Genetic associations on major depression: curation and functional analysis

Judith Pérez Granado

TESI DOCTORAL UPF / 2022

Thesis supervisors

Dra. Janet Piñero and Dra. Laura I. Furlong

DEPARTMENT OF MEDICINE AND LIFE SCIENCES

Universitat Pompeu Fabra Barcelona

Funding

This thesis research was conducted at the Integrative Bioinformatics Group, which belongs to the Research Programme on Biomedical Informatics (GRIB) at the Barcelona Biomedical Research Park (PRBB); a joint research programme of the Hospital del Mar Medical Research Institute (IMIM) and the Department of Medicine and Life Sciences of the Universitat Pompeu Fabra (UPF). This thesis includes the research conducted during a 6 months stay at the Laboratorio Internacional de Investigación sobre el Genoma Humano (LIIGH) at the Universidad Nacional Autónoma de México.

The research presented here has been supported by Instituto de Salud Carlos III cofinanced by Fondo Social Europeo, "El Fondo Social Europeo invierte en tu futuro" [FI18/00034] and Instituto de Salud Carlos III [MV20].

Acknowledgements

I would like to thank my thesis supervisors Dr Laura Furlong and Dr Janet Piñero, for their help and insightful comments during these years. This thesis would not have been possible without them. Thanks to Laura for always making me go beyond and providing alternative perspectives. Thanks to Janet for being an example of resilience and passing on her passion.

Thanks also to Dr Alejandra Medina-Rivera, for hosting me at the Laboratorio Internacional de Investigación sobre el Genoma Humano (LIIGH) at the Universidad Nacional Autónoma de México, for her support despite the challenging circumstances and to her group, for making me feel part of the team.

Thanks also to the Integrative Biomedical Informatics group: Juanma, Angela, Francesco, Emilio, Josep, Alexia, Miguel Àngel and of course Ferran; for all their advices along this journey. I would like to thank the GPCR Drug Discovery group too: Mariona, Tomeck, David, Adri, Brian and of course Jana, for letting me invade their office during these years.

I cannot forget about my GRIB fellow researchers, with whom I experienced this journey. Especially, Mariona, a friend who I take with me after all our teatime talks, roomscapes and what is still to come. Also thanks to Juanma, Coté, Giulia, Patricia, Quim, Alberto, Carolina, Eric; these would not have been the same without all the lunchtime talks and beer sessions.

I would like to thank Carina, Chus, Alfons, Miguel and Vicent for always being helpful, making paperwork and installations easier, for all the little talks and their goodwill. I cannot forget about Neus and Nàtalia, from the UPF PhD Programme in Biomedicine who are always there patiently answering all my questions.

Thank you to my friends as well, who have always been there for me, for their advice and the fun times. Thanks to my university friends, especially to María and her little boys. Thanks to my lab friends&co: Judit, Laura, Loreto, Àlex, Miguel, Luka and Marc. Thanks to Dr MªÀngels Calvo for introducing me to the world of science and research; as well as Leo for creating such a good working environment

which became a friendship. Thanks also to Laura, Sat and Marie for their friendship over the years. Thanks to Yu, my sister-in-law turned friend.

Sobretot vull donar-li les gràcies al Quim, per sempre fer-me costat, donar-me l'empenta que a vegades necessito i treure'm un somriure cada dia. Encara que no ho puguin llegir gràcies al Nazgul i la Nami, per transmetrem paciència, tranquil·litat i un amor incodicinal.

Por último pero no menos importante agradecérselo a mi familia, por su apoyo, por creer en mi y animarme a perseguir mis sueños siempre. A mi padre por esa mente analítica y esas charlas en las que solamente los dos nos entenemos. A mi madre por ser tan mami, por cuidarme, escucharme, aconsejarme, motivarme y mimarme. Y a mi hermana por siempre estar ahí para un café o unas bravas, para hablar y animarnos la una a la otra.

Abstract

Major depression (MD) is the leading cause of impairment worldwide. The lack of understanding of its biological underpinnings hampers the development of better diagnostic tools and treatments. Thanks to the advances in genetic association studies, multiple genetic variants significantly associated with MD have been identified. In this thesis, we aim to leverage this knowledge to advance in the understanding of MD and unravel its molecular mechanisms. For that, we developed curation guidelines to evaluate available genetic association data on MD of diverse nature, and created an expert-curated database of genetic variants associated with MD. Then, we leveraged these data and functional genomic tools to unravel the role of these variants in disease pathogenesis and propose mechanistic hypotheses. In light of the plethora of tools available to perform such analyses, we conducted a benchmarking analysis to evaluate their performance and compare their outcomes; highlighting the need for guidelines for method selection and evaluation. Overall, this thesis contributes to filling the gap regarding the quality assessment of genetic studies on MD, and to advance in discovering the functional role of MD-associated variants by using in silico approaches.

Resum

La depressió major (DM) és la principal causa d'incapacitat en tot el món. La falta de comprensió dels seus fonaments biològics dificulta el desenvolupament de millors diagnòstics i tractaments. Gràcies als avanços en estudis d'associació genètica, s'han identificat múltiples variants genètiques significativament associades a la DM. En aquesta tesi, volem aprofitar aquests coneixements per avançar en la comprensió de la DM i descobrir els seus mecanismes moleculars. Per a això, hem desenvolupat unes directrius de curació per avaluar l'ampli ventall de dades d'associació genètica disponibles sobre la DM i hem creat una base de dades de variants genètiques associades a la DM que ha estat curada per experts. Un cop finalitzada, vam aprofitar aquestes dades i diverses eines de genòmica funcional per entendre el paper d'aquestes variants en la patogènesi de la malaltia i proposar hipòtesis mecanístiques. Davant de la plètora d'eines disponibles, vam dur a terme una anàlisi de referència per avaluar el seu funcionament i comparar els seus resultats, on destaquem la necessitat de directrius per seleccionar I avaluar els mètodes. Globalment, aquesta tesi contribueix a omplir el buit que existeix pel que fa a l'avaluació de la qualitat dels estudis genètics sobre la DM, I avançar en el descobriment del paper functional de les variants associades a la DM mitjçant l'ús de mètodes in silico.

Preface

Major depression (MD) is the most common disabling disorder worldwide. MD diagnosis is based on its symptomatology, and its treatment is based on psychotherapy and pharmacology. However, these are not effective in 50% of cases. The lack of better treatments and the absence of biomarkers that could aid in its diagnosis or treatment reflects the need to better understand MD biological underpinnings.

Huge advances have been made thanks to genetic association studies on MD, which have identified genetic variants (GVs) significantly associated with the disease. However, to the best of our knowledge, no guidelines exist to evaluate multiple evidences from different types of studies in the context of MD or complex disease in general. Therefore, in this thesis, we have developed curation guidelines to evaluate associations of diverse nature and have created an expert-curated database of GVs associated with MD.

As a complex disease, MD arises from the interaction of multiple environmental and genetic factors. The latter have been attributed to multiple GVs with minor effects and mostly lying in non-coding regions. These GVs do not usually have a direct link to their target genes but are expected to impact disease by altering regulatory mechanisms. Therefore, posterior functional analyses are required to elucidate how these GVs ultimately impact disease pathogenesis.

In this context, we have contributed to the development of a bioinformatics pipeline that leverages genome-wide association studies (GWAS) data, several bioinformatic tools and genomic annotation data of diverse nature. The application of this pipeline has enabled the identification of GVs potentially relevant for MD as well as the proposal of mechanistic hypotheses on how these GVs impact MD pathogenesis.

Currently, there is a plethora of methods and tools available for the analysis of GWAS data. These have adapted to data accessibility constraints, being the most commonly available data type summary statistics, followed by full genome summary statistics. However, there is no available criteria to aid in their selection or validation. In searching for tools that could help identify GVs' underlying biological mechanisms, we have developed a workflow that systematically

compares different tools' outcomes. This comparison also considers the results' biological impact, illustrating a high divergence between their outcomes. Overall, these findings revealed an important issue when implementing post-GWAS analysis to unveil disease pathophysiological mechanisms for drug prioritisation or biomarker research.

Table of contents

Abbreviations

CGS	candidate gene studies
CNS	central nervous system
eQTL	expression quantitative traits loci
GV	genetic variant
GWAS	genome-wide association studies
HPA	hypothalamic-pituitary-adrenal
LD	linkage disequilibrium
MD	major depression
QTL	quantitative traits loci
SNP	single nucleotide polymorphisms
TF	transcription factor
TFBS	transcription factor binding site
TM	text mining
TWAS	transcriptome-wide association studies
WGS	whole-genome sequencing

1. INTRODUCTION

1.1 Major Depression

Major depression (MD) is one of the most common psychiatric disorders worldwide. Its prevalence varies by region and country, being 6.9% in the EU[1], 6.7% in the USA[2] and between 1-2% in Japan[3]. MD has the highest morbidity burden in low- and middle-income countries, but it affects all countries regardless of their gross domestic product[4]. According to the World Health Organization (WHO), MD affects 350 million people worldwide and is expected to be the leading cause of disease burden by 2030[4].

MD is a complex disease caused by the interaction of multiple genetic and environmental factors, with an estimated heritability from twins of 37%[5]. Environmental factors can vary in their nature as well as in their timing in life. MD polygenicity results from the interaction of multiple genetic variants (GVs)[6], where multiple single nucleotide polymorphisms (SNPs) with minor individual effects have been identified[7,8]. In contrast, the role of structural variants in MD has been debated. While rare copy number variants (CNVs) may not play a significant role in MD compared to other psychiatric disorders[9], short deletions are more common in MD patients and most likely alter gene expression regulation[10].

The diagnosis of MD is based on symptomatology, grouping different etiologies, severity levels and treatment responses under the same diagnosis. The diagnostic criteria provided by the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders - 5th Edition (DSM-5) characterises MD by symptoms of depressed mood or anhedonia (loss of interest or pleasure) (Table 1). These must be accompanied by at least four of the following symptoms: appetite or weight changes, fatigue or loss of energy, difficulty concentrating, psychomotor agitation or retardation, feelings of worthlessness or guilt and suicidality[11]. This diagnostic system highly overlaps with the one from the WHO's International Classification of Diseases (ICD)[12] (Table 1). Both require the symptoms to be present most of the time for over two weeks and not be better explained by other conditions. Because of this symptomatic and diagnostic heterogeneity, the Psychiatric Genomics Consortium (PGC) decided to use the term "Major Depression" to cover a broader phenotype. This term includes both

lifetime major depressive disorder and depressive symptoms, which a practitioner has diagnosed according to diagnostic criteria or correspond to minimally phenotyped cases (e.g., self-reported MD). We use this terminology throughout this thesis to encompass all these phenotypes.

DSM-5	ICD-10
Five or more symptoms, at least one of which must come from the "A" criteria: "A" criteria 1. Depressed mood 2. Markedly diminished interest or pleasure in almost all activities "B" criteria 1. Significant weight loss/gain or decrease/increase in appetite 2. Insomnia or excessive sleep 3. Psychomotor agitation or retardation 4. Fatigue or loss of energy 5. Feelings of worthlessness or excessive/inappropriate guilt 6. Diminished concentration or indecisiveness 7. Recurrent thoughts of death, suicidal ideation, plans or an attempt	Six or more symptoms, including two from the following: 1. Depressed mood 2. Loss of interest and enjoyment 3. Reduced energy leading to increased fatigability and diminished activity Three or more typical symptoms from the following: 1. Reduced concentration and attention 2. Reduced self-esteem and self-confidence 3. Ideas of guilt and unworthiness (even in mild type of episode) 4. Bleak and pessimistic views of the future 5. Ideas or acts of self-harm or suicide 6. Disturbed sleep 7. Diminished appetite

Table 1. Diagnostic criteria for major depressive disorder. American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders - 5th Edition (DSM-5) and International Classification of Diseases (ICD)-10 diagnostic criteria for major depressive disorder. Adapted from McIntosh AM, Sullivan PF, Lewis CM. Uncovering the Genetic Architecture of Major Depression. Neuron. 2019 Apr 3;102(1):91-103. doi: 10.1016/j.neuron.2019.03.022.

Significant genetic overlap between MD and other psychiatric disorders, such as schizophrenia or anxiety, point to pleiotropic GVs[13,14]. Moreover, MD is frequently comorbid with chronic diseases, including cancer and cardiovascular, metabolic, inflammatory or neurological disorders[15]. These co-occurrences could be due to direct mechanisms such as biological processes (e.g., higher cortisol in MD contributing to metabolic diseases) or indirect mechanisms such as treatment-induced (e.g., immunotherapy in cancer), psychosocial factors (e.g., childhood

maltreatment), or behavioural reasons (e.g., smoking or physical activity) [15].

According to the WHO, the MD burden is 50% higher for females than males[4]. However, we are still in the early stages of understanding sex differences in neural circuits and how they relate to observed differences in disease prevalence between males and females[16]. In addition, progress toward a more inclusive data collection system that disambiguates data by gender is required before closing the knowledge gap.

The main treatment strategies for MD are psychotherapy and pharmacological therapy[17]. The latter takes a long time to become effective, and 50% of patients do not fully respond[18]. These can also have adverse effects, ranging from headache, insomnia, and diarrhoea to somnolence and constipation[19]. Identifying objective clinical measurements could help clinicians diagnose MD and make informed treatment decisions[20]. All in all, the absence of more effective MD treatments reflects the lack of understanding of MD physiopathology and the need for biological markers[21].

Biomarkers for MD are still at the research stage[20,22], which is the reason why significant efforts have been made to better understand the physiopathological mechanisms of MD by leveraging genetic information. Diverse genetic study designs, such as candidate gene studies (CGS) or genome-wide association studies (GWAS), have been used to identify GVs that play a potential role in disease pathogenesis. These studies, along with clinical observations of pharmacological response in patients with MD and other diseases, have resulted in the development of diverse MD theories to explain its physiopathology.

1.1.1. Neurobiological hypothesis

The complexity of MD pathogenesis motivates the integration of different theories that consider the interaction between diverse neurobiological mechanisms.

1.1.1.1. Monoamine theory

This was the first proposed theory back in the 1960s. It was based on the therapeutical effect of monoamine oxidase inhibitors and tricyclic antidepressants[23,24]. It was hypothesised that a deficit of monoamine

neuromediators (i.e., serotonin, norepinephrine and dopamine) in the central nervous system (CNS) could contribute to the development of MD[25]. Indeed, current first-line treatments for MD include selective serotonin reuptake inhibitors and norepinephrine reuptake inhibitors. However, the underlying pathology goes beyond the deficit of monoamine neuromediators, as reflected by the lack of response for 30-60% of MD patients[26].

1.1.1.2. Stress and HPA axis

Early stressful life events and chronic stress have been linked to the onset of MD, as well as other diseases such as heart disease or obesity[27]. A dysregulation of the hypothalamic-pituitary-adrenal (HPA) axis, which responds to and adapts to environmental changes, promotes the release of different hormones, including cortisol, adrenocorticotropic hormone, corticotropin-releasing hormone and vasopressin. The abnormal functioning of the HPA axis is a common feature in MD patients, resulting in higher cortisol levels[28,29], which have been linked to disease severity, particularly in melancholic depression[30]. Unfortunately, treatments that regulate the HPA axis have not been very effective[31].

