genome-wide significance threshold, rs77601382 and rs74825248 (p=4.64e-8). View using Integrated Genome Viewer (v2.3.93[73,74]). The red bar indicates the region with increased SNP-level association.

c. *FBLN1* transcription in the human brain through the lifespan. Data from BrainSpan[41]. Log-transformed normalized expression is shown for cerebellar cortex (CBC), central ganglionic eminence (CGE) and lateral ganglionic eminence (LGE), dorsal frontal cortex (DFC), and hippocampus (HIP).

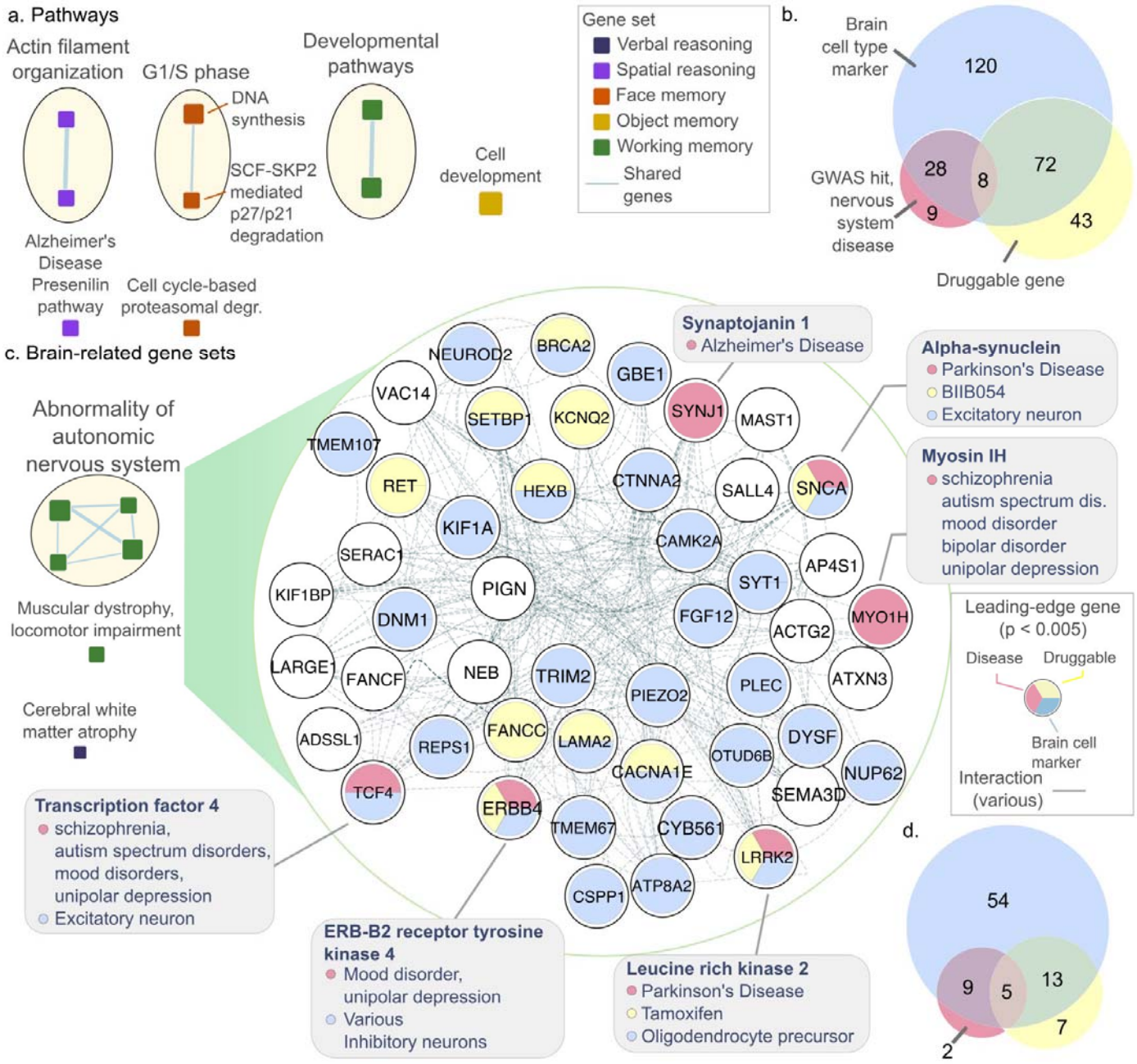


Figure 3. Pathway and gene set enrichment analysis for neurocognitive task performance

a. Pathways significantly enriched for genetic variation in neurocognitive task performance (GSEA, 100 permutations, q < 0.1, Supplementary Tables 8, 9, 10). Nodes indicate pathways, with fill indicating phenotype and yellow bubbles denoting clusters of related gene sets; edges indicate shared genes.

