Comprehensive analyses of RNA-seq and genome-wide data point to enrichment of neuronal cell type subsets in neuropsychiatric disorders

Olislagers M[1*], Rademaker K[1], Adan RAH[1,2,3], Lin BD[1,4#] & Luykx JJ[1,4,5*#]

[1] Department of Translational Neuroscience, Brain Center Rudolf Magnus, University Medical Center Utrecht
[2] Institute of Neuroscience and Physiology, Sahlgrenska Academy
[3] Institute of Neuroscience and Physiology, The Sahlgrenska Academy at the University of Gothenburg
[4] Department of Psychiatry, Brain Center Rudolf Magnus, University Medical Center Utrecht
[5] Outpatient Second Opinion Clinic, GGNet Mental Health

*To whom correspondence should be addressed. E-mail: j.luykx@umcutrecht.nl (Jurjen J. Luykx), mitchellolislagers@gmail.com (Mitchell Olislagers).
[#] Shared supervision.

Keywords: Cell type enrichment; GWAS; single-cell RNA sequencing; LDSC; MAGMA; DEPICT; FUMA

Abstract

Neurologic, psychiatric and substance use disorders share a range of symptoms, which could be the result of shared genetic background. Many genetic loci have been identified for these disorders using genome-wide association studies, but conclusive evidence about cell types wherein these loci are active is lacking. We aimed to uncover implicated brain cell types in neuropsychiatric traits and to assess consistency in results across RNA datasets and methods. We comprehensively employed cell-type enrichment methods by integrating single-cell transcriptomic data from mouse brain regions with an unprecedented dataset of 42 human genome-wide association study results of neuropsychiatric, substance use and behavioral brain-related traits (n=12,544,007). Single-cell transcriptomic datasets from the Karolinska Institute and the 10x Genomics dataset were used. Cell type enrichment was determined using Linkage Disequilibrium Score Regression, Multi-marker Analysis of GenoMic Annotation, and Data-driven Expression Prioritized Integration for Complex Traits. The largest degree of consistency across methods was found for implication of pyramidal cells in schizophrenia and cognitive performance. For other phenotypes, such as bipolar disorder, two methods implicated the same cell types, i.e. medium spiny neurons and pyramidal cells. For autism spectrum disorders and anorexia nervosa, no consistency in implicated cell types was observed across methods. We found no evidence for astrocytes being consistently implicated in neuropsychiatric traits. We provide comprehensive evidence for a subset of neuronal cell types being consistently implicated in several, but not all psychiatric disorders, while non-neuronal cell types seem less implicated.

Introduction

It is well documented that several neuropsychiatric disorders, including substance use disorders (SUDs), share symptoms, which could be the result of shared genetic underpinnings[1,2]. Much of the heritability (h^2) of genetically complex or polygenic brain disorders -e.g. schizophrenia (SCZ), Parkinson's disease and alcohol use disorder- is due to common genetic variation[3]. In addition, genome-wide association studies (GWASs) have deepened the understanding of such disorders, unravelling thousands of associated loci[4,5]. However, elucidating disease mechanisms has remained challenging. One reason is missing heritability, meaning the gap between twin-based and SNP-based h2 estimates, which may result from limited statistical power, GWASs not probing associations with rare variants, epigenetics, genomic interactions, and structural genomic alterations[6]. Another reason is that over 90% of identified variants are located within non-coding regions of the genome, indicating that regulatory elements -e.g. promoters, enhancers and insulators- may explain part of the underlying genetic mechanisms in some polygenic disorders[4,7]. Due to extensive linkage disequilibrium (LD), it is also challenging to identify a causal variant within a given associated locus[4].

To overcome gaps between associated and causal genetic association, their functional effects and ultimately the biological pathways, extensive research has been performed to identify brain tissues having a role in neuropsychiatric disease. Functional genomic studies using macroscopic brain samples point to enrichment in phylogenetically conserved areas of the brain in psychiatric disorders and brain-related behavioral phenotypes, whereas enrichment in neurological disorders is typically found for fewer brain regions[3]. However, identification of specific cell types within brain tissues is considerably less well studied. Specific cell types that are associated with SCZ and anorexia nervosa (AN) have previously been identified by integrating GWAS findings with mouse single-cell RNA (scRNA) brain data: while medium spiny neurons (MSNs), cortical interneurons, hippocampal CA1 pyramidal cells (pyramidal CA1), and pyramidal cells from the somatosensory cortex (pyramidal SS) seem implicated in SCZ[8], suggestive findings were reported for enrichment of MSNs and pyramidal cells (CA1) in AN[9]. Recently, more extensive cell type enrichment analysis was performed for 28 phenotypes using mouse gene expression from the entire central nervous system (CNS)[10]. In psychiatric disorders, enrichment was found for MSNs, cortical interneurons, striatal interneurons, neuroblasts, pyramidal cells (CA1), and pyramidal cells (SS)[10]. In neurological disorders, fewer cell types were identified and these were dissimilar across disorders[10]. These cell type enrichment analyses have mainly been performed using Linkage Disequilibrium Score Regression (LDSC) and/or Multi-marker Analysis of GenoMic Annotation (MAGMA). However, in this landmark and other studies, MAGMA versions <1.08 have been employed[8-10], of which it was recently reported that its SNP-level P-value aggregation into gene-level P-values might result in type-I errors[11]. In addition, cell type enrichment in SUDs and several other disorders, such as anxiety disorders, has to the best of our knowledge not been studied.

Here, we systemically investigated cell type enrichment in an extensive set of brain-related phenotypes by integrating mouse scRNA brain data from the Karolinska Institute (KI) and 10x Genomics with summary statistics from 42 phenotypes related to neuropsychiatry, SUDs, and brain-related behavior. Our goals were to perform cell type enrichment for a more comprehensive set of brain-related traits than previously studied and to assess consistency in results across a wider array of methods. We went beyond previous studies by systematically performing cell type enrichment analyses using the most recent releases of different methods that rely on different assumptions and

algorithms, i.e. LDSC, MAGMA v1.08, Data-driven Expression Prioritized Integration for Complex Traits (DEPICT) and Functional Mapping and Annotation (FUMA). We found evidence for a subset of neuronal cell types being consistently implicated in several, but not all psychiatric disorders, while non-neuronal cell types seem less implicated.