1.1.1.3. Glutamate signalling pathway

The main excitatory neurotransmitter is glutamate, and increased levels have been reported in the blood and brain of MD patients. Stress factors can increase glutamate, which binds to ionotropic glutamate receptors, including N-methyl-D-aspartate receptors (NMDARs) and α-amino-3-hydroxy-5-methyl-4-isoxazole-propionic acid receptors (AMPARs)[6]. NMDAR antagonists have proven antidepressant effects, whereas traditional antagonists reduced glutamate release, esketamine changed the paradigm[32]. This effective fast-acting antidepressant increases transient prefrontal glutamate for rapid restoration of synaptic connectivity, which is reduced in MD[32,33]. Nonetheless, this antidepressant lacks retrospective studies due to the recentness of its approval in clinical practice and is not exempt from side effects such as nausea or headache[32].

1.1.1.4. Gamma-Aminobutyric Acid (GABA)

The main inhibitory neurotransmitter is gamma-Aminobutyric Acid (GABA). Contrary to glutamate, MD patients show lower GABA levels,

and it has been proposed that a functional imbalance of both systems influences the pathophysiology of MD [34].

1.1.1.5. Disturbance of Neurogenesis and Neuroplasticity

A wealth of studies supports the disruption of brain neurogenesis and neuroplasticity in MD. Neurotrophic factor disturbances, primarily lower brain-derived neurotrophic factor (BDNF) levels, are thought to cause neuronal atrophy, decreased neurogenesis, and glia support destruction[35]. BDNF has been linked to serotoninergic neuron function, maintaining their differentiation and survivability, and has been tagged as a promising synaptic regulator[35,36].

1.1.1.6. Neuroinflammation and cytokine theory

The immune-inflammation hypothesis focuses on the interaction between increased cytokine production, which activates the HPA axis, altering neurotransmitters' synthesis and metabolism[37]. This, in turn, affects neuronal apoptosis, neurogenesis, and neuroplasticity. Furthermore, the gut microbiota, which interacts with the CNS and the immune system, can produce neuroactive substances that mimic host-signalling molecules, which may contribute to MD development[38].

1.1.1.7. Others

The pathophysiology of MD is also associated with alterations of other mechanisms and biological pathways as well as their interaction. For instance, oxidant-antioxidant imbalance[39], mitochondrial dysfunction[40], circadian rhythm disturbances[41], or gut microbiome alterations[42].

1.2. Unravelling the genetic basis of MD

The polygenic nature of MD and its interaction with environmental factors have challenged the identification of genes and GVs that increase disease susceptibility[6]. In recent years, great advances in molecular genetic research have been made thanks to more affordable genotyping tools and a better understanding of human genetic variation. Initially, candidate genes thought to be involved in MD neurobiology were identified and evaluated by linkage and CGS. However, these were generally conducted on small samples, usually leading to ineffective or

unreplicable findings[43]. As a result, there was a shift toward GWAS studies, which examine millions of common GVs in an unbiased and hypothesis-free manner.

1.2.1. Candidate gene studies (CGS)

CGS evaluate the effects of GVs on genes that may contribute to disease susceptibility by affecting their protein product or gene expression regulation (Box 1). CGS were first introduced in 1995 and had been running since then, but their popularity dropped around 2005[44]. Despite the benefit of prioritising potentially relevant genes, CGS are intrinsically biased toward genes and biological pathways that researchers select based on prior knowledge. Furthermore, their validity has been questioned due to the lack of reproducibility and the typically small sample size used in these studies, which generally ranged between tens and hundreds of individuals[43,45]. For instance, due to the widespread use of serotonin reuptake inhibitors in MD treatment, the gene SLC6A4, which encodes for the serotonin transporter in charge of its reuptake, has been extensively examined[46]. Some studies have associated a polymorphism in its promoter region with lower serotonin reuptake and an increased risk of MD[47,48], while other meta-analyses have reported no significant association[49]. Also derived from the monoamine theory, the catechol-O-methyltransferase (COMT) Val158Met polymorphism has been evaluated. COMT degrades catecholamines such as dopamine, and both Val and Met alleles have been associated with MD risk, raising conflicting findings[50,51]. In contrast, other analyses have even failed to detect a significant association[52,53]. Another frequently studied polymorphism is Val66Met, which is associated with decreased BDNF activity and higher MD risk with contradictory findings[54].

Recently, Border et al. re-evaluated the association of 18 historical candidate genes with MD in samples ranging from 62,138 to 443,264 individuals[43]. The genes under evaluation included those mentioned above and others which have been reported to have large genetic effects in much smaller samples via CGS. However, the authors could not replicate the original findings, failing to identify the large effect GVs in sample sizes orders of magnitude larger; and hypothesising that those were potential false positive results[43]. Similarly, historical candidate genes for schizophrenia have been reported as controversial[55]. Based on genetic association findings for MD and other psychiatric disorders, Duncan et al. address the distribution of GVs across the entire genome,

with many of them falling in poorly understood regions of the genome, which had not previously been examined by CGS[45]. The genetic and non-genetic heterogeneity of MD, as well as the typically small sample size of CGS, undermine their validity[56]. It has been estimated that at least 1,000 cases and controls would be required to identify GVs with a 1.5 odds ratio[57]. According to the National Institute of Mental Health, CGS should be left in the past and move towards more reproducible and statistically rigorous studies, especially when considering psychiatric disorders[58].

Box 1. Candidate gene studies (CGS)

- CGS are hypothesis-driven and focus on a particular gene in the genome (candidate gene) to evaluate its association with disease.
- Candidate genes are chosen based on an a priori hypothesis about their role in the disease. These may be biologically relevant, given underlying physiopathological mechanisms or pharmacological evidence.
- CGS examine the association between a gene's specific allele (or set of alleles) and the disease. These GVs are common in the population.
- CGS generally follow a case-control design, testing the gene in randomly selected subjects with and without the disease.
- CGS compare the allelic frequencies of selected GVs between cases and controls.

1.2.2. Genome-wide association studies (GWAS)

GWAS have supported the widely accepted premise that MD is a complex polygenic disorder in which multiple common GVs contribute to disease susceptibility. Generally, these GVs are in non-coding regions of the genome, each playing a minor role (odds ratio around 1.3)[59] (Box 2). There are two types of GWAS studies, the ones that use SNP arrays followed by imputation and those based on whole-genome sequencing (WGS). In the former, imputation estimates the effects of GVs that have not been directly genotyped before assessing the association between GVs and the trait under investigation[60]. It requires summary statistics data and linkage disequilibrium (LD) information from reference population data, such as the 1000 Genomes Project, which can differ between ethnic groups. Altogether, imputation increases the power of downstream analyses[61]. Regarding WGS, several studies have been performed on MD[62,63], with a focus on pharmacogenomics as well[64]. However, these studies are not a common practice yet, because large sample sizes (i.e., 1,000,000) are required to produce reliable

results, as evidenced by other complex disorders such as schizophrenia[65]. Furthermore, WGS, compared to SNP arrays, are more expensive[66]. This thesis focuses on GWAS that employ SNP arrays and imputation.

The first GWAS on MD dates back to 2009 and included 1,738 patients and 1,802 controls[67]. Even with sample sizes comparable to other psychiatric disorders, initial research was fruitless, with SNPs failing to pass genome-wide significance (i.e., p-value $5x10^{-8}$)[68]. Since then, the GWAS success rate raised thanks to: 1) an increase in sample size due to larger study cohorts and meta-analysis; and 2) a reduction in phenotype heterogeneity[69]. Due to the minor contribution of numerous SNPs to MD overall risk, estimates on the required GWAS sample size range from 3,000 to 75,000 to identify multiple MD associations[70,71]. These numbers may vary depending on the expected number of GVs identified and their effect size.

Box 2. Genome-wide association studies (GWAS)

- GWAS evaluate common GVs, which are present in at least 5% of the population.
- GWAS test the association to disease of one million or more common GVs known as single nucleotide polymorphisms (SNPs).
- GWAS are unlikely to detect rare variants and cannot identify CNVs.
- GWAS are typically case-control studies that compare the allele frequencies of cases with the disease to controls without.
- To account for the million independent tests conducted, the GWAS significance threshold is $5x10^{-8}$. This threshold could be considered overly conservative since SNPs are, to some extent, correlated due to linkage disequilibrium (LD) (i.e., the nonrandom association of alleles).
- GWAS require large sample sizes, in the thousands, to reach enough power to detect significant associations.

Adapted from Dunn, E. C., Wang, M.-J. & Perlis, R. H. A Summary of Recent Updates on the Genetic Determinants of Depression. in Major Depressive Disorder 1–27 (Elsevier, 2020). doi:10.1016/B978-0-323-58131-8.00001-X.

Regarding sample size, on the one hand, GWAS meta-analyses combine multiple GWAS from different studies to assess the relationship between SNPs and a trait or disease. These have contributed to the discovery of most of the recently identified GVs associated with MD. Nonetheless, they are not exempt from challenges such as phenotype

or ancestry heterogeneity, data availability or statistical approaches diversity[72]. On the other hand, initiatives such as the PGC, a consortium from over 20 countries, and the CONVERGE consortium (China, Oxford and Virginia Commonwealth University Experimental Research on Genetic Epidemiology), the biggest Han Chinese population sequencing cohort, have encouraged a community effort toward the collection of large cohorts[73,74]. Additionally, companies such as 23andMe, based on self-reported surveys, have also contributed with large amounts of individual data.

Additionally, MD phenotype heterogeneity poses a challenge to GWAS, where two different approaches can be adopted: very large sample sizes with minimal phenotyping (as assessed through patients' self-reports) or smaller but more fine-grained and rigorously phenotyped samples (as assessed by clinical practitioners)[46]. There have been reported differences in heritabilities, being higher for strict DSM-5 criteria compared to minimal phenotyping[75]. Another criticism stems from a lack of phenotype specificity, in which shared genetic liability with other psychiatric disorders is greater for minimal phenotyping, potentially due to misdiagnosis[75]. However, both could be complementary; while the first benefits from increased power to identify more GVs, the second can identify subphenotype and more severe phenotype differences[44,46].

Recent findings from GWAS on MD include 44 risk loci identified by Wray et al. in 135,458 cases and 344,901 controls[7]. This study and others were included in a meta-analysis from Howard et al. in which 102 independent GVs were associated with MD[8]. More recently, a GWAS on >1.2 million individuals, including data from the Million Veteran Program, reported 178 MD genomic risk loci[76]. Nevertheless, the GVs from this study, the largest GWAS on MD, explained only 11.2% of the MD GV heritability[76]. GVs with minor effects and rare variants are expected to explain the missing heritability once whole-genome sequencing becomes more affordable and sample sizes increase significantly[69]. It is estimated that millions of individuals would be needed to detect such small effects (i.e., odds ratio ≤ 1.1), given that the relationship between effect size and the sample size is non-linear among common GVs. Interestingly, CGS findings do not overlap with current GWAS results on MD[55,77].

The interpretation of GWAS findings, though, is challenging. It involves the identification of the gene or regulatory mechanism being

affected by GWAS SNPs, followed by the determination of their biological function and how this impacts the disease. Therefore, additional analysis and data are required to shed light on how the identified SNPs ultimately lead to the disease phenotype.

1.2.3. Preclinical studies

Preclinical models have been used for many years to study the aetiology and the pathophysiological mechanisms of MD as well as the effects of antidepressants. Among preclinical models, we can find cellular and animal models that mimic features of MD[78]. Neuronal cells are the most widely employed cellular model due to the effects of stress on neurons and neuronal brain networks, with a particular emphasis on cultures from the cortex and hippocampus. Additionally, glial cell cultures have been used because of their importance in maintaining synaptic connections and supplying nutrients to neurons. It has also been reported that the abnormal functioning of glial cells may contribute to MD. In recent years, patient-derived induced pluripotent stem cells have become popular as a strategy to gain insights into MD mechanisms and drug research[79]. In general, cellular models are combined with other studies to identify cell-specific mechanisms or confirm observations from human subjects and/or animal models. Although the physiological relevance of cellular models has been debated, when an antidepressant's mechanism of action is known, such models can be helpful as a first screening platform to identify toxic or side effects[78].

Animal models must show a similar disease phenotype, pharmacological sensitivity and pathophysiological mechanisms to be useful for studying neuropsychiatric disorders[80]. The latter may differ from human MD because multiple factors trigger the disease, and the exact mechanisms are not fully understood. In addition, rodents exhibit what's called a "depressive-like behaviour", which is determined by behavioural observation. A variety of tests have been designed to assess features of depressive-like behaviour, such as despair by creating an unescapable situation or anhedonia by evaluating the pleasure of previously pleasant things[81]. Animal models for MD are based on environmental, genetic, or pharmacological influences[82]. Environmentally induced models are based on the dysregulation of stress hormones in MD by applying stress factors at different time points (e.g., maternal separation in early life or chronic social defeat stress exposure in adulthood)[82]. Genetically based models can either arise from selective breeding or genetic manipulation via knockout or transgenesis[81,82]. Pharmacological models use a drug or

treatment to induce a depressive-like behaviour, although these may produce other alterations not present in human MD[81,82]. Additionally, the surgical ablation of the olfactory bulbs in rodents results in a chronic psychomotor agitated MD with significant cognitive impairment, mimicking a limited number of MD cases [81,82].

These models have contributed to a better understanding of the abnormal functioning of brain circuits in MD via identifying cellular and molecular changes associated with MD[81]. Although there is not a single animal model that perfectly replicates the complexity and heterogeneity of human MD in all its aspects (e.g., aetiology and treatment response), these mimic most features[81,82]. Additionally, it is common practice the use of several animal models to combine strengths and address weaknesses in order to gain insights into various disease pathogenesis aspects[82].

1.3. Assessing the validity of genetic associations for complex diseases

Almost 600 GVs have been identified as significantly associated with MD thanks to GWAS[83]. Despite such advances, few studies have sought to validate GWAS findings[84]. It has been argued that rather than more discovery studies, there is a critical need for their functional follow-up to better understand their role in disease pathogenesis[85]. Understanding the biological mechanisms by which GVs influence disease phenotype is critical to promote better diagnosis and treatment[85]. The challenge in complex diseases in general, and MD in particular, is that most of the identified GVs are common, lie in non-coding regions of the genome and lack mechanistic characterisation on how they are involved in disease pathogenesis. In order to leverage the use of genetic associations to support drug target identification or precision medicine applications, the validity of the association between the GVs and the disease must be assessed.

Several curation consortia have been created to support a standard procedure for assessing the validity of gene-disease and GV-disease associations. These curation criteria may serve different purposes: development of knowledge bases, drug research and discovery (e.g., identifying drug targets or developing new treatments), or clinical genomics (e.g., diagnostic gene panels, functional interpretation or genetic counselling); where different levels of evidence to assess the

association validity would be required accordingly. In a clinical setting, for instance, the validity criteria may differ depending on whether we are conducting a predictive test for a healthy individual, a diagnostic test for a disease patient, or a treatment response test[86].

Initiatives like the Gene Curation Coalition (GenCC) include several organisations from around the world that collect gene-disease association data with a focus on highly penetrant monogenic forms of disease, that is Mendelian and rare diseases[87]. GenCC has developed a unified validity system that promotes data-sharing and data consistency by utilising standard terminologies and a clinical validity classification system. The harmonised validity terms for gene-disease associations include: Definitive (repeatedly reported in both research and clinical diagnostic contexts, upheld over time and with no convincing contradictory evidence - the highest validity level); Strong (very similar to definitive), Moderate (lacks a large body of evidence), Limited (little evidence, where not all has been disproven- could be false positives), Dispute (equally weighted evidence supporting and refuting association), Animal Model Only (very little or inexisting evidence in humans but convincing evidence in animal models), Refuted Evidence (existence of association but with new evidence refuting it), and No Known Disease Relationship (no claim has been ever made).