b. Number of leading edge genes associated with transcription in specific brain cell types (blue), drug targets (yellow) or genetic associations with specific nervous system disorders (pink) (pathways with q < 0.10, N=355 genes).

c. Brain-related gene sets enriched for genetic variation in task performance. Left: Significant gene sets; legend same as panel a (Supplementary Tables 11,12, 13). Right: Top leading edge genes in enriched brain-related gene sets (N=48 genes, p < 5e-3, pathways with q < 0.05). Nodes show genes and fill indicates genes associated with brain cell types, drugs or genetic associations with nervous system disorders (white indicates absence of association). Edges indicate known interactions (GeneMANIA[47]). Genes with disease associations have been highlighted in grey pullout bubbles.

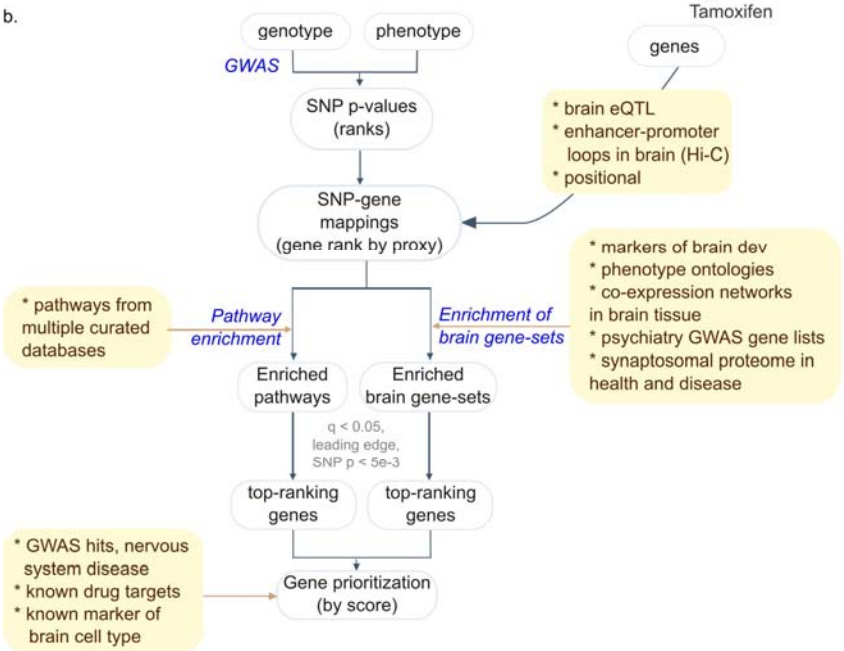
573 d. Leading edge genes in brain-related gene sets associated with disease, drugs or brain cell types
574 (N=134 genes); legend as in b.

a.

Increasing level of system organization

	Complex cognition (reasoning)				Executive function		Declarative memory		Social processing
Cognitive domain									
Task	Verbal	Nonverbal	Spatial	Attention	Working mem.	Object	Face	Word	Emotion Ident.
Variants #, $p < 1x10^{-5}$	75	75	219	24	153	24	16	43	24
Genes $p < 1x10^{-5}$ * $p < 5x10^{-8}$	*CEP162* *EEF1AKMT1* *MST1R* *NAV2* *NOTO* *NRXN1* *PGBD5* *PRKN* *PTPRQ* *RBM6* *XPO4*	*FBLN1* * *BTBD11* *C16orf96* *CDIP1* *CLSTN2* *CORO7* *FNBP1L* *MGRN1* *MREG* *NMRAL1* *PTPRD* *SLC26A3*	*ACADS* *BLOC1S5* *BLOC1S5-TXNDC5* *CDC45* *EEF1E1-BLOC1S5* *NBEA* *PRDM2* *SPPL3*	*INSC* *SRGAP3*	*CACNA2D3* *DNAH14* *FAT3* *KMT5A* *NCAM2* *SNAP91* *TECPR2* *THEMIS* *WISP3*	*BMP5* *CDH23* *FERMT1* *IGFBP7* *MBNL2* *PTPN13*	*ADAMTS14* *ETS1* *PCP4*	*FSTL5* *KCNC4* *TENM2*	*C6orf10* *CLEC16A* *EPG5* *KCNJ6* *NUP210L* *TOX2*
Pathway themes $q < 0.1$ * $q < 0.05$			Actin filament org.; Presenilin pathway		Developmental pathways *	Reg. of cell devel.	G1-S phase of cell cycle; NOD-like receptor sig.		
Brain-related gene set $q < 0.1$ * $q < 0.05$			White matter atrophy		Autonomic nervous system dysfunc. * Inability to walk *				
Cells (fetal brain), gene assoc.		Neural progenitors	Excitatory & inhibitory neurons	Excitatory neurons Radial glia					
Disease, gene assoc.		Schizoph. Bipolar disorder	Schizoph. Migraine	Alzheimer's	Parkinson's Autism Depression Alzheimer's Schizoph.				Conduct disorder
Disease-Task association				Attention[1] deficit; Psychosis risk	Schizoph.[1] Bipolar dis. TBI		Schizoph.[1,2] Epilepsy		Various neuropsych.[1] Depression Schizoph.
Drugs, gene assoc.					Schizoph. Bipolar dis. TBI AEE788 BIIB054 Chlorambucil Ergocalciferol Everolimus Imagabalin Nerispirdine Ocriplasmin Omeprazole Regorafenib Tamoxifen				