Methods

GWAS summary statistics

Our goals were to identify salient cell types that are implicated in more brain-related traits than previously studied and to assess consistency in results across methods. Brain-related GWAS summary statistics from predominantly European samples were obtained from publicly available sources. A total of 41 summary statistics from brain-related phenotypes (Table 1, Table S1) were obtained, among which 11 psychiatric disorders (486,142 cases and 1,002,695 controls), 11 neurological disorders (186,171 cases and 2,278,970 controls), and 8 substance use disorders (case/control: 11,569 cases and 34,999; cohorts with continuous substance use data: n=3,683,037). All psychiatric and neurologic disorders for which summary statistics were available had also been included in the Brainstorm project[3]. Substance use disorders were added because of the high comorbidity[12,13] and genetic covariance[14] with psychiatric traits. Eleven well-powered (N>250,000) brain-related behavioral/quantitative phenotypes (n=4,166,895) were additionally selected. Because of the association between BMI and brain structure, we considered BMI a brain-related trait[15]. Finally, to discriminate cell types that were specific to the brain, height (n=693,529) was included as a non-brain-related anthropomorphic trait.

Table 1. Phenotype descriptions

	Phenotype	Cases	Controls	Total number of participants	Source	Ancestry
Psychiatric disorders	Attention-deficit/hyperactivity disorder (ADHD)	19,099	34,194	53,293	PGC	European
	Anorexia nervosa (AN)	16,992	55,525	72,517	PGC	European
	Anxiety disorders	7,016	14,745	21,761	ANGST	European
	Autism spectrum disorder (ASD)	18,382	27,969	46,351	PGC	European
	Bipolar disorder (BIP)	20,352	31,358	51,710	PGC	European
	Cross disorders	162,151**	276,846**	438,997**	PGC	European
	Major depressive disorder (MDD)	170,756	329,443	500,199	PGC	European
	Obsessive-compulsive disorder (OCD)	2,688	7,037	9,725	PGC	European
	Post-traumatic stress disorder	23,212	151,447	174,659	PGC	European
	Schizophrenia (SCZ)	40,675	64,643	105,318	PGC	European
	Tourette syndrome	4,819	9,488	14,307	PGC	European
Substance use disorders	Alcohol use	N/A	N/A	121,604	PGC	European
	Alcohol dependence	11,569	34,999	46,568	PGC	European
	Drinks per week	N/A	N/A	941,280	GSCAN	European
	Cannabis use	N/A	N/A	162,082	PGC	European

	Age smoking initiation	N/A	N/A	341,427	GSCAN	European
	Ever smoked regularly	N/A	N/A	1,232,091	GSCAN	European
	Cigarettes per day	N/A	N/A	337,334	GSCAN	European
	Smoking cessation	N/A	N/A	547,219	GSCAN	European
Neurological disorders	Amyotrophic lateral sclerosis	20,806	59,804	80,610	AVS	European
	Alzheimer's disease	71,880	383,378	455,258	PGC	European
	Epilepsy	15,212	29,677	44,889	ILAE	Multi-ancestry
	Generalized epilepsy	3,769*	29,677*	33,446*	ILAE	Multi-ancestry
	Focal epilepsy	9,671*	29,677*	39,348*	ILAE	Multi-ancestry
	Stroke	40,585	406,111	446,696	MEGASTROKE	European
	Ischemic stroke	34,217*	406,111*	440,328*	MEGASTROKE	European
	Large artery stroke	4,373*	406,111*	410,484*	MEGASTROKE	European
	Cardioembolic stroke	7,193*	406,111*	413,304*	MEGASTROKE	European
	Small vessel stroke	5,386*	406,111*	411,497*	MEGASTROKE	European
	Parkinson's disease	37,688	1,400,000	1,437,688	IPDGC-PDWBS-SGPD	European
Behavioral/quantitative	Body mass index (BMI)	N/A	N/A	681,275	GIANT	European
	Chronotype	N/A	N/A	449,734	SDKP	European
	Excessive daytime sleepiness	N/A	N/A	452,071	SDKP	European
	Sleep duration	N/A	N/A	446,118	SDKP	European
	Short sleep duration	106,192*	305,742*	411,934*	SDKP	European
	Long sleep duration	34,184*	305,742*	339,926*	SDKP	European
	Insomnia	N/A	N/A	453,379	SDKP	European
	Intelligence	N/A	N/A	269,867	CTGLAB	European
	Educational attainment	N/A	N/A	766,345	SSGAC	European
	Cognitive performance	N/A	N/A	257,828	SSGAC	European
	Neuroticism	N/A	N/A	390,278	CTGLAB	European
Non-brain-related control	Height	N/A	N/A	693,529	GIANT	European
Total		**683,882**	**3,316,664**	**12,544,007**		

Detailed descriptions, including references, are listed in Table S1.

**Sample count of a phenotype that is part of larger group.*

***May include sample overlap with AN, ADHD, ASD, BIP, MDD, OCD, SCZ and Tourette syndrome.*

Abbreviations: PGC, Psychiatric Genomics Consortium; ANGST, Anxiety Neuro Genetics STudy, GSCAN; GWAS and Sequencing Consortium of Alcohol and Nicotine use, AVS; ALS Variant Server, ILAE; International League Against Epilepsy, IPDGC; International Parkinson's Disease Genomics Consortium, SGPD; Systems genomics of Parkinson's disease

consortium, PDWBS; Parkinson's disease web based study, SDKP; Sleep Disorder Knowledge Portal, CTGLAB; Complex Traits Genetics Lab, SSGAC; Social Science Genetic Association Consortium.

Single cell RNA sequencing datasets

All cell type enrichment analyses were conducted using the KI dataset[8,16-18] and the 10x Genomics dataset[19]. These datasets were selected because they cover brain regions that are generally accepted to be involved in the pathogenesis of brain-related disorders[20]. Additionally, their high coverage may enable the identification of different cell types. Detailed information about the 10x Genomics dataset, regarding quality control, necessity of a randomized representative subset of cells and cell type identification are reported in the Supplementary Methods.