Among GenCC organisations, the Clinical Genome Resource (ClinGen) focuses on clinically relevant genes and variants across Mendelian diseases[88–90]. Regarding gene-disease associations, ClinGen performs a semi-quantitative assessment of the strength of genetic and experimental evidence. It reviews molecular mechanisms, phenotypic variability and mode of inheritance to support or refute the existence of an association[89]. ClinGen gene-disease validity curation follows the GenCC classification system. As for variant-association data, ClinGen Variant Curation Interface (VCI) supports germline variant classification according to the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) guidelines[90,91]: pathogenic, likely pathogenic, uncertain significance, likely benign and benign. The focus is GV classification considering the variant data model, which involves population, experimental and computational data, as well as disease and mode of inheritance data models. Estimated variant pathogenicity is provided by biocurators' evaluations of the evidence criteria. The Online Mendelian Inheritance in Man (OMIM) is also under the GenCC umbrella. OMIM

platform collects genes and genetic phenotype association data, curating evidence from the literature[92]. Association data is extensively linked to other genomic resources and additional references to facilitate posterior annotation and analytical efforts.

Regarding rare diseases, Orphanet has developed a procedure for the selection, quality evaluation and dissemination of clinical practice guidelines[93]. Orphanet aims to assist in healthcare decisions by providing accurate and specific recommendations to doctors and patients. Their quality criteria are based on the Appraisal of Guidelines, REsearch and Evaluation (AGREE II) Instrument, which curates clinical practice guidelines examining scope and purpose, clarity and applicability, among other features[93]. In addition, Genomics England PanelApp has compiled a list of clinically and scientifically validated genes and variants with clear evidence of disease causation, referred to as gene panels[94]. It is based on integrating knowledge from diverse panels to reach a consensus. It uses a traffic-light system: green (3-4 sources - highest confidence level); amber (2 sources); or red (1 source - lower confidence level).

These curation guidelines are intended for gene-disease and GV-disease validity assessment in Mendelian and rare diseases. Therefore, they cannot be applied to diseases with different genetic architectures, such as complex diseases (Figure 1). Complex diseases involve multiple common GVs, where natural selection has favoured the removal of GVs with large effects, becoming rare and thus, making common GVs extremely unlikely to have large effects on disease[95]. A further limitation in gene and GV-disease validity assessment in complex diseases is the functional interpretation of the GVs' role in disease pathogenesis. Most of the associated GVs lie in non-coding regions of the genome with no clear target gene or gene expression regulatory mechanism. The greatest effects can be expected from GVs in or near protein-coding regions, affecting all or most cell types[85]. In comparison, GVs in non-coding regions are subject to weaker selection and may affect gene regulation in a cell-type-specific manner[96]. Many common GVs have been identified thanks to GWAS, but further research into their biological mechanisms is required to assess their role in disease pathogenesis.

There is a lack of curation guidelines for GV-disease associations that consider multiple evidences and study types in complex diseases, except for cancer. Available resources for cancer, which is primarily caused by

somatic GVs, include Clinical Interpretation of Variants in Cancer[97] or Cancer Genome Interpreter[98]. Both annotate GVs with clinical relevance and their effect on treatment response, promoting data normalisation and interoperability. As for other complex diseases in general, several publicly available repositories focus on GV-disease associations, particularly from GWAS. The databases GWAS Catalog, GWASdb or GWAS Central all collect and curate genetic association data from GWAS studies[83,99,100]. ClinVar accepts third-party submissions on GVs and their clinical significance in disease. Still, while GVs associated with Mendelian disorders should follow ACMG/AMP and ClinGen recommendations (e.g., benign or pathogenic), GWAS data is reported under "association"[101]. When focusing on psychiatric disorders, Psychiatric disorders Gene association NETwork (PsyGeNET) has developed curation guidelines for gene-disease associations extracted from the literature by text mining (TM). These guidelines evaluate the association of a gene to a disease per publication and annotate them using standard terminologies[102].

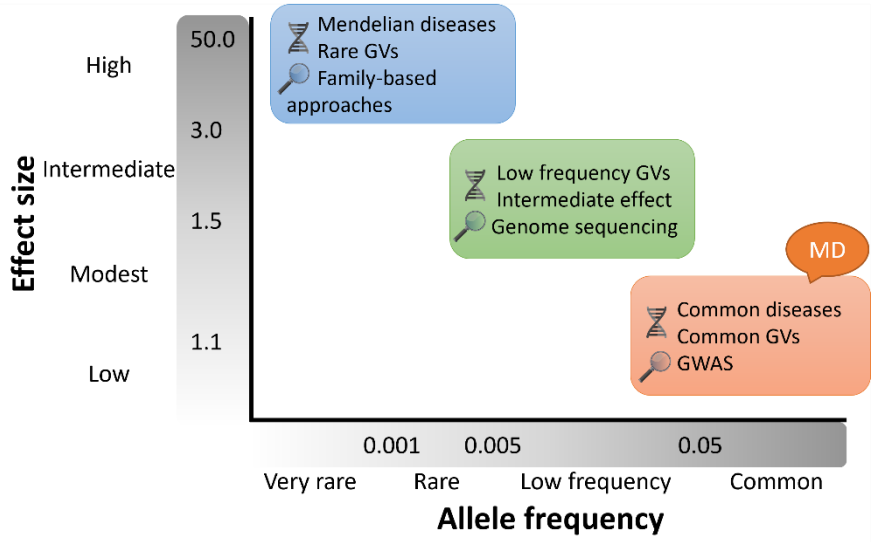


Figure 1. GVs and disease severity. The relationship between the GVs' effect size and severity, as well as the typical study type. Adapted from Manolio, T. A. et al. Finding the missing heritability of complex diseases. Nature 2009 461:7265 461, 747–753 (2009).

Overall, there is a clear need of guidelines for assessing gene and variant-disease associations in complex diseases to support downstream applications. These would help determine the GVs' role in disease pathogenesis and eventually translate this knowledge for its use in

clinical applications, precision medicine, and research. These guidelines should consider current experimental methodologies used to uncover gene and GV-disease associations in complex diseases; for instance, GWAS involving SNP arrays or WGS, post-GWAS functional analysis using different omics, as well as animal and cellular models. These should also include recommendations for evaluating evidence obtained by previous approaches, such as candidate gene studies.

1.4. From genome association to disease mechanisms

The main goal of genetic association studies in MD is to understand disease biology, support the identification of disease biomarkers and propose new therapeutic strategies. Instead of encouraging larger GWAS, which would produce GVs with smaller and smaller impacts, it has been suggested that the focus should be on downstream functional analysis of significant well-replicated GWAS GVs[96,104]. Genetic association studies on MD face several challenges in making biological inferences on causal mechanisms and prioritising GVs and target genes. First, multiple GVs are involved in disease pathogenesis, which may also interplay. Because of LD, identified associated GVs may not be the causal ones (i.e., influencing disease risk) but point to regions in the genome that are involved in disease pathogenesis. Furthermore, these GVs are usually in non-coding regions of the genome, and their target genes or regulatory mechanisms cannot always be straightforwardly determined. Coding GVs may have no effect or directly disrupt a protein function. In contrast, non-coding GVs are thought to influence gene function through diverse regulatory mechanisms, such as altering gene expression regulation by affecting promoter and enhancer activity or disrupting transcription factor (TF) binding sites.

To overcome these challenges, different strategies for performing an accurate functional analysis have been developed, which may shed some light on the role of disease-associated GVs (Figure 2). To that end, the availability of summary statistics of genetic studies (i.e., the list of genome-wide significant GVs and their effect size) is key to promoting open-access research that allows follow-up studies. Additionally, genomic annotation data such as regulatory elements or chromatin states' is required to assess these GVs' role in gene expression regulation. As a result, integrating these GVs with genomics data is critical for identifying their target genes and ultimately deciphering the underlying regulatory mechanisms. Currently, different bioinformatics

analyses have been developed to address this challenge and prioritise causal GVs and target genes.

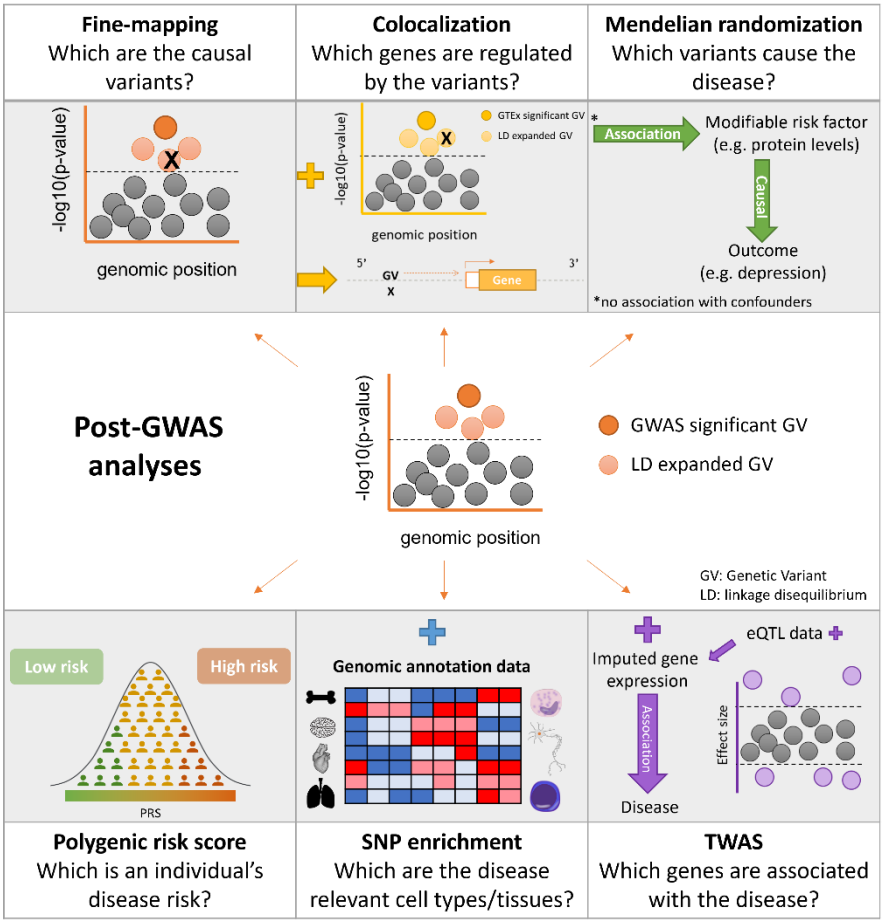


Figure 2. Summary of post-GWAS analyses. Starting with GWAS findings, an overview of post-GWAS analyses incorporating diverse genomic association data and conducting different statistical analyses to answer multiple questions.

Fine-mapping

The identified GWAS SNPs are not necessarily the causal ones but rather tag genomic regions of interest. SNPs in microarrays, also known as tag SNPs, are in high LD with neighbouring SNPs, serving as surrogates for large genomic regions that contain unmeasured SNPs[105]. Therefore, the most-significant SNPs or lead SNPs (p-value < 5×10^{-8}) may be in high LD with the true causal SNP[60]. Fine-mapping aims to identify the GVs that have a biological effect (i.e., are causal) given association data and the assumption that at least one causal GV exists[106].

Fine-mapping is limited by the experimental sample size, the LD structure, the number of causal SNPs per region, their effect sizes, and SNP density[106]. The former can be increased by pooling different studies or performing meta-analyses. As for the LD structure, patterns among SNPs can be complex and are influenced by recombination, mutation rates, natural selection, population subdivision and bottlenecks[107]. For instance, small LD blocks in certain populations, such as Africans, may reduce the number of candidate causal GVs and lead to statistical heterogeneity around lead SNPs[108]. Another controllable factor critical to capture causal GVs is SNP density, which can be increased by DNA sequencing, genotype imputation or additional genotyping[106].

Identifying causal GVs by fine-mapping approaches is done by integrating GWAS results with LD reference patterns. Heuristic fine-mapping considers pairwise correlation data to either retain potentially causal SNPs based on a set threshold or perform hierarchical clustering. However, caution should be taken because these methods do not account for SNPs' joint effects and do not provide an objective measure of the confidence that an SNP is causal but rely on arbitrary thresholds[106].

Bayesian methods consider predefined windows of SNPs and compute the posterior probability of a hypothesis or model conditional on observed data (trait and SNPs) and assumed prior distribution. Prior probabilities assumptions can be varied by treating GVs as independent and either likely to be causal or as a fixed number[109]. The resulting posterior probability can be used to compute each SNP's posterior inclusion probability (PIP), that is the sum of posteriors over different models that consider the SNP a causal GV. Alternatively, by PIP ranking and calculating the cumulative sum, posterior probabilities can be used to determine credible sets, that is, sets that capture likely causal SNPs. In this sense, Bayesian methods are advantageous because SNPs probabilities can be directly compared, unlike p-values. Furthermore, because these models are based on the combined effect of SNPs, they control for SNPs with high effects while increasing the power to detect those with small effects[106].

Colocalisation analysis

The integration of GWAS and functional annotations can shed some light on the underlying biological mechanisms of these GVs, which may

exert phenotypic effects through different molecular mechanisms. Multiple molecular quantitative traits loci (QTLs) datasets can be leveraged to evaluate the role of GVs in gene expression (eQTLs), protein expression (pQTLs), exon splicing (sQTLs), DNA methylation (mQTLs), and chromatin accessibility (caQTLs). These datasets on molecular quantitative traits associated with GVs (e.g., eQTL), along with GWAS GVs, can be used in colocalisation analysis to assess whether a GV is both associated with a disease phenotype and a molecular trait.

For instance, the application of colocalisation methods on GWAS and eQTL data enables the identification of genes whose expression is affected by the GVs associated with the disease phenotype. Indeed, SNPs associated with complex traits are more likely to be eQTL[110]. Nonetheless, overlapping eQTL and GWAS signals can be due to: 1) linkage, where two independent causal SNPs are in LD; 2) causality, where a single causal SNP modifies the expression of a gene to influence the trait; or 3) pleiotropy, where a single causal SNP has independent effects on gene expression and the trait. To correctly interpret GWAS results, it is essential to distinguish between these settings. Furthermore, because one cis eQTL has been identified for almost all known human genes, these overlaps are likely to occur by chance[111]; some colocalisation methods consider these probabilities while others do not.

Mendelian randomisation

Mendelian randomisation, or MR, is another approach to determine whether a single causal SNP influences gene expression and a disease or trait. MR uses GVs as instrumental variables to infer causality and exclude pleiotropy by considering the relationship between modifiable risk factors or exposures (e.g., protein levels) and an outcome (e.g., depression)[112]. MR is based on the following assumptions: 1) GVs are significantly associated with the risk factors; 2) GVs must not be associated with confounders of exposure and outcome association; and 3) GVs impact the outcome only by influencing the exposure and not by any other pathway. It is not easy to demonstrate the latter two assumptions, and new methods have been developed that are less reliant on them. MR results should be accompanied by other sources of evidence for their correct interpretation for clinical decisions[113]. For instance, whether the risk is cumulative or has an acute effect or whether there are more relevant lifetime periods.