b.



575

Figure 4. a. Association of top genes, gene sets, and pathways with different levels of brain organization. Each column shows data for an individual phenotype, grouped by domain; rows show associations at increasingly higher levels (from top to bottom), and finally with drug targets. All results are from this work unless otherwise cited. Circles indicate relative number of suggestive variant peaks ($p < 10^{-5}$) from GWAS (median=43; mean=73.4), with numbers indicated below (asterisk: $p < 5x10^{-8}$), and genes are those mapped to top-ranking SNPs ($p <= 1x10^{-5}$) (only protein-coding genes; noncoding genes listed in Supplementary Table 14). Pathways and brain-related gene sets shown are those passing $q < 0.1$ in enrichment analysis (red asterisk: $q <= 0.05$). Fetal brain cell associations are as shown in Figure 1d. Gene-disease associations combine those for top GWAS SNPs (Figure 1d) and from gene set enrichment analysis; drug associations are from the latter (Supplementary Tables 10 and 14). Prior associations of alterations in phenotype or task-based brain activation as described in [14](1) or [75](2).

b. Proposed workflow for gene prioritization, as used in this work. When provided with genotype-phenotype data, SNPs are first prioritized by assigning an association statistic (e.g. by GWAS). Gene set enrichment analysis is performed to identify groups of genes with subthreshold phenotype association. SNP-gene mappings use brain-specific maps of genome regulation, prioritizing evidence-based association over positional mapping. Enrichment of pathways and brain-related gene sets are simultaneously performed using a rank-based method such as GSEA, which provides a leading edge subset for subsequent prioritization. Leading edge genes are annotated with clinical attributes of interest, such as druggability, prior disease association and evidence for expression in particular brain cell types, and the combination of attributes can be turned into a prioritization score.

Tables

Phenotype	N	# lead SNPs (p < 1e-5)	Indiv. Sig. SNPs (p < 1e-6)	SNP p	Gene
Complex Cognition					
Verbal	2,068	83	-		
Non-verbal	2,047	75	rs77601382	$4.6X10^{-8}$	*FBLN1*
			rs76901846	$1.0 X10^{-7}$	*BTBD11*
			rs5765534	$1.4 X10^{-7}$	
Spatial	2,024	219	rs446816	$2.6 X10^{-7}$	*NBEA*
			rs7001721	$8.5 X10^{-7}$	
Executive Function					
Working memory	2,047	153	rs565936	$6.6 X10^{-7}$	*FAT3*
			rs2093484	$9.3 X10^{-7}$	
Attention	2,041	24	rs11992719	$5.1 X10^{-8}$	
			rs1792551	$9.3 X10^{-7}$	*INSC*
Social processing					
Emotion Identification°	2,068	24	rs73118294	$7.1 X10^{-8}$	*TOX2*
			rs4341378	$4.9 X10^{-7}$	
Declarative memory					
Face memory	2,066	16	rs6926533; rs148111284	$5.4 X10^{-7}$; $6.9 X10^{-7}$	*RBMXP1* *PCP4*
Word memory	2,073	43	-		
Object memory	2,070	24	rs56659368	$3.2 X10^{-7}$	

Table 1. Genetic variants significantly associated with neurocognitive phenotypes in the Philadelphia Neurodevelopmental Cohort (PNC) dataset. For each test in the PNC neurocognitive test battery, GCTA was run to obtain SNP-level (marginal) p-values associated with binarized (good or poor) performance. Top SNPs ($p < 1.0x10^{-6}$) are shown above (full list of suggestive SNPs in Supplementary Table 5). SNPs were mapped to genes based on expression modulation, chromatin interaction of positional information. Only protein-coding genes shown here; additional non-coding RNA associations shown in Supplementary Table 7).