Overview of cell type enrichment analyses

To identify cell types underlying various phenotypes, we employed four methods (Figure 1). LDSC was first used to estimate SNP-h^2 and bivariate genetic correlations across all traits. Then, LDSC[21] was used to test whether the 10% most cell type-specific genes, based on the specificity metric Sg,c described above, were enriched in h^2. MAGMA (version 1.08)[8,22] was used to identify whether gene-level association of summary statistics either linearly increased with cell type expression specificity or whether the top 10% specific gene-level association of the summary statistics were associated with cell type expression specificity. DEPICT was used to identify genes from associated GWAS loci that were significantly enriched in certain cell types. By comparing cell type enrichment results of LDSC, DEPICT and MAGMA, we evaluated the relative stringency of each method. Finally, external scRNA datasets were used to conduct additional cell type enrichment analyses, using FUMA. To allow comparison of all enrichment methods, we compared the P-values that refer to the strength of association of a given cell type with a given phenotype, as not all methods provide an enrichment score. KI level 1, KI level 2 and 10x Genomics cell types were identified as significant after passing a Bonferroni corrected significance level of $P<0.05/(24*42)$, $P<0.05/(149*42)$ and $P<0.05/(16*42)$, respectively. We then counted the number of methods pointing to significant enrichment of specific cell types and report that number for both KI levels and 10x Genomics as our main outcome measure. Phenotypes implicating similar cell types were then identified by hierarchal clustering. We discuss these methods more elaborately below.

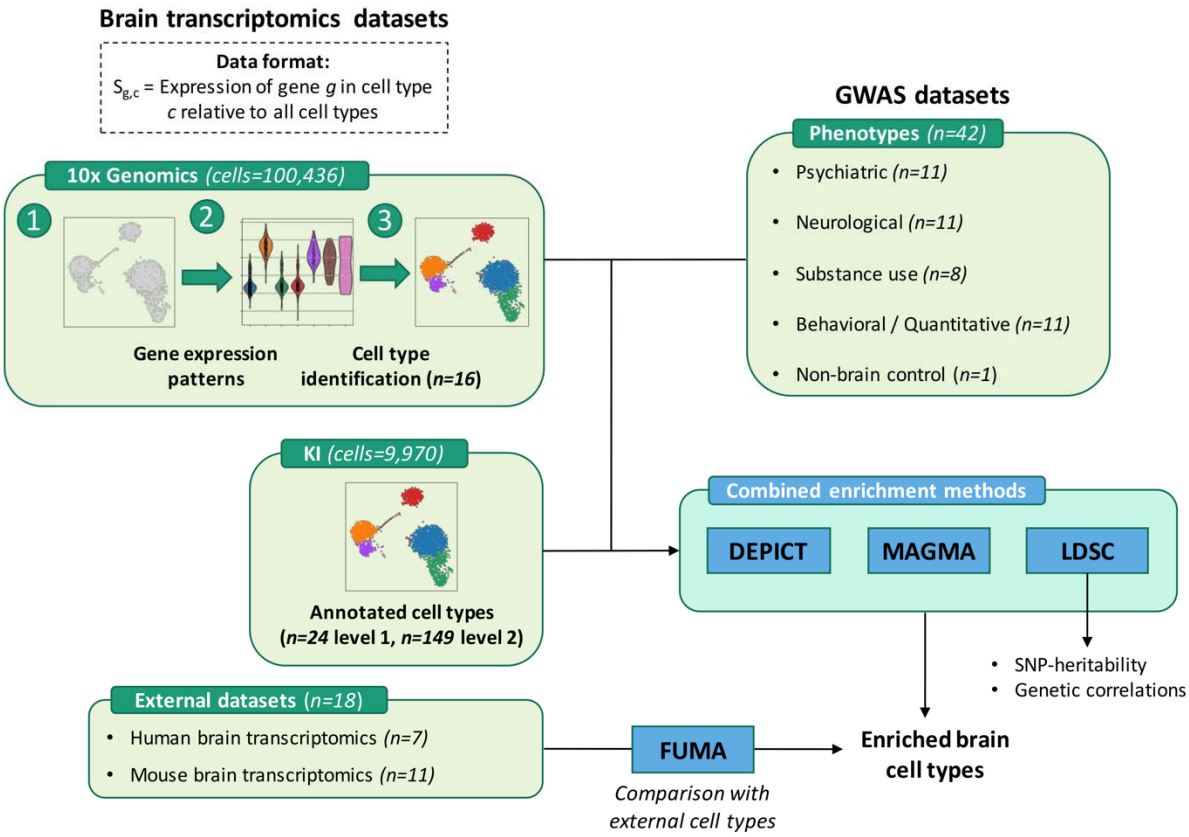

Figure 1. Overview of the approach of dataset integration as inputs for enrichment methods, in order to detect implicated brain cell types for various phenotypes.

Two mouse brain transcriptomic datasets (10x Genomics, KI) have the data format Sg,c of cell-type specificity for genes, which was calculated by dividing expression of gene G in cell type C by expression of G in all cell types of a given dataset. Custom cell type identification was performed for 10x Genomics (16 detected cell types), while existing annotation was re-used for KI (first level of 24 cell types and second level of 149 cell (sub-)types). The datasets were integrated with genome-wide association study (GWAS) data, and these were the input for cell type enrichment methods DEPICT, MAGMA and LDSC. External human and mouse brain transcriptomics data were used in cell type enrichment method FUMA, so that enriched cell types from any of the other three methods could be compared to FUMA-enriched cell types. Finally, LDSC was also used to estimate SNP-based heritability for each GWAS phenotype and to calculate genetic correlations across all phenotypes.