Polygenic risk score

Polygenic risk scores (PRS) predict an individual's disease risk. It is computed based on the number of risk alleles in an individual's genotype multiplied by the GVs' estimated effects (i.e., GWAS odds ratio multiplied by its direction[104]. The association with the disease is tested in a linear (continuous trait) or logistic (binary trait) regression adjusting for covariates (e.g., sex or age). PRS can provide insights into how GVs affect disease subtypes as well as determine the genetic overlap between disorders or stratify individuals for a more effective clinical intervention[13]. However, PRS are not yet suitable for healthy individuals but rather for distinguishing risk groups[13]. The frequently low predictive power of PRS has led to some scepticism about its clinical utility[114]. However, it is expected that PRS could benefit from larger GWAS sample sizes as well as more accurate clinical phenotyping[44,114].

SNP enrichment

SNP enrichment techniques aim to prioritise disease-relevant cell types, accelerating the challenging functional validation of GWAS GVs. The premise underlying SNP enrichment techniques is that GWAS GVs are overrepresented in genomic areas that are particularly active in the pathogenic cell types[104]. These approaches combine GWAS and genomic annotation data (e.g., cell type-specific gene expression or chromatin annotations) to identify cell types with associated GVs that overlap annotations more frequently than would be anticipated by chance. However, GWAS GVs are typically found in genomic regions with a high gene density and, therefore, a higher density of chromatin regulatory elements, which can confound enrichment estimates if not considered[104]. Additionally, when considering gene expression data, there is a required balance between failing to account for multiple causal genes and including many genes not relevant to the disease[115]. It is expected that when additional expression and chromatin data for more cell types and states become available, enrichment estimates will get more precise in identifying cell types causally involved in disease.

TWAS

Transcriptome-wide association studies (TWAS) are a gene-based strategy that examines the association between gene expression genetically regulated by GVs and disease risk[116,117]. TWAS first combine genotype data with the regulatory effects of eQTLs to impute genetically regulated gene expression levels[116]. Then, they examine the

association between the imputed expression levels and disease risk. The result is an interpretable transcription hypothesis between a gene and a disease. TWAS have some advantages over other variant-based approaches like colocalisation with eQTLs, such as providing a functional understanding of disease mechanisms, independent steps for predicting gene expression levels and its association with disease, and reduced multiple testing[116]. However, results may require further processing since TWAS signals may not all be independent and biologically relevant[117].

1.4.1. Integration with genomic annotation data

GVs can be further assessed and prioritised according to their genetic regulatory functions by leveraging publicly available genomic annotations to arrange posterior more expensive and time-consuming functional laboratory research. Disease-associated GVs are enriched in cis-regulatory elements (CREs), where cis means that the effects are caused by GVs in the same DNA molecule as the target gene. Thus, these GVs are likely to influence disease risk by altering the genetic regulation of one or more target genes[118]. Among the approaches for assessing these GVs' regulatory activity, these can be overlapped with accessible chromatin regions, TF binding sites, or histone marks. The latter can be used to determine the type or regulatory element (e.g., promoter or enhancer) and advise posterior functional assays.

Several publicly available resources characterise epigenetic marks, for instance, the Encyclopedia of DNA Elements (ENCODE)[119], NIH Roadmap Epigenomics[120] or FANTOM[121]. These include CREs annotations and chromatin states in hundreds of cell types and tissues. Indeed, tissue type selection is critical when studying complex diseases since several tissues might be dysregulated, and gene expression varies across tissues. The Genotype-Tissue Expression project (GTEx) is one of the most comprehensive eQTL resources, with samples from 54 non-diseased tissues collected from over 1000 people[122]. While cis-eQTLs contribute to expression heritability, the largest amount of variance is estimated to regulate via trans mechanisms (i.e., effects are due to GVs affecting diffusible elements such as TF)[123]. However, identifying trans regulatory mechanisms is much more complex, requiring tens or hundreds of thousands of individuals[111].

Furthermore, GVs within TF binding site (TFBS) play a central role in complex traits in general and major depression in particular[124]. These

GVs change the affinity of TFs, resulting in the creation or disruption of a binding site. Experimental approaches include chromatin immunoprecipitation assay with sequencing (ChIP-Seq) and electrophoretic mobility shift assays (EMSAs). However, ChIP-Seq, identifies regions of 100-1000 base pairs (bp), whereas the actual TFBS is a shorter region. (9-15pbs)[125]. EMSA has some limitations too, because it may identify non-specific binding proteins[126]. Several resources that scan the DNA sequence with a position-specific scoring matrix for a TF of interest have been developed to predict the TFBS [125]. Common approaches are based on pattern-matching and machine learning. The latter integrates functional annotation, epigenomics and transcriptomics data.

Although there are many GVs that overlap genomic regulatory regions, not all of them are necessarily functionally relevant. The advancement of bioinformatics has encouraged the development of functional prediction algorithms, which can also guide GV prioritisation[127,128]. These algorithms predict the likelihood of a GV influencing regulatory functions and causing disease using sequencing data as well as evolutionary and genomic annotation data. Generally, prediction validation is done by comparing against datasets that include both true pathogenic and non-pathogenic GVs. Prediction results are typically in the form of scores, which quantify how likely these GVs are to be deleterious or pathogenic. These results can be used to increase the power of posterior analyses. Still, available resources may use different data, each with its own set of strengths and weaknesses, resulting in a variety of outcomes. As a result, it has been proposed to reach a consensus by combining the findings of various algorithms to produce a more accurate prediction[129].

Focusing on non-coding GVs, popular tools are the Combined Annotation Dependent Depletion (CADD)[127] and the Deleterious Annotation of genetic variants using Neural Networks tool (DANN)[130]. CADD considers annotations mostly coming from ENCODE and computes a C-score via a machine-learning model that uses annotated and simulated GVs to measure their effects. On the other hand, DANN employs a Deep Neural Network algorithm capable of capturing linear and non-linear relationships between annotations.

Overall, the limitations of genomic annotation data integration are inherent in existing epigenomic and eQTL data, which are context and

tissue-specific, making some cases more challenging to assess. Nonetheless, the real bottleneck is the lack of disease-focused functional biological studies downstream of GWAS to aid in our understanding of the trait.

1.4.2. Experimental evaluation of GVs functions

Once GVs have been prioritised, the focus shifts to evaluating their function. Among the most common experimental approaches, we highlight cell culture-based reporter assay and genome editing[131]. Cell culture-based reporter assays allow the comparison of reference and alternative alleles when cloned to reporter genes and transfected into relevant cell types. Additionally, thousands of GVs can be tested using massively parallel reporter assays (MPRAs), a practical approach since several GVs in LD may impact multiple enhancers and cooperatively affect gene expression[132,133]. Limitations may be found when the context is inappropriate (e.g., not relevant cell type or environmental conditions) or there is transcriptional noise, resulting in false positive results. Finally, genome editing is a more physiologically relevant method that allows for specific changes not only in the DNA sequence but also in the epigenetic state. Available techniques are zinc finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs) and clustered regularly interspaced short palindromic repeat (CRISPR)-based systems[134].

The physical interaction of transcriptional elements (promoters, enhancers, and distal regulatory elements) may prove the potential regulatory function of these GVs and how these influence their target genes[135]. High-throughput sequencing and 3D chromosome structure have shed some light on the genome's long-distance contacts even across time[136]. However, because these technologies are still in the early stages of development, validation and annotation rules are still required[137].

1.4.3. From gene regulatory mechanisms to disease

But how do these GVs and genes ultimately impact disease phenotype? Despite the vast amount of genomic annotation data available for identifying regulatory GVs and understanding their functions, little is known about how these affect disease risk. Approaches considering expression differences (for example, overexpression or knockout) are hampered by a lack of eQTL effect sizes and technical difficulties in

addressing these expression changes. Correlation studies between gene expression and trait have been proposed to aid in this matter. Genome editing, on the other hand, is a better approach because it allows a more physiologically relevant assessment and causal association. Moreover, when molecular, cellular, and organismal phenotypes are tailored to each disease, disease-specific knowledge will be obtained[131].

In the meantime, biological networks and gene ontologies may shed some light on prioritised genes and their role in complex diseases. Instead of being the result of one gene's effect, complex diseases are frequently caused by gene interaction. Network approaches are based on biomolecular knowledge and gene-based association test that can capture biological interactions between different molecules. These networks could reveal pathways of interest for disease phenotype by highlighting regulatory, metabolic and signalling processes[138]. Gene ontologies can also provide information about relevant biological processes, molecular functions, and cellular components.

2. OBJECTIVES

MD poses a significant burden on society, mainly due to our lack of understanding of its pathogenesis involving the interaction between genetics and environmental factors. Several hypotheses about its origin have been proposed, and GVs significantly associated with MD have been identified. However, these GVs usually lie in non-coding regions of the genome, and functional analyses are required to uncover their role in disease pathogenesis. Thus, the primary goal of this thesis is to gain insights into the role that GVs play in the pathogenesis of MD by applying bioinformatic approaches and leveraging publicly available genetic association data and post-GWAS functional analysis resources.

The following objectives were established to achieve our goal:

1. To develop curation guidelines for evaluating the quality of MD genetic association data.
2. To create a database of GVs associated with MD following the developed guidelines.
3. To uncover potential regulatory mechanisms by which GVs associated with MD may contribute to disease pathogenesis using GWAS summary statistics.
4. To benchmark post-GWAS analysis tools by systematically evaluating their performance using full genome summary statistics and selecting the most suitable ones for the interpretation of GWAS findings on MD.

The first objective was addressed by evaluating existing curation criteria for other diseases as well as considering the genetic architecture of MD and types of genetic association studies performed (**Chapter 3.1**). The second objective was achieved by applying the developed guidelines on publicly available genetic association data on MD scattered throughout the literature and databases (**Chapter 3.1**). The third goal was accomplished by conducting diverse functional analyses, including fine-mapping, colocalisation, and transcription factor binding site analysis, to determine the role of GVs in MD (**Chapters 3.2 and 3.3**). Finally, the fourth goal was addressed by developing and implementing a workflow that compared the outcomes of different post-GWAS analysis tools and their biological implications (**Chapter 3.3**).

3. RESULTS

3.1. Building a data curation pipeline for complex diseases: the case of major depression

The complexity of the genetic architecture of MD fades the huge progress of genetic association studies on MD. Multiple GVs with a minor role have been identified, which mainly lie in non-coding regions of the genome. Compared to Mendelian and rare disorders, these features challenge the development of curation guidelines for clinical and research applications of these data. In this chapter, we review available guidelines and evaluate the diversity of genetic association study types and designs. Then, we develop and apply expert curation guidelines for genetic association data, focusing on the specific case of MD. Finally, we functionally analyse these data and rank GVs according to supporting evidence.

Building a data curation pipeline for complex diseases: the case of Major Depression

Judith Pérez-Granado[1], Janet Piñero[1,2] and Laura I. Furlong[1,2*]

1. Research Programme on Biomedical Informatics (GRIB), Hospital Del Mar Medical Research Institute (IMIM), Department of Medicine and Life Sciences (MELIS), Universitat Pompeu Fabra (UPF), Barcelona, Spain
2. MedBioinformatics Solutions SL, Barcelona, Spain

[*]Correspondence: Laura I. Furlong, laura.furlong@upf.edu

Abstract

Major depression (MD) is the leading cause of impairment worldwide, and despite huge efforts put into understanding its biological underpinnings, these are not yet fully understood. Genetic association studies have made great progress in the last years identifying multiple genetic variants associated with different traits and disease risks. Evaluating gene-disease association is key for enabling the identification of disease biomarkers, assisting patients' diagnosis, or supporting drug development efforts. While curation criteria for assessing the validity of gene-disease associations for Mendelian and rare disorders have been established, there is a need for comparable assessment criteria for complex diseases. The fact that multiple genetic variants influence disease risk, which mainly lie in non-coding regions with unclear target genes or regulatory mechanisms, challenges complex diseases' understanding. Here, we review existing protocols and propose curation guidelines tailored to MD where we consider different types of experimental evidence and their study design. The result of their application is an expert-curated database of genetic variants associated with MD collected from various repositories and the literature. We then conducted a functional enrichment analysis to unravel their potential role in MD pathogenesis. The proposed curation guidelines could be applied to other diseases with similar genetic architecture.

Keywords: major depression, curation, variant-disease association, GWAS, candidate gene studies.

1. Introduction

Major Depression (MD) is a highly frequent mental disorder and the leading cause of disability worldwide[1,2]. It is caused by the interaction of multiple genetic and environmental factors and is characterised by a wide range of clinical traits and features such as sleep disturbance or thoughts of guilt[2]. Despite the numerous genetic association studies conducted on MD, it is still unclear how identified genetic variants (GVs) affect disease risk. Twin studies estimate that GVs can explain almost 40% of MD[3]. However, no biomarker has been identified to aid in diagnosis or treatment, with pharmacological treatments being ineffective in 40% of patients[2,4]. A better understanding of both MD genetic architecture and the role that GVs play in disease pathogenesis will lead to new developments to improve the care of patients. Different databases collect GVs associated with MD[5,6], and multiple methods have been developed for their functional analysis[7–9]. However, no curation guidelines that combine multiple evidences from diverse study types (e.g., human and animal models) exist to validate genetic association data on MD or complex diseases in general.

Several curation consortiums have been established in the last few years to promote systematic methods for evaluating the validity of gene-disease and GV-disease associations. Their purposes can range from creating a knowledge resource to supporting pharmacological research and clinical genomics, all of which call for evidence of varying strengths. For instance, the Gene Curation Coalition (GenCC) focuses on Mendelian and rare diseases and has developed a standard clinical validity classification system to promote interoperability[10]. GenCC considers evidences from diverse study types as well as contradictory evidences to assign association data to the established categories. Regarding Mendelian diseases, it comprises the Clinical Genome Resource (GlinGen) and the Online Mendelian Inheritance in Man (OMIM), among others. ClinGen focuses on clinically relevant genes and GVs and performs both gene-disease and GV-disease association validity[11–13]. The former is a semi-quantitative evaluation of the weight of genetic and experimental data[11] and the latter considers data models for the variant, disease and mode of inheritance[13]. OMIM curates gene and phenotype association data from the literature and cross-references it with external resources to aid annotation and analytical efforts[14]. As for rare diseases, Orphanet focuses on clinical practice guidelines, selecting, qualitatively evaluating, and disseminating them to assist in healthcare decisions[15]. Furthermore, the Genomics England PanelApp

has compiled and integrated a consensus list of clinically and scientifically validated genes and variants with evidence of disease causation, known as gene panels[16].

To the best of our knowledge, there are no well-established curation protocols to validate GV-disease association data for complex diseases that consider different evidences and study types (e.g., human and animal models). The curation guidelines presented above are intended for highly penetrant monogenetic disorders. These are characterised by a small number of altered genes, with GVs having large effects on disease and mostly lying in coding regions of the genome. In contrast, complex diseases involve multiple common GVs, each with a small effect on disease risk[17]. Additionally, most GVs identified as associated with complex diseases lie in non-coding regions of the genome, challenging the interpretation of the GVs' impact on gene function and, ultimately, disease pathogenesis.

Cancer could be the exception to the absence of this type of curation protocols for complex disease with guidelines to annotate clinically relevant GVs with an effect on treatment response[18,19]. Another example of available guidelines for complex disorders are the ones developed for PsyGeNET (Psychiatric disorders Gene association NETwork)[20]. This resource offers curated gene-disease association data, which has been automatically extracted from the literature using text mining (TM) by annotating it to standard terminologies and expert-reviewing the existence of an association in text snippets.