Cell type enrichment using LDSC

Human orthologs were obtained using the One2One R package that is incorporated in the MAGMA_Celltyping R package[8]. SNPs were annotated to the human genome (hg19, version 33) of the GENCODE project[23]. Binary annotations files were created for each chromosome, containing 11 sub-annotations. In the first sub-annotation, SNPs that mapped to genes without a human ortholog were coded as 1. The other 10 sub-annotations represented the SNPs in specificity deciles for a particular cell type in increasing order (1 = SNP belongs to a sub-annotation). These specificity deciles were obtained by restructuring the specificity metric $S_{g,c}$, described in the Methods ("Single cell RNA sequencing dataset quality control and preparation") using the 'prepare.quantile.groups' function in the MAGMA_Celltyping package[8]. LD scores were then calculated for each annotation file using a 1 centimorgan (cM) window, 1000 Genomes Project Phase 3 files[24] and restricted to

1,217,311 Hapmap3 SNPs. For each summary statistics dataset, we generated munged summary statistics by applying previously described quality control steps[25] (Supplementary Methods), implemented in the LDSC 'munge_sumstats.py' script. Finally, SNP-h^2 was partitioned, using the munged summary statistics, 1000 Genomes Project Phase 3 MAF files and both the 1000 Genomes Project phase 3 baseline model and all sub-annotations as independent variables. For the regression weights, we used the LD weights calculated for HapMap3 SNPs, excluding the MHC region (chr6: 25-34 Mb) using the 'overlap-annot' to account for SNPs grouped into multiple deciles. In addition to the settings described above, we performed sensitivity analyses, including removing the HapMap3 SNPs restriction, using only SNPs that pass a genome-wide significance threshold, changing the software version and changing the reference genome version to determine differences in KI-derived cell type enrichment results in SCZ. To allow comparison of all enrichment methods, cell type enrichment figures show the P-value associated with the most specific decile for each cell type as not all methods provide an enrichment score.

Cell type enrichment using MAGMA

MAGMA[8] version 1.08 was used to identify whether gene-level association of summary statistics linearly increased with cell type expression specificity or whether the top 10% specific gene-level association of the summary statistics were associated with cell type expression specificity. The analysis was restricted to Hapmap3 SNPs and the MHC region was excluded. The KI and 10x Genomics specificity metric $S_{g,c}$ were transformed into 41 bins using the 'prepare.quantile.groups' from the MAGMA_Celltyping R package. Mouse genes were then mapped to human orthologs. A 10kb upstream and 1.5kb downstream window was used to include SNPs surrounding GWAS hits to compute single gene-level P-values. The association tests were performed with the 'calculate_celltype_associations' function in linear and top 10% mode. The top 10% mode is more specific as it only takes into account the top 10% most specific gene-level P-values, whereas the linear mode provides more power by taking into account gene-level P-values in all binned fractions in a linear regression model. As a sensitivity analysis, MAGMA was also run unrestricted to HapMap3 SNPs and including the MHC region to determine their effects on cell type enrichment results for SCZ using both the KI and 10x Genomics dataset. The analyses were additionally performed in MAGMA (version 1.07b) to detect the implications of the recently corrected statistical foundation in the 'snp-wise mean model' of version 1.07b that is used to aggregate SNP-level P-values into a gene-level test statistic[11].

Cell type enrichment using DEPICT

DEPICT[26] (version 1, release 194) was used to identify cell types wherein genes from associated loci were significantly enriched using the specificity metric $S_{g,c}$. Only SNPs that passed a significance threshold ($P<1x10^{-5}$) were included in the analysis. As DEPICT does not allow more than 1,000 associated SNPs to be included in the analysis, a stricter significance threshold was applied for body mass index (BMI) ($P<5x10^{-8}$), educational attainment ($P<5x10^{-8}$) and height ($P<1x10^{-15}$) to reduce the number of associated SNPs that pass the significance threshold. Default parameters were used for the remaining settings. To evaluate the effect of restricting to HapMap3 SNPs on KI-derived cell type enrichment, DEPICT was run both restricted and unrestricted to HapMap3 SNPs for SCZ as a sensitivity analysis. Additionally, the MHC region was excluded.

Additional cell type enrichment analyses using additional mouse and human scRNA datasets

To confirm our cell type enrichment findings and to investigate whether these findings could be replicated using human data, additional cell type specificity analyses were performed using FUMA[27,28] (version 1.3.6a), which applies MAGMA (version 1.08) to test for positive relationships between cell type expression specificity of various external mouse (n=11) and human (n=7) scRNA datasets (Table S2) and associated SNPs that are aggregated to a gene-level P-value. The scRNA datasets originate from various regions of the brain, such as cortices (cerebral, frontal, somatosensory), basal ganglia (striatum, substantia nigra, globus pallidus), hippocampus, cerebellum and hindbrain. SNPs that passed the genome-wide significance threshold ($P<5x10^{-8}$) were included in the analyses. Because no SNPs were independently associated with obsessive-compulsive disorder (OCD) and small vessel stroke, a lenient significance threshold ($P<1x10^{-5}$) was applied for OCD and small vessel stroke. Conditional pair-wise analyses were applied both per dataset and across datasets to identify independently associated cell types.

Results

Cell type-specific gene expression in the 10x Genomics dataset

In the quality control of a randomized representative subset (n=108,844) of the 10x Genomics dataset, 8,408 cells and 6,419 non-expressed genes were removed from further analyses (Figure S1-S3). Altogether, the subset consisted of a matrix with 21,579 genes and 100,436 cells with 16 cell clusters (Table S3-S4). All cell clusters were subsequently mapped to biological brain cell types by specifically expressed marker genes (Figure S4).

LDSC, MAGMA and DEPICT sensitivity analyses and quality control

For cell type enrichment analyses using LDSC, we initially adopted the same parameters that were previously described[8]. Additionally, to optimize the cell type enrichment pipeline we tested various settings (Figure S5, Table S5). The parameters described in the methods section 'LDSC to target cell types' provided cell type enrichment results that were most consistent with DEPICT and MAGMA. For MAGMA, we found that restricting to HapMap3 SNPs and excluding the MHC region increased statistical power to identify associated cell types, whilst not inflating cell types that were not associated (Figure S6, Table S6). In addition to MAGMA version 1.08, we also performed cell type enrichment analyses using MAGMA version 1.07b (Table S7). We found that, although cell type associations follow similar patterns using both versions, the updated SNP-wise mean gene analysis model exerts effects on cell type enrichment results, resulting in differently associated cell types (Figure S7). However, no consistent unidirectional differences in cell type enrichment results were observed. Finally, for DEPICT, we found that not restricting to HapMap3 SNPS increased statistical power to identify associated cell types, without an upwards bias for non-associated cell types (Figure S8, Table S8).

Cell type enrichment analyses using the 10x Genomics dataset

Consistent with SNP-h^2 estimate patterns (Table S9, Figure S9-10, Supplementary Methods, Supplementary Results) and genetic correlations (Table S10, Figure S11-12, Supplementary Methods, Supplementary Results), we found that cell type association patterns of neurological disorders were distinct from psychiatric, substance use and behavioral association patterns by hierarchal clustering (Figure S13-15). For brain-related phenotypes, no cell types in the 10x Genomics dataset were identified to which implicated genomic loci consistently mapped using all three methods (Figure S13-S15, Table S11). By comparing LDSC, MAGMA and DEPICT, we found that MAGMA (linear) was too lenient (Figure S16) and susceptible to bias due to sample size (Figure S17). Therefore, MAGMA (linear) was excluded from our analyses.

Two methods provided evidence of neurons as group to be implicated in cross-disorders (8 psychiatric disorders jointly studied)[5]. Neurons as a group were, according to two methods, also associated with educational attainment, along with certain interneurons. Additionally, neurons as group were associated with intelligence. We also found evidence by two methods that implicated genomic loci of cognitive performance specifically mapped to certain neuroblasts. Suggestive findings are reported in the Supplementary Methods.

Cell type enrichment analyses using the KI dataset

After integrating GWAS findings with the 10x Genomics dataset, the analyses were expanded by using the KI dataset, which includes more cell types and thereby improves the resolution of the analysis (Figure 2 & 3, Figure S18-S19, Table S12). The largest degree of consistency across methods in brain-related traits was found for SCZ and cognitive performance (Figure 2 & 3). Genetic loci that are associated with SCZ consistently mapped to excitatory pyramidal cells (CA1) and pyramidal cells (SS), while those associated with cognitive performance only mapped to pyramidal cells (SS). For SCZ, we found evidence by two methods that MSN were the main implicated inhibitory neurons. MSNs and both types of pyramidal cells were found to be associated with cross-disorders by two methods, while only pyramidal cells were associated with educational attainment and MSNs and pyramidal cells (CA1) were implicated in BIP. Suggestive findings are reported in the Supplementary Methods.

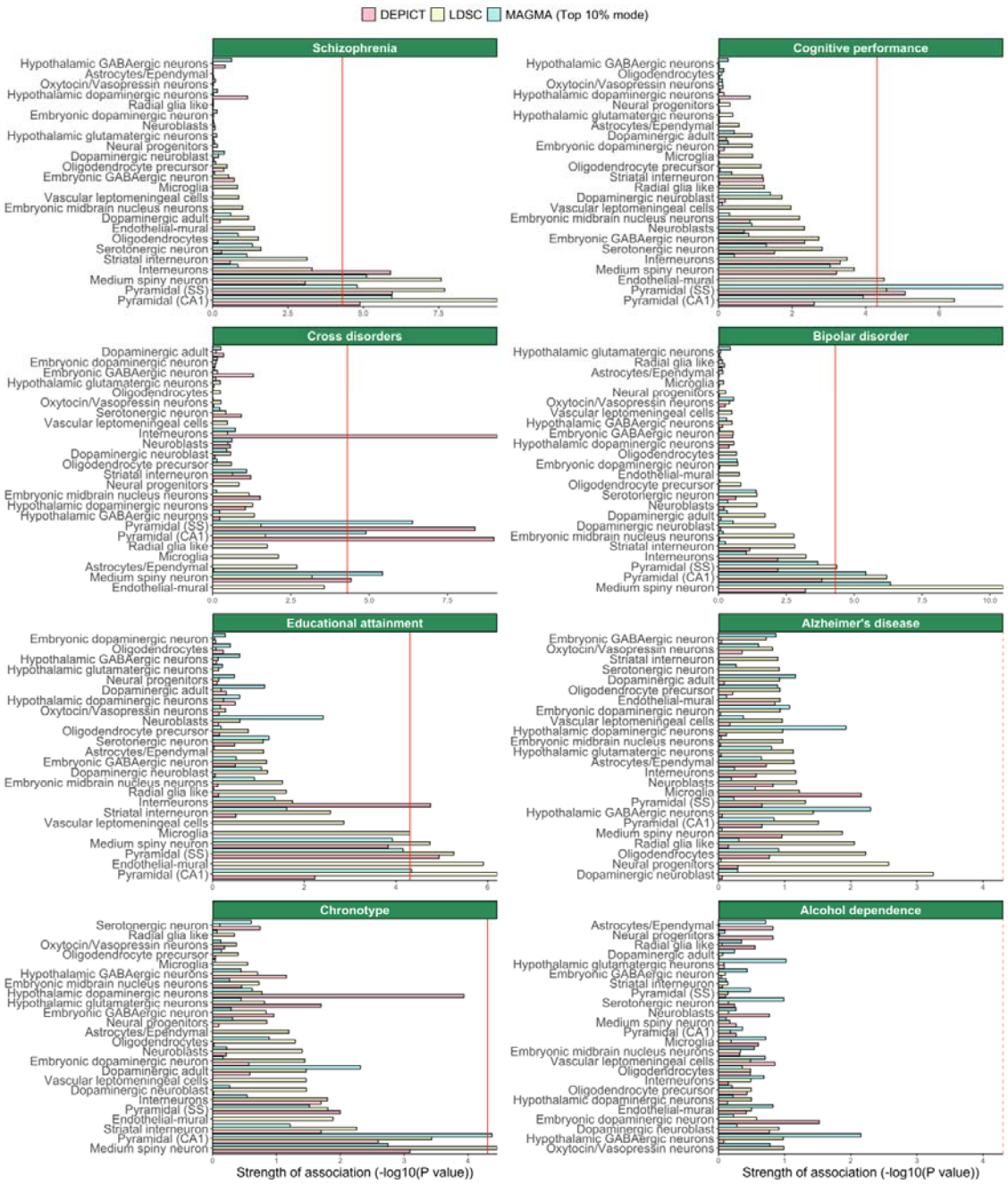

Figure 2. Cell type enrichment estimated by DEPICT, LDSC and MAGMA (top 10% mode) in selected brain-related phenotypes.