The evolution of genetic association studies and technological advancements has promoted the identification of GVs associated with MD via candidate gene studies (CGS) and genome-wide association studies (GWAS). CGS have been running for more than two decades and analyse genes hypothesised to be involved in MD neurobiology. These are based-on proposed theories, such as the monoamine or serotonergic theories[2,21]. However, due to the polygenic nature of MD and the small sample size of most CGS, their findings have been criticised for their lack of replicability[22,23]. GWAS on MD were introduced in the mid-2000 and consist of arrays of hundreds of thousands of GVs. Thanks to international research consortia and an increase in sample size, common GVs significantly associated with MD have been identified[24,25]. However, these GVs' target genes or the regulatory mechanisms they influence cannot always be

straightforwardly determined, requiring a comprehensive genomic analysis to understand how they influence disease risk[26].

Several gene-environment (GxE) interaction and treatment response (TR) studies on MD have been conducted[27,28]. Both approaches have been primarily based on CGS, missing the complexity of MD and conducted on small samples, which has resulted in inconsistent and unreplicable outcomes[23,29,30]. It is expected that as genome-wide environment interaction studies approaches are performed for MD, more reliable findings will arise[29]. Similarly, pharmacogenomic studies on MD are expected to boost with the increment of sample size and the analysis of more homogeneous groups[30,31].

In addition to approaches to identifying the contribution of genetic variability to MD risk, a variety of cellular and animal models have been proposed to shed light on different aspects of MD pathogenesis. For instance, animal models for MD allow researchers to study "depression-like" behaviour in genetically modified and environmentally or pharmacologically influenced rodents[32,33]. These models, however, are best suited to assessing the role of coding GVs with a clear effect on gene function[34]. Currently, there is no one-size-fits-all model for MD, and while symptoms like anhedonia and social interaction can be modelled, others like guilt and suicidal thoughts cannot be captured in such models[32,33]. As for *in vitro* MD modelling, it has progressed from tumour-derived cells to immortalised cell lines and patient-derived neural cells[35]. These are adequate pre-screening methods to test the toxic and side effects of antidepressants with known mechanisms of action[36]. The challenge is to capture both environmental and genetic variability effects, as well as to determine cell type relevance[35].

Here, we propose a set of curation guidelines to evalute GVs associated with MD taking into account evidences from different types of experimental approaches. We built an expert-curated database of GVs associated with MD, and characterised the MD-associated GVs in terms of genomic and functional features. Finally, we made the database available to the research community.

2. Methods

2.1. Data collection

We obtained a list of terms representing different types of major depression (MD) (e.g., major depression, single episode) and their phenotypes (e.g., anhedonia) from PsyGeNET[20]. Then, using the Unified Medical Language System (UMLS v.2021AA)[37], we expanded this list by searching for semantically similar terms that an expert ultimately curated.

We used this list to collect genetic variants (GVs) associated with MD from different resources. We collected these data from genome-wide association studies (GWAS) publicly available from the repositories GWAS Catalog and GWASdb (Supplementary Table S1), as well as from the scientific using text-mined associations from DisGeNET. A publication reference (e.g., a PMID identifier or equivalent document identifiers) was required for association data to allow its evaluation in the publication context. Likewise, a dbSNP identifier for the GVs was required to facilitate their subsequent functional analysis.

2.2. Data analysis and validation

2.2.1. Functional analysis

We identified the genes that overlapped or were nearby the curated set of GVs using the Variant Effect Predictor tool through the SNPNexus platform[38]. Then, we performed a functional analysis of the genes using the R package gprofiler2[39]. It integrates several resources and annotates significantly enriched biological and cellular processes, molecular functions, pathways, miRNAs and phenotypic features. We applied a term size filter to the resulting terms (terms with <1500 genes) to eliminate more general terms.

2.2.2. Data scoring

We have developed a score to rank the GVs associated with MD considering the study type and the number of publications reporting the association. The score (S) ranges from 0-1 and it is computed using the following formula.

$$S = ST + P + AT$$

where ST is study type, P is publication and AT is association type

$$ST = \begin{cases} 0.25 & \text{if } N_{ST} = 1 \\ 0.50 & \text{if } N_{ST} = 2 \\ 0.75 & \text{if } N_{ST} = 3 \end{cases}$$

where N_{ST} is the number of study types (i.e., GWAS, CGS or preclinical models) supporting the association

$$P = \begin{cases} 0.1 & \text{if } N_{PUBS} > 1 \\ 0 & \text{Otherwise} \end{cases}$$

where N_{PUBS} is the number of publications supporting the association

$$AT = \begin{cases} 0.15 & \text{if } N_{AT} > 1 \\ 0 & \text{Otherwise} \end{cases}$$

where N_{AT} is the number of association types (i.e., VDA, TR, E) supporting the association

3. Results

We have developed curation guidelines for assessing the quality of genetic association data from diverse study types, focusing on the particular case of MD. By following these guidelines, we analysed over 2000 publications and obtained a curated dataset of 709 GVs associated with MD. We present the analysis of these GVs in terms of the evidences that support them, as well as their genomic and functional implications. The resulting database is made available for the research community to support future applications in precision medicine for MD.

3.1. Data collection

We recovered 37 major depression or MD-related terms from PsyGeNET, which became 104 terms after semantic expansion. This list was expert-curated, with 96 terms kept to retrieve GVs associated with MD from various resources (Supplementary Table S2).

A thousand and fourteen GVs were collected from 2026 publications (DisGeNET: 1651 GVs from 994 publications, GWASdb: 72 GVs from 22 publications; GWAS Catalog: 873 GVS from 39 publications). These led to 2911 GV-publication pairs recovered from the literature and genome-wide association studies (GWAS) repositories that have undergone the data curation.

3.2. Data curation

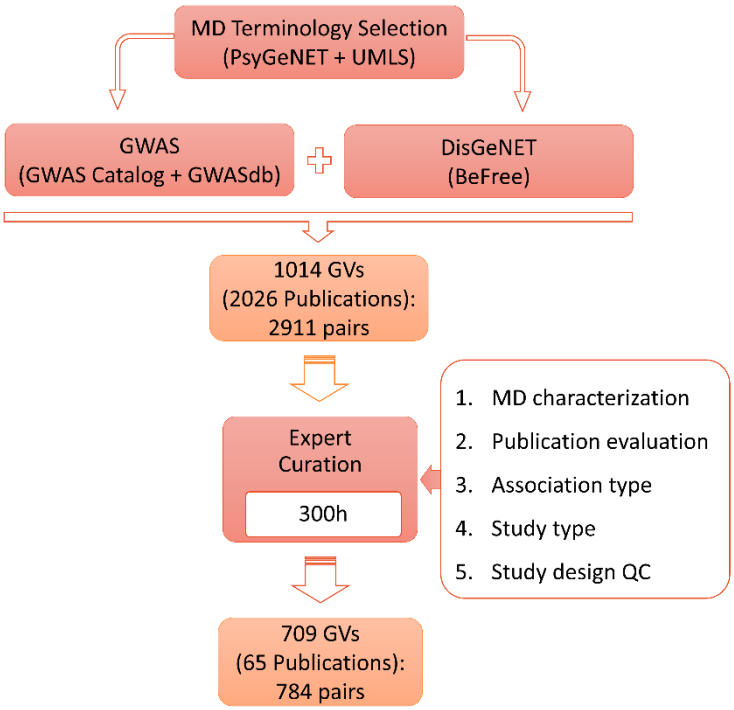


Figure 1. Implemented pipeline and guidelines. Schema of the pipeline followed and the developed guidelines. MD: major depression; UMLS: Unified Medical Language System; GWAS: genome-wide association studies; GV: genetic variants; QC: quality control.

3.2.1. Guidelines

1. Characterisation of MD context

MD is a complex disease in which genetic, social and environmental factors play an important role. It is characterised by a loss of interest and symptoms such as sleep difficulties, low energy or fatigue, appetite changes, feelings of worthlessness and even suicidality[3]. MD is also a common comorbidity in chronic diseases such as cancer and neurological disorders[40]. Therefore, the association of GVs with MD might be studied in the context of the assessment of environmental factors or other comorbid conditions[2]. In addition, the mention of MD does not always imply that it is being studied or involved in the association, but it could be background information for another disease setting. Thus, all captured associations were reviewed for their true association with MD and classified according to their context (Table 2 and Supplementary Table S3). Regular expressions were used to capture and tag evidences containing words such as "bipolar", "schizo", "life", "suicide" or "environment". Nevertheless, all of them were individually reviewed by an expert.

MD	The R allele of PON1 Q192R was associated with depression: per-allele odds ratio 1.22 (95% confidence interval: 1.05 to 1.41) in this population. (PMID: 17183021)
Features/Traits	Relationship between G1287A of the NET Gene Polymorphisms and Brain Volume in Major Depressive Disorder: A Voxel-Based MRI Study. (PMID: 26960194)
Environmental factors (E)	The present study suggests that the combined effect of rs2242446 and rs5569 in the NET gene could modify the response to the negative life events in triggering MD. (PMID: 18779921)
Comorbidities	Association analysis of the 5-HT6 receptor polymorphism C267T with depression in patients with Alzheimer's disease. (PMID: 11442897)
No MD-related	From among this cohort, we studied the chloride currents generated by G190S (associated with pronounced transitory depression), F167L (little or no transitory depression), and A531V (variable transitory

	depression) hClC-1 mutants in transfected HEK293 cells using patch-clamp. (PMID: 23933576)

Table 2. MD context classification system. Sentences showcasing different MD contexts used in the developed classification system. MD: major depression.

The main focus of this research paper is on GVs that are strictly associated with MD. Thus, studies involving comorbidities or evaluating specific features that may be present in MD were not further assessed here but were set aside for future projects.

2. Evaluation of publication

The first step of the publication evaluation was to remove reviews because we would be unable to assess the study quality design in the subsequent curation steps. Then, we checked the dbSNP normalisation process so that the normalised variant matched the reported in the association. Erroneous and multiple normalisations are evaluated and corrected or removed accordingly. Some examples of these types of publications are in Table 3.

Reviews	In this review, we bridge evidence from neuroimaging, behavioural and clinical studies that have examined the role of COMT variants on depression-relevant phenotypes. (PMID: 23792050) [from abstract]
Incorrect variants	However, dimensional analyses showed significant associations of the HADS depression severity scores with Gln460Arg (rs2230912) and Ala348Thr (rs1718119) in the depressed and diabetic patient groups. [identified as: rs755302767] (PMID: 30664971)

Table 3. Publication abstract evaluation. Sentences capturing the type of publication abstract and dbSNP normalisation process assessment.

3. Classification of the association type

There are different types of studies that report GVs associated with MD. We have classified them accordingly: variant-disease association (VDA), treatment response (TR), environmental (E) or combinations (Table 4 and Supplementary Table S3).

Variant-disease association (VDA)	Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. (PMID: 29700475) OR Genetic variants from two previously unreported loci (rs10457592 on 6q16.2 and rs2004910 on 12q24.31) showed significant associations with MDD ($P < 5 \times 10^{-8}$) in a total of 336,753 subjects. (PMID: 29728651)
Treatment response (TR)	Genome-wide pharmacogenetics of antidepressant response in the GENDEP project. (PMID: 20360315)
Environmental (E)	The Val1483Ile polymorphism in the FASN was associated with depressive symptoms under the influence of psychological stress. (PMID: 21641044)

Table 4. Association type classification system. Examples of different types of associations.

4. Characterisation of the study type

VDAs captured by DisGeNET can come from a variety of studies, including preclinical models (including cell and animal models), candidate gene studies (CGS), and GWAS. We classified them accordingly. Note that we removed from DisGeNET the publications already captured by GWAS Catalog or GWASdb because the latter provides the summary statistics from the GWAS (i.e., all GVs significantly associated with MD) as opposed to retrieving only the GVs mentioned in the abstract. Some examples are shown in Table 5.

Preclinical studies- cell culture	We show that 5-HT3AB(Y129S) receptors exhibit a substantially increased maximal response to serotonin compared with WT receptors in two fluorescence-based cellular assays … inversely correlated to the incidence of major depression… (PMID: 18184810)
Candidate gene studies (CGS)	Three SNPs (rs10008257, rs2433320 and rs2452600) were identified in the PDLIM5 gene and genotyped in patients diagnosed with recurrent MDD and in matched control subjects. (PMID: 18197271)

Genome-Wide Association Studies (GWAS)	Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways (PMID: 29662059)

Table 5. Study type characterisation. Examples of different types of studies that were retrieved.

5. Quality control of the study design of CGS and GWAS

One of the big reasons behind the lack of CGS replication can be attributed to the sample size[23,41]. Several articles on GWAS for MD have also been published in that regard, and it is clear that the larger the sample size, the greater the number of significantly identified GVs. Based on different sample size estimations considering study type, GVs and effect size expected to be identified, a sample size cut-off compromise was reached for the particular case of MD[41–43] (Table 6). In addition, we filtered out non-significant associations. The established p-value cut-off was 0.05 for CGS and $5x10^{-8}$ for GWAS to account for the greater number of GVs tested[44]. Note this filter was applied to GWAS repositories at the moment of data collection thanks to its availability via the summary statistics.

Study design	**Case-Control**			**Case-only**	
Study type	**CGS**		**GWAS**	**CGS**	
GV	1	20	500K	1	20
Sample size	1500	3000	3000	1000	2000

Table 6. Sample size cut-off. The minimum sample size number considered for evidence evaluation. This number varies depending on whether the analysis is a case-control or a case-only study, and in the former case, whether it is a CGS or a GWAS. Furthermore, the number of GVs is ultimately considered in each case to set the sample size cut-off. CGS: Candidate gene studies; GWAS: genome-wide association studies; GV: genetic variants.

These guidelines are also available as a standalone document (see Supplementary Material).

3.2.2. Curation results

Following the described guidelines, we conducted an expert curation process. From an initial set of 1014 GVs from 2026 publications, we built a database of 709 GVs associated with MD supported by 65 publications, corresponding to 784 GV-publication pairs (Supplementary Table S4). Table 7 shows the curated GV-publication pairs resulting from each curation step, along with the number of GVs. Additional results can be found in Supplementary Table S5. Note that numbers may not add up because there can be pairs involving combinations of association and study types. The 22% (546) of DisGeNET associations were related to MD comorbidities, for GWAS Catalog was 33.4% (334) and 30.6% (22) for GWASdb. The 50% (1232) of association pairs in DisGENET were VDAs, but 45% (558) and 16% (200) were removed because these did not meet the sample size and significance criteria, respectively. Similarly, 23.6% (17/72) of GWAsdb association pairs were VDAs, and 64.7% (11) did not pass study design QC. In contrast, GWAS Catalog VDAs were 65.3% (653/1000), and only 1.4% (9) did not meet the sample size criteria. As for associations involving TR (317), only 4% (12) of associations from DisGeNET and 17% (1) from GWAS Catalog passed the curation criteria. Regarding those involving environmental factors (188), the associations passing curation criteria corresponded to 12% (21) of DisGeNET and 83% (5) of GWAS Catalog associations.