*Cell type enrichment results are generated using KI data. Bars represent the mean strength of association (-log10(P)) of LDSC, DEPICT and MAGMA (top 10%). The red line indicates the Bonferroni threshold P<0.05/(24*42). The red line is solid if any of the methods identified any cell type as significantly associated, and if none of the methods identified any of the cell types as significantly associated, the red line is dashed. A complete overview of cell type enrichment results using KI data, including MAGMA (linear) is available in the supplementary information (Figure S18).*

To identify cell types on a deeper cellular level, the analysis was further expanded by using the KI level 2 dataset (Figure S20-S22, Table S13), which includes 149 cell types that were subtypes of the cell types identified in the level 1 dataset. We were able to identify subgroups of the KI level 1 cell types, allowing the identification of specific gene-expressing cell types (Figure S20-S22, Table S13). Using both the KI level 1 and level 2 datasets, we again found that MAGMA (linear) was relatively lenient (Figure S23-S24) and prone to inflated results due to sample size (Figure S25).

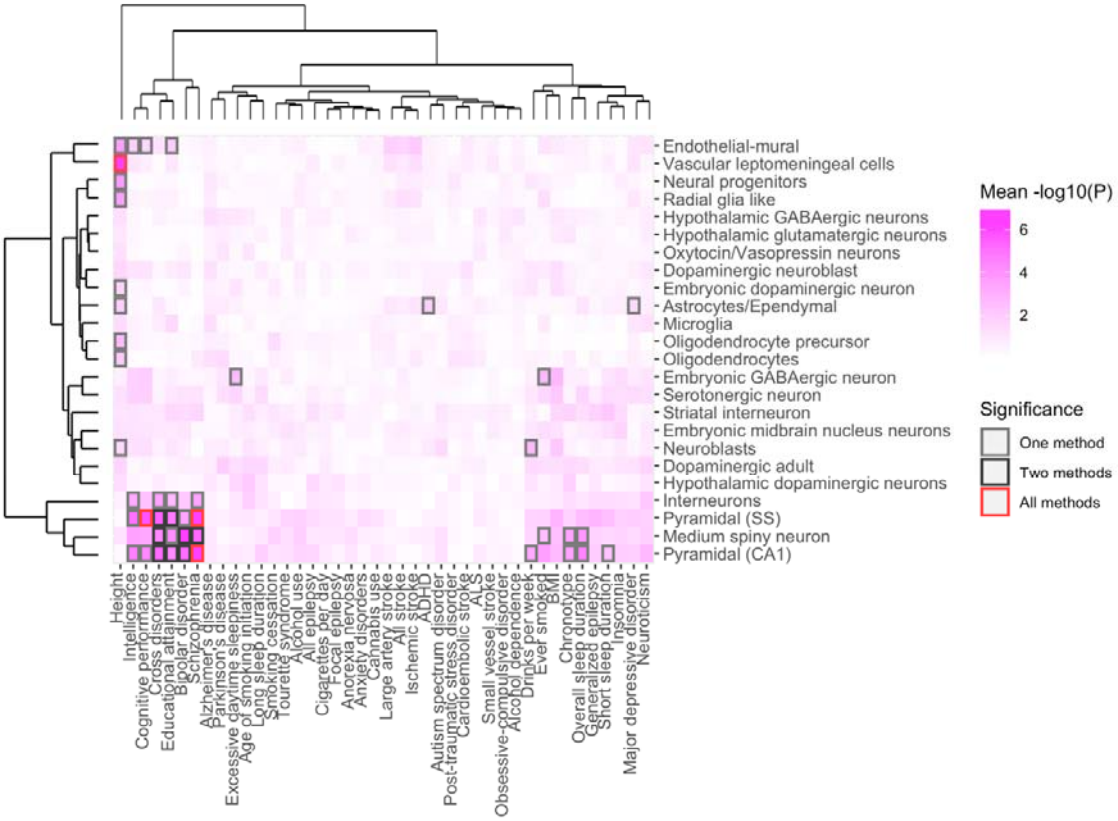

Figure 3. Overview of enriched cell types of 42 common-variant psychiatric, neurologic and behavioral/quantitative GWAS results in the KI dataset.

Abbreviations: ADHD; attention deficit hyperactivity disorder, ALS; amyotrophic lateral sclerosis, BMI; body mass index.

*Analyses from LDSC, DEPICT and MAGMA (top 10% mode), referred to as 'methods' in the graph, show enrichment in MSNs and pyramidal cells (CA1) and pyramidal cells (SS) across brain-related phenotypes. The largest degree of consistency was found in SCZ and cognitive performance. Phenotypes and cell types are grouped by hierarchal clustering Shades of pink are proportional to the mean strength of association (-log10(P)) of all methods. The color of the frames refers to the number of methods that identified a given cell type as significant in a given phenotype, after Bonferroni correction (P<0.05/(24*42)). Grey frames: one method (intelligence, excessive daytime sleepiness, ADHD, drinks per week, ever smoked, chronotype, overall sleep duration, short sleep duration, MDD). Black frames: two methods (cross-disorders, educational attainment, BIP). Red frames: all three methods (human height, cognitive performance, SCZ).*

Cell type enrichment analyses using additional scRNA datasets

Finally, we performed additional analyses with FUMA using additional mouse (n=11) and human (n=7) gene expression datasets to compare our findings and to assess consistency between rodent and human data. Using mouse and human gene expression datasets, we were able to identify at least one implicated cell type in 22 phenotypes (Figure S26, Table S14). Using human gene expression data, 24 cell types were enriched in at least one phenotype, while those were 70 cell types using mouse scRNA data. Consistent with findings from 10x Genomics and KI datasets, pyramidal cells from various mouse brain regions, among which pyramidal cells (CA1) and pyramidal cells (SS), were implicated in SCZ and cognitive performance. Pyramidal cells were also enriched in numerous psychiatric disorders, SUDs and behavioral/quantitative phenotypes. Along with pyramidal cells, inhibitory GABAergic and MSNs were consistently enriched in psychiatric disorders, SUDs and behavioral/quantitative phenotypes. Using human datasets, enrichment of pyramidal cells (CA1) was replicated in SCZ, cognitive performance, intelligence, cross-disorders, and overall sleep duration. Additionally, pyramidal cells (CA1) were enriched in cigarettes per day. However, the strongest consistent evidence was found for enrichment of GABAergic neurons from the prefrontal cortex and midbrain in various psychiatric disorders, SUDs and behavioral/quantitative phenotypes. Consistent with findings using 10x Genomics and KI data, fewer enriched cell types in neurological disorders were identified, with exclusively enriched human microglia in Alzheimer's disease and human inhibitory GABAergic neurons from the prefrontal cortex and midbrain in generalized epilepsy. Generalized epilepsy was the only neurological phenotype in which implicated mouse cell types were identified, namely certain pyramidal neurons and certain inhibitory neurons. We thus largely replicated the main cell type enrichment findings from the 10x Genomics and KI dataset in mouse datasets and in human datasets using FUMA.

Discussion

Here, we provide a comprehensive overview of specific brain cell types implicated in a range of brain-related phenotypes using both mouse and human brain scRNA data. We show that results from brain-related GWAS data consistently map to excitatory pyramidal neurons (CA1), pyramidal neurons (SS) and inhibitory MSNs and less so to glial and embryonic cells. The largest degree of consistency across methods and tissue origins (rodent and human) was found for implication of pyramidal cells in schizophrenia and cognitive performance.

Our SNP-h^2 and genetic correlation findings confirm that neurological disorders are genetically distinct from one another and from psychiatric and SUDs, as well as from behavioral/quantitative phenotypes, which is in line with previous evidence[3,10]. Consistent with these findings, we found that GWAS findings from psychiatric disorders, SUDs and brain-related behavioral/quantitative phenotypes, but not neurological disorders, consistently map to excitatory hippocampal pyramidal neurons (CA1), excitatory pyramidal neurons (SS) and inhibitory MSNs and much less to glial and embryonic cells. Alzheimer's disease was the only malady targeted here that showed evidence of exclusively human glial cells being implicated, underscoring the importance of key transcriptomic differences between human and mouse microglial signatures[29]. We replicated our main findings with multiple external scRNA datasets using FUMA. This provides further evidence that genetic underpinnings of neurological disorders are distinct from those of psychiatric, SUDs and behavioral/quantitative phenotypes[8,10]. Our main findings were based on the identification of cell types by LDSC, DEPICT and MAGMA top 10% mode. MAGMA linear mode was omitted because it was deemed too lenient and thus prone to type I error inflation. This concurs with previous studies reporting that binned MAGMA analyses in linear mode inflate results since the binned scores can have strong correlations with the average gene expression across cell types[28]. Also in agreement with previous lines of evidence, we confirm that the statistical foundation of the SNP-wise mean gene analysis model MAGMA <1.07 may result in biased associations of cell types[11].

One discrepancy between KI- and 10x Genomics-derived cell types was that neuroblasts were commonly enriched for psychiatric and behavioral phenotypes using the 10x Genomics dataset, while using the KI dataset, hippocampal and striatal neurons together with interneurons mapped to genomic results of the same phenotypes. This discrepancy could be a consequence of a lower sequencing depth in the 10x Genomics dataset (approximately 18,500 mapped reads per cell) than in the KI dataset (approximately 1.2 million mapped reads per cell). Notably, the minimum sequencing depth is generally considered to be between 25,000 and 50,000 mapped reads per cell[30]. This suggests that the relatively low sequencing depth of the 10x Genomics dataset led to overlapping cell clusters. Moreover, this could explain why KI-derived cell types were more specific. Additionally, although k-means clustering is commonly used for single cell data, selecting the right value of k is challenging[30]. PCA-based clustering methods would be particularly well-suited for low sequencing depth[31], and for instance could be expanded to initially select significant principal components with PCA and use these for subsequent clustering[32].

Although we provide new insight with the largest and most comprehensive study of cell type enrichment in brain-related disorders, our results should be interpreted in light of inevitable limitations. First, we have employed methods that bin gene sets based on the specificity for a particular cell type. Using MAGMA, it is possible to test whether the genes specific to a phenotype

are enriched in genetic associations of that phenotype while controlling for genetic associations of another phenotype[10]. However, as our main goal was to identify enriched cell types, such conditional analyses are beyond the scope of this study. Second, gene expression data were obtained from adolescent mice. However, we found that microglia associated with age-induced neuroinflammation were exclusively found to be enriched in Alzheimer's disease using human scRNA datasets, whereas no enriched glial cells were identified using mouse scRNA datasets. Therefore, mouse gene expression data from not only a spatial, but also a temporal resolution is warranted for future research to identify cell types implicated in disease during development. Additionally, improved coverage of brain-related regions, such as the entire CNS[10], is warranted for future research.

The identification of a specific subset of brain cell types being implicated in various brain disorders only marks the beginning of elucidating causal biological pathways. One question future research should address is what the effects of genetic variants in the noncoding genome are. One way to address this question is using an activity-by-contact model[33]. This model allows for the identification of cell type-specific enhancers and their target genes by leveraging single-cell chromatin accessibility and enhancer activity data. Additional insight could be obtained by performing cell prioritization analyses from human post-mortem brain samples and/or induced pluripotent stem cells from individuals with relevant genetic backgrounds using LDSC, MAGMA and DEPICT to identify genes that are predicted to be functionally similar to causal genes. Additionally, the recently developed computational toolkit CELLECT can provide additional insight in cell type enrichment[34]. CELLECT builds upon gene prioritization models, such as LDSC, DEPICT and MAGMA and subsequently performs cell type prioritization analyses using a continuous representation of cell type expression, rather than binary representation. Finally, statistical power is currently a major challenge in genetic studies. Future studies might benefit from multi-trait analysis of GWAS (MTAG)[35], which is a method for analysis of multiple GWASs, thereby increasing the statistical power of each trait analyzed to identify genetic associations.