Regarding the curated GVs (709), there is a surprisingly low overlap between GWAS Catalog and GWASdb (2) and a much higher overlap with GWAS Catalog and DisGeNET (364) (Supplementary Figure S1). Additionally, the overlap across association studies was very small, VDA-TR: rs5569, VDA-E: rs3800373; and TR-E: rs6265. When evaluating the study type of these GVs, 6 were supported by preclinical models, 32 by CGS and 672 by GWAS (Figure 2). Only rs3101339 and rs17759843 were common to more than one study type. Specifically, rs3101339 was reported by Li S et al., who conducted a GWAS and a CRISPR gene editing experiment[45]; while rs17759843 was reported from a transcriptomic and genomic analysis conducted in both humans and mice[46]. Moreover, the different study types reporting GVs associated with MD passing the QC criteria were not far apart in time (2007 for preclinical, 2008 for CGS and 2010 for GWAS) (Figure 3). On the other hand, the mean number of GVs reported per publication varies by study type and ranges from 1 to 117, with a general trend toward an increase over time.

	DisGeNET	GWAS Catalog	GWASdb
Initial dataset	1014 GV (2026 P): 2911 GV-P		
	1651 (994): 2449	873 (39): 1000	72 (22): 72
MD characterisation	1249 (699): 1767	611 (26): 665	49 (12): 49
Publication evaluation	1211 (667): 1680	611 (26): 665	49 (12): 49
Association type			
VDA –	985 (4231): 1232	599 (21): 653	17 (6): 17
TR –	219 (149): 317	6 (2): 6	28 (4): 28
E –	122 (122): 188	6 (3): 6	6 (3): 6
Study design QC			
VDA –	448 (39): 474	590 (17): 644	6 (3): 6
TR –	12 (7): 12	1 (1): 1	
E –	20 (15): 21	5 (2): 5	
Final	477 (60): 506	596 (20): 650	6 (3): 6
	709 (65): 784		

Table 7. Summary of curated GVs and publications. Number of GVs that advanced through the curation process in each step, along with the number of publications reporting the evidence for association in parenthesis, followed by the number of GVs-publications pairs. GV: genetic variant; P: publication; MD: major depression; VDA: variant-disease association; TR: treatment response; E: environmental; QC: quality control.

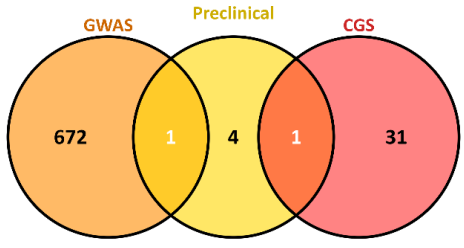


Figure 2. Curated set of GVs. Overlap of curated GVs associated with MD by study type (i.e., GCS, GWAS or Preclinical). CGS: candidate gene studies; GWAS: genome-wide association studies.

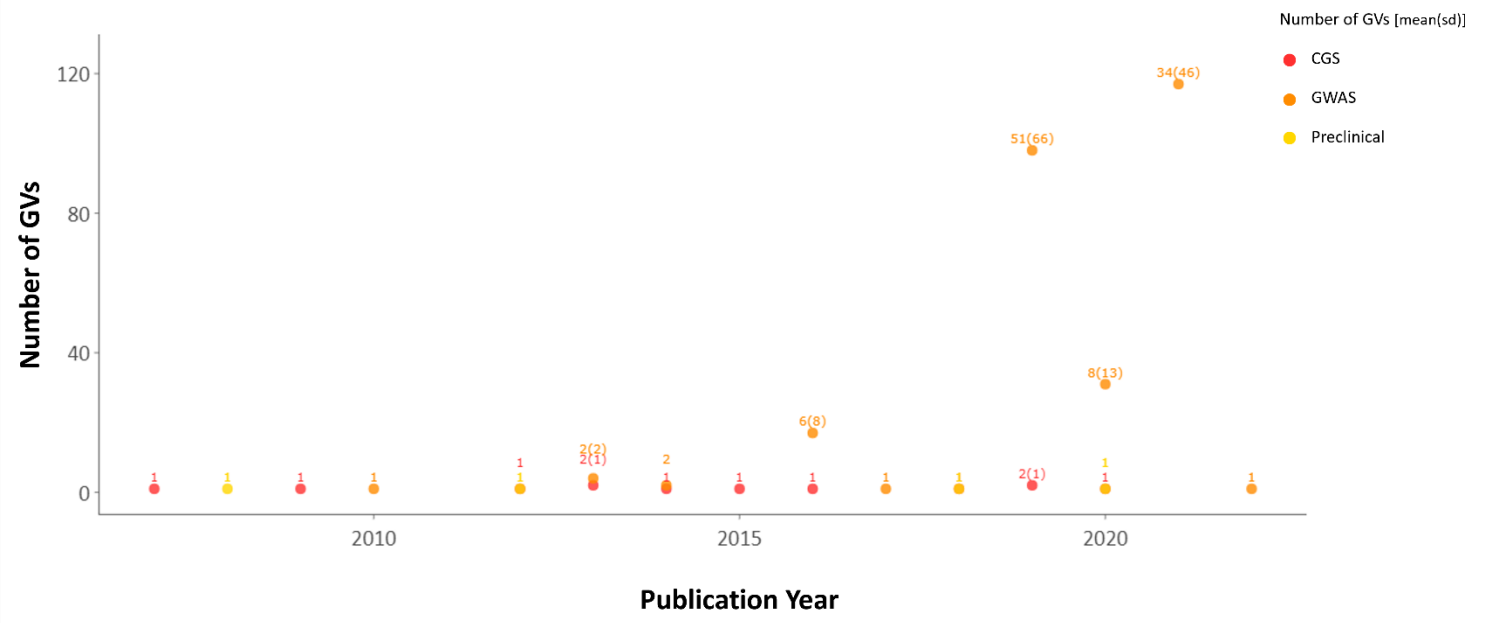


Figure 3. Chronology of publications per study type. Maximum number of GVs associated with MD per year per study type. The mean and standard deviation of GVs identified yearly are on top of each dot. The dot size is proportional to each publication's mean number of genetic variants. The mean is written above, with the standard deviation in parentheses.

Despite the broad set of MD-related terms available, reflecting different manifestations of the disease, almost 60% of these associations were referred to "Major depressive disorder". We could see a usage tendency of a narrow set of 13 terms to refer to MD from the 96 considered (Figure 4).

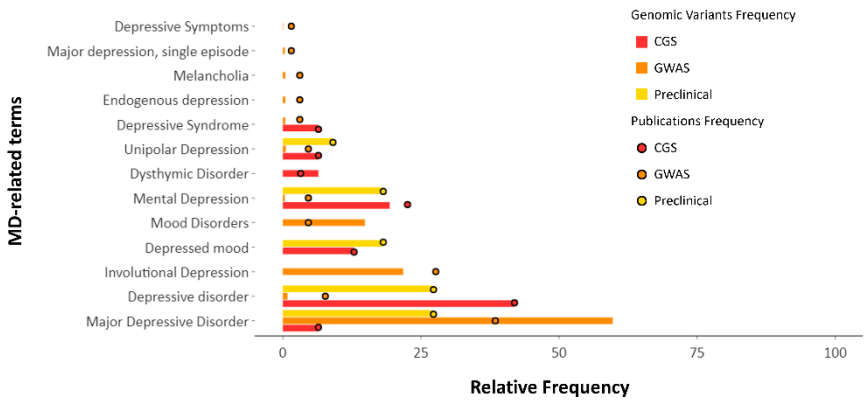


Figure 4. MD terminology. Relative frequency of the genetic variants (bars) and publications (dots) associated with each MD-related term divided by study type.

3.3. Data analysis and validation

3.3.1. Functional analysis

We assigned the curated set of GVs to 563 genes based on their proximity to the closest gene. Four genes were shared between CGS and GWAS (LINC02210-CRHR1, MAPT-AS1, RPL12P8, ESR1) and 4 between GWAS and preclinical studies (RPL31P12 and NEGR1) (Figure 5). Most GVs reported by CGS and preclinical studies were in coding regions of the genome, compared to GVs reported by GWAS (Figure 6).

The functional analysis of these genes revealed their role in neuron development and differentiation as well as synapse assembly, central nervous system and brain development (Supplementary Table S6). Also, there are genes related to immune and inflammatory responses. Phenotypic traits such as behaviour, stress and sleep disturbances are also associated with these genes.

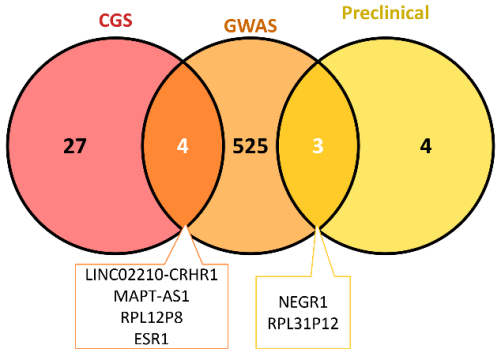


Figure 5. Gene overlap between study types.

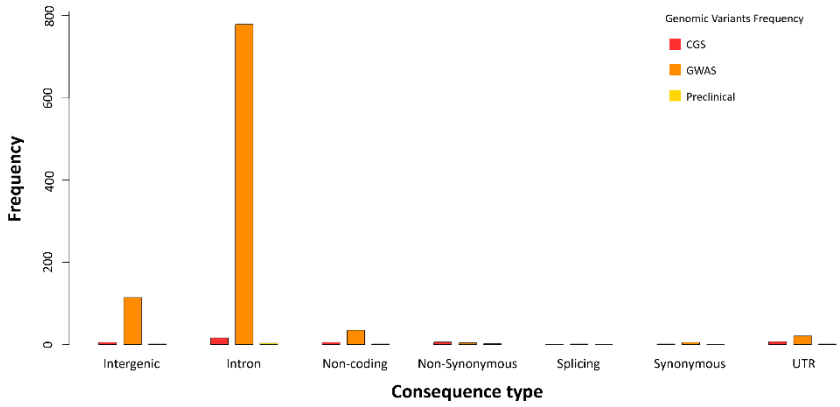


Figure 6. Consequence type of GVs. Consequence type of the curated set of GVs. Note that a GV can have multiple consequence types; thus, the number of consequence types may not necessarily match the total number of GVs. GV: genetic variant; UTR: untranslated region.

3.3.2. Evaluation of GVs from CGS

We reviewed the evidences involving GVs reported in a recent publication by Border et al.[23] (Table 1) as potential false positives because they were identified by CGS and could not be replicated by larger studies. Almost 13% (316/2449) of GV-publications pairs captured by DisGeNET involved these GVs, 59% (187) of which did not meet the sample size curation criteria. As for the GVs, 31% (5/16) did not reach the significance level. Three GVs from this set (rs1801133, rs6265 and rs4680) remained in the final curated set, all evidenced by different CGS.

3.3.3. Data scoring

We ranked the GVs associated with MD according to the study types and publications reporting the association (Supplementary Table S7). Most GVs (642/709) are reported in a single publication, by a single association, and in a single study, getting a score of 0.25. Only 9% of GVs (64) were reported by more than one publication, particularly 3 GVs (rs1021363, rs11135349 and rs12967855) by up to 4 publications. One GV, rs5569, reaches 0.4 by being reported by more than one association type. And 4 GVs reach 0.5 by either being reported by more than one study type or more than one publication and association type.

4. Discussion

Currently, there are publicly available repositories that collect genetic variants (GVs) associated with major depression (MD). Nonetheless, despite the development of several curation guidelines for gene-disease and GV-disease association, their scope is limited to Mendelian and rare diseases with a particular focus on clinical relevance. To the best of our knowledge, no standards exist for assessing multiple evidences for GV-disease associations from different study types in the context of complex disorders. Here, we have developed curation guidelines based on current literature on MD, other diseases' curation guidelines and the collected association data. Ultimately, we have built an expert-curated dataset of GVs associated with MD that could help us understand the complex genetic regulation of MD.

MD is clinically variable with phenotype heterogeneity, which results in a broad terminology used for referring to MD, adding up to 96 UMLS terms. This terminology includes semantically similar terms, MD subtypes and characterising features (e.g., major depression, unipolar depression and anhedonia) (Supplementary Table S2). All these terms were considered for association data collection from DisGeNET and GWAS repositories (i.e., GWAS Catalog and GWASdb).

Our focus for these guidelines are GVs associated with MD. Thus, different features that could influence the study perspective and outcome were considered in the curation steps. Firstly, since MD may be described by different features or traits and is comorbid with many diseases, which would influence the association study design, MD context was characterised (Supplementary Table S3)[29,40]. In parallel, associations wrongly captured by text mining (TM) and that had no relationship with MD were removed. Secondly, the publication

reporting the association was evaluated to keep only those original studies conducted on MD patients, which reported correctly captured GVs. Thirdly, MD can be studied in a particular environmental context (E) or in association with treatment response (TR), so we characterised the association type. Fourthly, the study type was characterised to assess its design quality regarding the minimum sample size requirements and significance threshold.

The developed guidelines are intended for a semi-automatic curation process (Figure 1). The curation can be programmatically boosted by setting up revision flags using regular expressions to capture comorbidities, negations or preclinical models, among others. However, it requires a manual inspection of the publication's full text, especially for evidences extracted from the literature, to determine the number of GVs tested, the sample size and significance.

Only 27% of collected GV-publications pairs made it through the curation process, corresponding to 3% of publications and 70% of GVs (Table 7 and Supplementary Table S5). These numbers reflect that GVs that were tested under different contexts or without a large enough sample size were recovered by other publications. Most CGS did not pass the quality control filters, which is consistent with previous criticisms of CGS's lack of replicability due to small sample sizes[22,23]. Nonetheless, 18 CGS passed the curation criteria meeting quality control requirements reporting significant associations. The relatively small overlap between resources emphasises the importance of combining all of these data sources to develop a comprehensive database of MD-associated GVs (Supplementary Figure S1). The small overlap between GWAS repositories could be because GVs identified by GWAS may not be the ones impacting the disease phenotype, but these could be in linkage disequilibrium (i.e., correlated association between nearby variants). There is also a small overlap between association types (i.e., VDA, TR or E), which could be due to different GVs impacting different aspects of the diseases.

Regarding TR evidences, the evaluated treatments target different neurotransmitters (e.g., dopamine or serotonine) and also include electroconvulsive therapy. As for the environmental factors considered, these involve diverse types of stress as well as traumatic experiences.

The evaluation of GVs according to study type revealed that two GVs were reported by more than one (Figure 2). Li S. et al. identified rs3101339 in a GWAS study and further validated the potential effects of its disruption in the transcription factor binding site of NEGR1 with CRISPR-Cas9-mediated genome editing[45]. Meanwhile, rs17759843 was reported in both human and mice via transcriptomic and genomic analysis[46]. No CGS finding was replicated in GWAS, emphasising the lack of power of CGS, which were generally conducted on samples of tens to hundreds of individuals, and thus supporting these findings being potentially false positives[22,23]. Additionally, this could also be explained because CGS generally target genes and GVs in coding regions compared to GWAS, which tend to identify GVs in non-coding regions with potential regulatory roles (Figure 6). Furthermore, the first CGS performed dates back to 1999, but no study conducted before 2008 passed our curation criteria. Unlike preclinical and CGS, the number of GVs identified by GWAS increased over time, probably due to better sequencing technologies and bigger sample sizes (Figure 3)[3].

Despite the broad terminology characterising MD, only a small set of terms (13.5%) is used in the curated association data (Figure 4). We believe that one reason could be the normalisation process conducted by GWAS repositories, which we have observed generally tends to homogenise towards a set of broader terms[5,6]. Although this broader phenotype annotation and phenotyping process may facilitate access to larger sample sizes it may preclude GVs associated with more fine-granular phenotypic descriptions[47]. Thus, a balance between both approaches is required.