In sum, by incorporating different tools that rely on different assumptions and algorithms we provide robust evidence for a subgroup of neuronal cell types consistently implicated in several brain-related phenotypes. We thus provide a framework that furthers the understanding of cell types involved in brain-related phenotypes at a cellular level that can serve as a basis for future, more hypothesis-driven research.

Data availability

All scRNA datasets used in this study are publicly available. All summary statistics are publicly available and the sources are listed in Table S1. We have made our code publicly available at https://github.com/mitchellolislagers/cell_type_enrichment_pipeline so that with the advent of new GWASs researchers may readily apply our pipeline to new data.

Acknowledgements

Funding was provided through a personal Rudolf Magnus Talent Fellowship (H150) to Jurjen Luykx. The authors thank Kevin Kenna, Mark Bakker and Nathan Skene for fruitful input on statistical analyses.

Disclosure

All authors declare they have no conflict of interest.

References

1. Armstrong, T.D. & Costello, E.J. Community studies on adolescent substance use, abuse, or dependence and psychiatric comorbidity. *J Consult Clin Psychol* **70**, 1224-39 (2002).
2. Couwenbergh, C. *et al.* Comorbid psychopathology in adolescents and young adults treated for substance use disorders: a review. *Eur Child Adolesc Psychiatry* **15**, 319-28 (2006).
3. Brainstorm, C. *et al.* Analysis of shared heritability in common disorders of the brain. *Science* **360**(2018).
4. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat Rev Genet* **20**, 467-484 (2019).
5. Cross-Disorder Group of the Psychiatric Genomics Consortium. Electronic address, p.m.h.e. & Cross-Disorder Group of the Psychiatric Genomics, C. Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell* **179**, 1469-1482 e11 (2019).
6. Young, A.I. Solving the missing heritability problem. *PLoS Genet* **15**, e1008222 (2019).
7. Giral, H., Landmesser, U. & Kratzer, A. Into the Wild: GWAS Exploration of Non-coding RNAs. *Front Cardiovasc Med* **5**, 181 (2018).
8. Skene, N.G. *et al.* Genetic identification of brain cell types underlying schizophrenia. *Nat Genet* **50**, 825-833 (2018).
9. Watson, H.J. *et al.* Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. *Nat Genet* **51**, 1207-1214 (2019).
10. Bryois, J. *et al.* Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson's disease. *Nat Genet* **52**, 482-493 (2020).
11. Yurko, R., Roeder, K., Devlin, B. & G'Sell, M. H-MAGMA, inheriting a shaky statistical foundation, yields excess false positives. *bioRxiv*, 2020.08.20.260224 (2020).
12. Adamson, S.J., Todd, F.C., Sellman, J.D., Huriwai, T. & Porter, J. Coexisting psychiatric disorders in a New Zealand outpatient alcohol and other drug clinical population. *Aust N Z J Psychiatry* **40**, 164-70 (2006).
13. Chan, Y.F., Dennis, M.L. & Funk, R.R. Prevalence and comorbidity of major internalizing and externalizing problems among adolescents and adults presenting to substance abuse treatment. *J Subst Abuse Treat* **34**, 14-24 (2008).
14. Carey, C.E. *et al.* Associations between Polygenic Risk for Psychiatric Disorders and Substance Involvement. *Front Genet* **7**, 149 (2016).
15. Medic, N. *et al.* Increased body mass index is associated with specific regional alterations in brain structure. *Int J Obes (Lond)* **40**, 1177-82 (2016).
16. Zeisel, A. *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138-42 (2015).
17. La Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* **167**, 566-580 e19 (2016).
18. Romanov, R.A. *et al.* Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat Neurosci* **20**, 176-188 (2017).
19. 10X Genomics. Transcriptional Profiling of 1.3 Million Brain Cells with the Chromium Single Cell 3' Solution. (2017).
20. Moustafa, A.A., Phillips, J., Keri, S., Misiak, B. & Frydecka, D. On the Complexity of Brain Disorders: A Symptom-Based Approach. *Front Comput Neurosci* **10**, 16 (2016).

21. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
22. de Leeuw, C.A., Mooij, J.M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* **11**, e1004219 (2015).
23. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766-D773 (2019).
24. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
25. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236-41 (2015).
26. Pers, T.H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* **6**, 5890 (2015).
27. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
28. Watanabe, K., Umicevic Mirkov, M., de Leeuw, C.A., van den Heuvel, M.P. & Posthuma, D. Genetic mapping of cell type specificity for complex traits. *Nat Commun* **10**, 3222 (2019).
29. Galatro, T.F. *et al.* Transcriptomic analysis of purified human cortical microglia reveals age-associated changes. *Nat Neurosci* **20**, 1162-1171 (2017).
30. Andrews, T.S. & Hemberg, M. Identifying cell populations with scRNASeq. *Mol Aspects Med* **59**, 114-122 (2018).
31. Menon, V. Clustering single cells: a review of approaches on high-and low-depth single-cell RNA-seq data. *Brief Funct Genomics* **18**, 434 (2019).
32. Shekhar, K. *et al.* Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell* **166**, 1308-1323 e30 (2016).
33. Fulco, C.P. *et al.* Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat Genet* **51**, 1664-1669 (2019).
34. Timshel, P.N., Thompson, J.J. & Pers, T.H. Genetic mapping of etiologic brain cell types for obesity. *Elife* **9**(2020).
35. Wu, Y. *et al.* Multi-trait analysis for genome-wide association study of five psychiatric disorders. *Transl Psychiatry* **10**, 209 (2020).