The GVs-to-gene mapping identified 563 genes potentially associated with MD (Figure 5). ESR1 or estrogen receptor α, common to CGS and GWAS GVs, has been associated with MD risk as estrogens influence neurotransmitters turnover and regulate serotonergic neuron activity[48]. LINC02210-CRHR1 encodes a protein involved in the hypothalamic-pituitary-adrenal axis, which is associated with MD pathophysiology[49]. Regarding the GVs' mapped genes common between GWAS and preclinical models, NEGR1 controls the neurological development of neurons[50]. As expected, most GVs reported by GWAS were in non-coding regions of the genome. Therefore, additional analyses of these GVs, such as expression, colocalisation or fine-mapping, are required to assess their role in disease pathogenesis. Additionally, the functional enrichment analysis of these genes showed their role in neuron

development as well as immune and inflammatory responses; all in all, features altered or that characterise MD[2].

The revision of association data, considering the list of GVs published by Border et al.[23] as potentially unsupported by larger studies, revealed only one GV that remained in the final curated set. These analyses also revealed that most of the studies involving these GVs were not properly designed in terms of sample size or did not reach significance when passing the QC, which is consistent with Border et al. findings[23]. However, as before-mentioned, some studies (7/305) did pass the curation criteria yielding valid associations. Despite the criticism levelled at CGS, if properly designed and supported by high-quality prior knowledge, this approach could still be applied to identify the association of particular GVs with MD.

The GVs ranking is based on the number of studies and publications reporting the association (Supplementary Table S7). The results reveal a poor replication of findings across different study types, as seen by GVs overlap evaluation. They also show that only the association with MD of 9% of GVs is supported by more than one publication. In more detail, 3 of these GVs are replicated across up to 4 GWAS on MD publications, and 2 of these are shared by the three GVs. These results are an example of GWAS's replicative power compared to CGS.

Overall, to advance the knowledge of MD genetic architecture, we have developed curation guidelines that consider both features of the disease and different experimental approaches. We have applied these guidelines and created an expert-curated database of GVs associated with MD. The MD-associated GVs are mapped to genes involved in processes related to MD pathogenesis The developed dataset is provided for the community to facilitate downstream analysis of MD-associated GVs by bioinformatic approaches. We believe the developed curation guidelines could be useful for other psychiatric and complex diseases with a similar genetic architecture.

Data availability statement

The original contributions presented in the study are included in the https://doi.org/10.5281/zenodo.7348795, further inquiries can be directed to the corresponding author.

Author contributions

JP-G, JP, and LF designed the curation guidelines. JP-G conducted the curation process and wrote the manuscript with the support and guidance of JP and LF. All authors reviewed the manuscript. The authors read and approved the final manuscript.

Funding

IMI2-JU resources which are composed of financial contributions from the European Union's Horizon 2020 Research and Innovation Programme and EFPIA (GA: 116030 TransQST and GA: 777365 eTRANSAFE), and the EU H2020 Programme 2014–2020 (GA: 676559 Elixir-Excelerate); Project 001-P-001647—Valorisation of EGA for Industry and Society funded by the European Regional Development Fund (ERDF) and Generalitat de Catalunya; Agència de Gestió d'Ajuts Universitaris i de Recerca Generalitat de Catalunya (2017SGR00519), and the Institute of Health Carlos III (project IMPaCT-Data, exp. IMP/00019), co-funded by the European Union, European Regional Development Fund (ERDF, "A way to make Europe"). The Research Programme on Biomedical Informatics (GRIB) is a member of the Spanish National Bioinformatics Institute (INB), funded by ISCIII and ERDF (PRB2-ISCIII (PT13/0001/0023, of the PE I + D + i 2013–2016)). The MELIS is a 'Unidad de Excelencia María de Maeztu', funded by the MINECO (MDM-2014-0370). JP-G was supported by Instituto de Salud Carlos III-Fondo Social Europeo (FI18/00034). This statement is a requirement from our funding agencies and therefore has to be included in the Funding section.

Conflict of interest

Competing interest reported. LF and JP are co-founders and hold shares of Medbioinformatics Solutions SL.

Supplementary material

The Supplementary Material for this article can be found online at: https://doi.org/10.5281/zenodo.7348795.

Abbreviations

CGS	candidate gene studies
ClinGen	Clinical Genome Resource
E	environmental factors

GenCC	Gene Curation Coalition
GV	genetic variant
GWAS	genome-wide association studies
MD	Major Depression
OMIM	Online Mendelian Inheritance in Man
PMID	PubMed unique identifiers
QC	quality control
S	score
SNP	single nucleotide polymorphism
TM	text mining
TR	treatment response
UMLS	Unified Medical Language System
VDA	variant-disease association

References

1. WHO. Depression. https://www.who.int/en/news-room/fact-sheets/detail/depression Who Depression Fact Sheet https://www.who.int/en/news-room/fact-sheets/detail/depression (2018).
2. Li, Z., Ruan, M., Chen, J. & Fang, Y. Major Depressive Disorder: Advances in Neuroscience Research and Translational Applications. Neurosci Bull 37, 863–880 (2021).
3. Unal-Aydin, P., Aydin, O. & Arslan, A. Genetic Architecture of Depression: Where Do We Stand Now? Adv Exp Med Biol 1305, 203–230 (2021).
4. Preskorn, S. H. Drug Development in Psychiatry: The Long and Winding Road from Chance Discovery to Rational Development. in Handbook of Experimental Pharmacology. 307–324 (2018). doi:10.1007/164_2018_169.
5. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 47, D1005–D1012 (2019).
6. Li, M. J. et al. GWASdb: a database for human genetic variants identified by genome-wide association studies. Nucleic Acids Res 40, D1047–D1054 (2012).
7. Hormozdiari, F. et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet 99, 1245–1260 (2016).
8. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. Hum Mol Genet 24, R111–R119 (2015).

9. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. Front Genet 11, 424 (2020).
10. The GenCC Home Page. https://thegencc.org/ https://thegencc.org/ (2020).
11. Thaxton, C. et al. Utilizing ClinGen gene-disease validity and dosage sensitivity curations to inform variant classification. Hum Mutat 43, 1031–1040 (2022).
12. Strande, N. T. et al. Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. Am J Hum Genet 100, 895–906 (2017).
13. Preston, C. G. et al. ClinGen Variant Curation Interface: a variant classification platform for the application of evidence criteria from ACMG/AMP guidelines. Genome Med 14, 1–12 (2022).
14. Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: leveraging knowledge across phenotype–gene relationships. Nucleic Acids Res 47, D1038–D1043 (2019).
15. Pavan, S. et al. Clinical Practice Guidelines for Rare Diseases: The Orphanet Database. PLoS One 12, e0170365 (2017).
16. Martin, A. R. et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat Genet 51, 1560–1565 (2019).
17. Zeng, J. et al. Widespread signatures of natural selection across human complex traits and functional genomic categories. Nat Commun 12, 1164 (2021).
18. Griffith, M. et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet 49, 170–174 (2017).
19. Tamborero, D. et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med 10, 1–8 (2018).
20. Gutierrez-Sacristan, A. et al. PsyGeNET: a knowledge platform on psychiatric disorders and their genes. Bioinformatics 31, 3075 (2015).
21. Shadrina, M., Bondarenko, E. A. & Slominsky, P. A. Genetics Factors in Major Depression Disease. Front Psychiatry 9, 334 (2018).
22. Duncan, L. E., Ostacher, M. & Ballon, J. How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. Neuropsychopharmacology 44, 1518–1523 (2019).
23. Border, R. et al. No Support for Historical Candidate Gene or Candidate Gene-by-Interaction Hypotheses for Major Depression

Across Multiple Large Samples. American Journal of Psychiatry 176, 376–387 (2019).
24. Howard, D. M. et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci 22, 343–352 (2019).
25. Levey, D. F. et al. Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. Nat Neurosci 24, 954–963 (2021).
26. Visscher, P. M. et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. The American Journal of Human Genetics 101, 5–22 (2017).
27. Zastrozhin, M. et al. Effect of Genetic Polymorphism of the CYP2D6 Gene on the Efficacy and Safety of Fluvoxamine in Major Depressive Disorder. Am J Ther 29, E26–E33 (2021).
28. Jiang, T., Wumaier, G., Li, X., Yang, X. & Liu, J. Research on the Effects of Occupational Stress and the DRD2 Gene on the Psychological Health of Workers in the Xinjiang Desert Oil Field. Front Psychiatry 12, (2021).
29. Dunn, E. C., Wang, M.-J. & Perlis, R. H. A Summary of Recent Updates on the Genetic Determinants of Depression. in Major Depressive Disorder 1–27 (Elsevier, 2020). doi:10.1016/B978-0-323-58131-8.00001-X.
30. Corponi, F., Fabbri, C. & Serretti, A. Pharmacogenetics and Depression: A Critical Perspective. Psychiatry Investig 16, 645–653 (2019).
31. Serretti, A. & Fabbri, C. Genetics of Treatment Outcomes in Major Depressive Disorder: Present and Future. Clinical Psychopharmacology and Neuroscience 18, 1–9 (2020).
32. Planchez, B., Surget, A. & Belzung, C. Animal models of major depression: drawbacks and challenges. J Neural Transm 126, 1383–1408 (2019).
33. Wang, Q., Timberlake, M. A., Prall, K. & Dwivedi, Y. The recent progress in animal models of depression. Prog Neuropsychopharmacol Biol Psychiatry 77, 99–109 (2017).
34. Sullivan, P. F. & Geschwind, D. H. Defining the Genetic, Genomic, Cellular, and Diagnostic Architectures of Psychiatric Disorders. Cell 177, 162–183 (2019).
35. Borsini, A. & Zunszain, P. A. Advances in Stem Cells Biology: New Approaches to Understand Depression. in Research and Perspectives in Endocrine Interactions 123–133 (Springer Verlag, 2016). doi:10.1007/978-3-319-41603-8_10.

36. Jantas, D. Cell-Based Systems of Depression: An Overview. in Herbal Medicine in Depression (ed. Grosso, C.) 75–117 (Springer International Publishing, 2016). doi:10.1007/978-3-319-14021-6_3.
37. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 32, D267–D270 (2004).
38. Oscanoa, J. et al. SNPnexus: A web server for functional annotation of human genome sequence variation (2020 update). Nucleic Acids Res 48, W185–W192 (2020).
39. Raudvere, U. et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 47, 191–198 (2019).
40. Gold, S. M. et al. Comorbid depression in medical diseases. Nat Rev Dis Primers 6, 1–22 (2020).
41. Zondervan, K. T. & Cardon, L. R. Designing candidate gene and genome-wide case-control association studies. Nat Protoc 2, 2492–2500 (2007).
42. Gauderman, W. J. Sample Size Requirements for Association Studies of Gene-Gene Interaction. Am J Epidemiol 155, 478–484 (2002).
43. Sullivan, P. F. The Psychiatric GWAS Consortium: Big Science Comes to Psychiatry. Neuron 68, 182–186 (2010).
44. Nishino, J., Ochi, H., Kochi, Y., Tsunoda, T. & Matsui, S. Sample Size for Successful Genome-Wide Association Study of Major Depressive Disorder. Front Genet 9, 227 (2018).
45. Li, S. et al. Regulatory mechanisms of major depressive disorder risk variants. Mol Psychiatry 25, 1926–1945 (2020).
46. Wingo, A. P. et al. Expression of the PPM1F gene is regulated by stress and associated with anxiety and depression. Biol Psychiatry 83, 284 (2018).
47. Mcintosh, A. M., Sullivan, P. F. & Lewis, C. M. Uncovering the Genetic Architecture of Major Depression. Neuron 102, 91–103 (2019).
48. Ryan, J. & Ancelin, M. L. Polymorphisms of estrogen receptors and risk of depression. Drugs 72, 1725–1738 (2012).
49. Normann, C. & Buttenschøn, H. N. Gene–environment interactions between HPA-axis genes and stressful life events in depression: a systematic review. Acta Neuropsychiatr 31, 186–192 (2019).
50. Pischedda, F. et al. A cell surface biotinylation assay to reveal membrane-associated neuronal cues: Negr1 regulates dendritic arborization. Molecular and Cellular Proteomics 13, 733–748 (2014).

3.2. Functional genomics analysis to disentangle the role of genetic variants in major depression

Our knowledge of GVs associated with MD has largely increased thanks to GWAS. However, most of these GVs lie in non-coding regions of the genome, and functional genomics analyses are required to further understand the underlying biological mechanisms. In this chapter, we aim to gain insights into how MD-associated GVs influence disease pathogenesis. We develop a bioinformatics pipeline that overcomes the limitations of some current GWAS datasets, for which full genome summary statistics are unavailable. Then, we apply this pipeline to a recent GWAS meta-analysis on MD to prioritise putative causal GVs either altering gene expression (i.e., eQTLs) or TFBS. Finally, we propose mechanistic hypotheses for these GVs in MD.

Pérez-Granado, J., Piñero, J., Medina-Rivera, A. & Furlong, L. I. Functional Genomics Analysis to Disentangle the Role of Genetic Variants in Major Depression. *Genes (Basel)* **13**, 1259 (2022).

Functional Genomics Analysis to Disentangle the Role of Genetic Variants in Major Depression

Judith Pérez-Granado[1], Janet Piñero[1,2] Alejandra Medina-Rivera 3,* and Laura I. Furlong[1,2*]

3. Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Medical Research Institute (IMIM), Department of Medicine and Life Sciences (MELIS), Universitat Pompeu Fabra (UPF), Dr. Aiguader 88, 08003 Barcelona, Spain; jperez2@imim.es (J.P.-G.); janet.pinero@upf.edu (J.P.)
4. MedBioinformatics Solutions SL, Almogàvers 165, 08018 Barcelona, Spain
5. Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Campus Juriquilla, Blvd Juriquilla 3001, Santiago de Querétaro 76230, Mexico

*Correspondence: amedina@liigh.unam.mx (A.M.-R.); laura.furlong@upf.edu (L.I.F.); Tel.: +52-555-623-4331 (A.M.-R.); +34-933-160-540 (L.I.F.)

Abstract

Understanding the molecular basis of major depression is critical for identifying new potential biomarkers and drug targets to alleviate its burden on society. Leveraging available GWAS data and functional genomic tools to assess regulatory variation could help explain the role of major depression-associated genetic variants in disease pathogenesis. We have conducted a fine-mapping analysis of genetic variants associated with major depression and applied a pipeline focused on gene expression regulation by using two complementary approaches: cis-eQTL colocalization analysis and alteration of transcription factor binding sites. The fine-mapping process uncovered putative causally associated variants whose proximal genes were linked with major depression pathophysiology. Four colocalizing genetic variants altered the expression of five genes, highlighting the role of SLC12A5 in neuronal chlorine homeostasis and MYRF in nervous system myelination and oligodendrocyte differentiation. The transcription factor binding analysis revealed the potential role of rs62259947 in modulating P4HTM expression by altering the YY1 binding site, altogether regulating hypoxia response. Overall, our pipeline could prioritize putative causal genetic variants in major depression. More

importantly, it can be applied when only index genetic variants are available. Finally, the presented approach enabled the proposal of mechanistic hypotheses of these genetic variants and their role in disease pathogenesis.

Keywords: major depression; genetic variants; eQTL; colocalization analysis; transcription factors; genetic regulation.

1. Introduction

Major Depression (MD) is the leading cause of impairment around the world [1]. It is mainly treated with both psychotherapy and drugs, but the latter is only effective in 40% of the patients [2]. Currently, there are no available biomarkers or tests that can aid in either MD diagnosis or personalized treatment. As a complex disease, multiple genetic variants (GVs) have been associated with MD in Genome-Wide Association Studies (GWAS), most of them falling within non-coding regions of the genome [3,4].

Functional follow-up studies to unravel the regulatory mechanisms by which these GVs play a role in the disease are key to understanding the molecular underpinnings of the disease and identifying biomarkers or new drug targets. Some authors propose that the efforts should be centered on the interpretation of GWAS signals to identify the causal GVs, meaning those with a biological effect on a disease, and their regulatory potential, instead of pursuing more GWAS [5].

In this study, we have focused on the GWAS meta-analysis on MD performed in 2019 by Howard et al. [3]. Full-genome summary statistics are not publicly available for this GWAS, so we have leveraged available data on index GVs. Ninety-seven loci were identified as significantly associated with MD, and these underwent the classic post-GWAS analysis: a gene-set enrichment analysis, the computation of polygenic risk score, and genetic correlation with other traits, as well as drug-gene interaction analysis. In line with previous GWAS findings, most GVs lie in non-coding regions, thus having no obvious direct effect on a gene.

A necessary step forward to disentangle the role of GVs identified in GWAS requires the evaluation of functional regulatory variation. Here, we have pursued two complementary analytical approaches geared toward the use of index GVs: (1) identification of candidate

susceptibility genes using expression quantitative trait loci in cis (cis-eQTLs), which are enriched among disease-associated loci [6], and (2) characterization of transcription factor (TF) binding sites modified by GVs, which are key to understanding their potential impact on regulatory mechanisms [6–8].

In the present study, we aim to advance the understanding of MD molecular underpinnings. We have designed and applied a regulatory variation analysis pipeline and conducted a functional enrichment analysis of the GVs, either acting as eQTLs or altering the transcription factor binding site (TFBS), along with the proximal (pGenes) and regulated genes (eGenes). Our findings provide biological insights into the functional role of MD GVs and enable the proposal of mechanistic hypotheses.

2. Materials and methods

2.1. MD GWAS Dataset and LD expansion

In order to obtain a comprehensive and reliable set of genetic variants (GVs) associated with major depression (MD), we focused our analysis on the GWAS meta-analysis from Howard et al. [3]. This meta-analysis evaluated 807,553 European individuals (246,363 cases and 561,190 controls) and identified 102 genetic variants (GVs) associated with MD. We retrieved these data from the summary statistics available at GWAS Catalog [9] (Accession Study: GCST007342, note that the full-genome summary statistics for this GWAS were not publicly available; downloaded in December 2020). We filtered the GVs by genome-wide significance (p-value $\leq 5 \times 10^{-8}$ and proceeded with the analysis with this set. We then fine-mapped MD-associated GVs to prioritize the causal ones using the Probabilistic Identification of Causal SNPs (PICS) algorithm [10]. In brief, PICS takes the most significant variant per association locus and performs LD expansion using the 1000 Genomes Project linkage disequilibrium (LD) information data for the study population and then identifies the GVs more likely to be causal (PICS probabilities). Using the PICS2 Data portal, we downloaded the precomputed PICS GVs for this study. This data constituted our full dataset of GVs.

2.2. GVs Annotation: VEP, CADD and ENCODE

We annotated the full set of GVs with Variant Effect Predictor (VEP) [11] and Combined Annotation Dependent Depletion (CADD) [12].

VEP annotates GVs' consequence type using the Sequence Ontology, its allele frequency from the 1000 Genomes Project Phase 3 along with the genomic coordinates, chromosome, and mapped gene at ±5000 bp distance (from now on pGenes). Combined Annotation Dependent Depletion (CADD) assesses GVs' potential pathogenicity by evaluating the PHRED-like scaled C-score; the recommended cut-off ≥ 15 was set to identify potentially pathogenic variants.

We analyzed the GVs with the Encyclopedia of DNA Elements (ENCODE) [13] to identify those potentially lying in transcription factor binding sites (TFBS). ENCODE data analysis was performed using SNPNexus [14], an online platform that allows a comprehensive annotation of GVs by integrating multiple tools.

2.3. Fine-Mapping and Colocalization of GWAS and cis-eQTLs

PICS2, in addition to GWAS PICS GVs, has precomputed PICS GVs for all Genotype- Tissue Expression (GTEx) V8 best eQTLs per gene, per tissue type. We overlapped the extracted GWAS PICS for MD GVs with GTEx cis-eQTL PICS GVs, filtering both sets by a PICS probability greater than 10% to narrow down the set to the most likely causal GVs without being overly permissive, as previous applications of this method have done [15]. We performed a Fisher test to assess the enrichment of GVs in eQTL regions. Finally, to identify colocalizing GWAS and eQTL GVs, we computed the products of PICS probabilities following the colocalization posterior probability (CLPP) method, which assumes independence of causal probabilities for GWAS and eQTL GVs [16]. The genes regulated by these eQTLs from now on will be referred to as eGenes.

2.4. TF Binding Analysis with RSAT Variation Tools

We predicted those GVs affecting the TFBS using the Regulatory Sequence Analysis Tools (RSAT) suite, which evaluates cis-regulatory elements. First, we used ENCODE ChIP-seq data to keep only the GVs lying in TFBS and, therefore, have a more biologically relevant set of GVs and reduce the number of tests. However, ChIP-seq data retrieve regions of around 100–1000 bp, but the actual binding site corresponds to 9–15 bp [17,18]. Thus, we proceeded with the RSAT analysis for a more robust and accurate assessment of the GVs potentially altering the TFBS. RSAT provides tools that evaluate cis-regulatory elements to

predict GVs affecting the TFBS by modifying the transcription factor (TF) binding affinity.

RSAT modular structure allowed the concatenation and independent execution of programs, each with a different goal. Before scanning the GVs and in order to account for their different nucleotide composition, we created four sets of background models according to the GV's functional impact obtained with VEP (i.e., intergenic and UTR, intronic, regulatory, and non-coding GVs). The subsequent steps were performed for each set separately. The module *create-background-model* was executed using the sequences obtained with *fetch-sequences-from-UCSC*, with the peak regions retrieved by ENCODE as input. In parallel, the module *retrieve-variation-sequence* was used to obtain the flanking sequence (30 bp per side) of the GVs of interest, using the dbSNP, genomic coordinates, reference, and alternative allele.

To assess the TFBS alterations, position weight matrices (PSSM) for TFs expressed in brain tissues (filtering them using GTEx expression data, ≥2 transcripts per million (TPM)) [19] were retrieved from the following databases: JASPAR [20], ENCODE, HOCOMOCO [21], footprintDB [22], and hPDI [23]. In all cases, the non-redundant Homo sapiens database version was used.

Finally, the module *variation-scan* was run with the previously built background Markov models (order 2 to account for CpGs without overfitting), the PSSM matrices, the GVs with their flanking sequences (see above), and the following parameters: weight of predicted sites (>1), weight difference between variants (>1), p-value of predicted sites (<1 e-3), and p-value ratio between variants (>10). The weight represents the binding affinity and the p-value of a score is the probability of observing a score of at least weight given a background model.

In addition, two control datasets, one randomizing TF motifs and one randomizing GVs, were built to validate the results obtained running RSAT with the GVs of interest. On the one hand, the TF's PSSMs matrices were permuted using permute-matrix -perm 5 to get randomized matrices with the same nucleotide composition and information content. On the other hand, a control set of GVs (1:10) was built using vSampler [24] with the following parameters: distance to closest transcription start site (TSS) deviation (±5000), gene density

deviation (±5 in 100 kbp), number of variants in LD (±50 and $r^2 = 0.1$), controlling for coding/non-coding match and variant type specificity, excluding for input GVs and sampling across the chromosome. Both controls were analyzed with the described RSAT pipeline.

We compared our set of GV-TF motif pair p-value ratios against the distribution of p-value ratios for the given motif in both control datasets. A Wilcox test was used to evaluate the results obtained from the controls because normality of p-value ratio distribution could not be assumed for most motifs after running a Shapiro–Wilk test. The alternative hypothesis tested was "greater".

In addition, to further confirm statistically significant GVs, a larger negative control dataset of GVs (1:1000) was generated. Again, vSampler was used with relaxed parameters to get a bigger control set (i.e., controlling for coding/non-coding match and variant type specificity, excluding for input SNPs, and sampling across chromosomes). The same non-parametric test was used to evaluate the results.

2.5. Identification of TF Active Regions with ChromHMM

We used chromatin state annotations from ChromHMM [25,26], available from ENCODE (v3), to evaluate whether GVs significantly altering the TFBS were lying in active transcription sites of brain regions. Under a 18-state ChromHMM model, we consider the following states annotations as active regulatory regions [26]: TssA, TssFlnk, TssFlnkU, TssFlnkD, Tx, TxWk, EnhG1, EnhG2, EnhA1, EnhA2, EnhWk, ZNF/Rpts. The available brain regions and cell types were: Brain Angular Gyrus, Brain Inferior Temporal Lobe, Brain Cingulate Gyrus, Brain Anterior Caudate, Brain Substantia Nigra, Brain Dorsolateral Prefrontal Cortex, Brain Hippocampus Middle, and Astrocytes. Additionally, the resulting TFs whose binding was altered were filtered by their expression in the specific brain region using GTEx matched data when available; otherwise, data for all brain regions were considered.

2.6. Retrieval of Regulation Evidence

We looked for evidence of gene expression regulation of TFs by matching GVs-TFs pairs from the TF binding analysis using RSAT with eQTL PICS GVs. We further explored the hTFtarget database [27] to

identify specific mechanistic regulation evidence of those TFs whose binding is altered by our set of GVs to regulate the expression of the target eGenes. The hTFtarget database contains associations of TFs and their targets from chromatin immunoprecipitation sequencing (ChIP-seq) in a specific tissue. We considered evidences for mechanistic regulation when eQTL and ChIP-seq data tissues matched.

2.7. pGenes, eGenes, and GVs Characterization

We conducted a gene-set enrichment analysis using the tool g:Profiler via the R package gprofiler2 [28], which integrates different resources and annotates enriched terms at the following levels: (1) biological processes, molecular functions, and cellular processes annotated with the Gene Ontology (GO); (2) pathways from Reactome (REAC) and WikiPathways; (3) miRNA annotations from MIRNA; (4) phenotypic features associated to disease from Human Phenotype, which is mainly focused on rare Mendelian disorders. In addition, we included DISGENET plus [29,30] association data (v16) in this analysis to evaluate the annotation of complex diseases and phenotypic traits; note that the study by Howard et al. was removed from this dataset to avoid circularity. Variant-set functional enrichment analysis was performed using variant association data from DISGENET Plus. We considered the set of known genes as the domain scope for the analysis. Furthermore, we characterized tissue expression using GTEx gene expression data (v8).

We performed these analyses for the following two gene-sets: (1) genes mapped to by MD-associated GVs (pGenes) and (2) genes regulated by cis-eQTLs (eGenes), and two variant-sets: (1) causal GVs and (2) colocalizing GVs.

3. Results

3.1. Major Depression Associated Genetic Variants Lie in Non-Coding Regions of the Genome

The GWAS study by Howard et al., 2019, reported 102 risk loci associated with major depression (MD), 97 with a p-value $\leq 5 \times 10^{-8}$, which were the starting point of our analysis. After LD expansion, we obtained a set of 5723 potentially causal genetic variants (GVs) (Supplementary Scheme S1 and Table S1). We annotated these GVs with VEP [11] and CADD [12] (Supplementary Figure S1). The identification of probable causal GVs using PICS fine-mapping GWAS

data [10] revealed 172 GVs (PICS>10%) in LD with the 97 GWAS risk loci (Supplementary Table S2). These GVs are located in different regions of the genome, but most of them are in non-coding regions, being mainly annotated as intronic (30%), intergenic (30%), or located in non-coding transcript regions (17%) (Figure 1A). Only two GVs lie in exonic regions (i.e., synonymous and nonsynonymous consequence types). The median allele frequency of these GVs was 0.364 (with more deleterious consequence types having lower allele frequencies) (Figure 1B). Only 4% (7) of the GVs were predicted by CADD as potentially pathogenic (Figure 1C). The fine-mapped GVs were assigned to 95 proximal genes (5000 bps), from now on referred to as pGenes. pGenes were classified based on their expression across tissues based on GTEx gene expression data [19]. Using hierarchical clustering, genes were divided into three roughly equally distributed clusters that seem to correspond to constitutively, lowly expressed in all tissues, and highly expressed in brain tissues (Supplementary Figure S2).

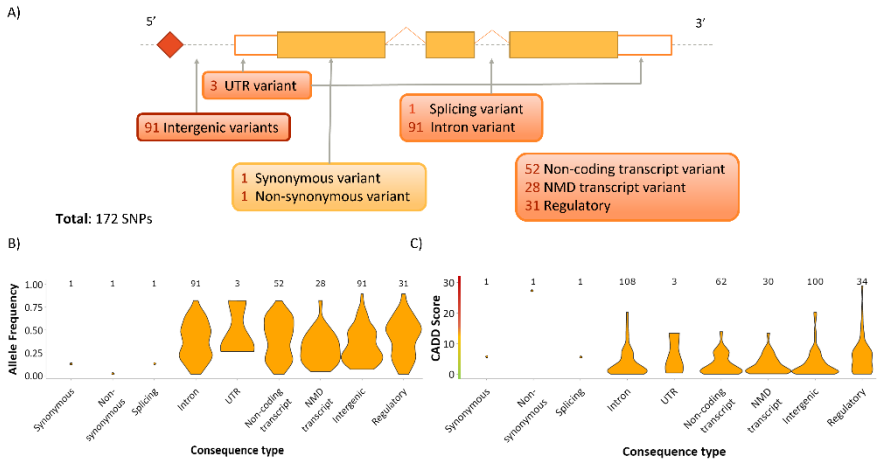


Figure 1. MD GVs are mostly non-coding, common, and potentially not pathogenic. GVs distributed along the genome according to their consequence type predicted with VEP. (B) Allele frequency density, according to GV's consequence type, also predicted with VEP. (C) Pathogenicity score (predicted by CADD) density per consequence type. Please note that a GV can have multiple consequence types; thus, the number of consequence types may not necessarily match the total number of GVs. MD: major depression; GV: genetic variant; VEP: Variant Effect Predictor; CADD: Combined Annotation Dependent Depletion; SNP: single nucleotide polymorphism; UTR: untranslated region; NMD: nonsense- mediated decay.

The pGenes are functionally enriched in GO terms related to nervous system development, neuron differentiation, synaptic signaling, and

different cellular components of the neuron such as dendrite, axon, or synapse (Supplementary Table S3); these biological processes and molecular functions are involved in the pathophysiology of MD [31]. pGenes are associated with an abnormal nervous system morphology and physiology according to the Human Phenotype ontology. Disease enrichment analysis shows enrichment for the association of both pGenes and causal GVs with major depressive disorder and other related mental disorders such as schizophrenia or bipolar disorder (Figure 2 and Supplementary Table S4). pGenes are also associated with comorbid phenotypes and conditions in MD, such as smoking behavior, body mass index, and duration of sleep [32]. Notably, 37% of pGenes and 42% of GVs have no previous evidence of association with depression or other mental disorders.

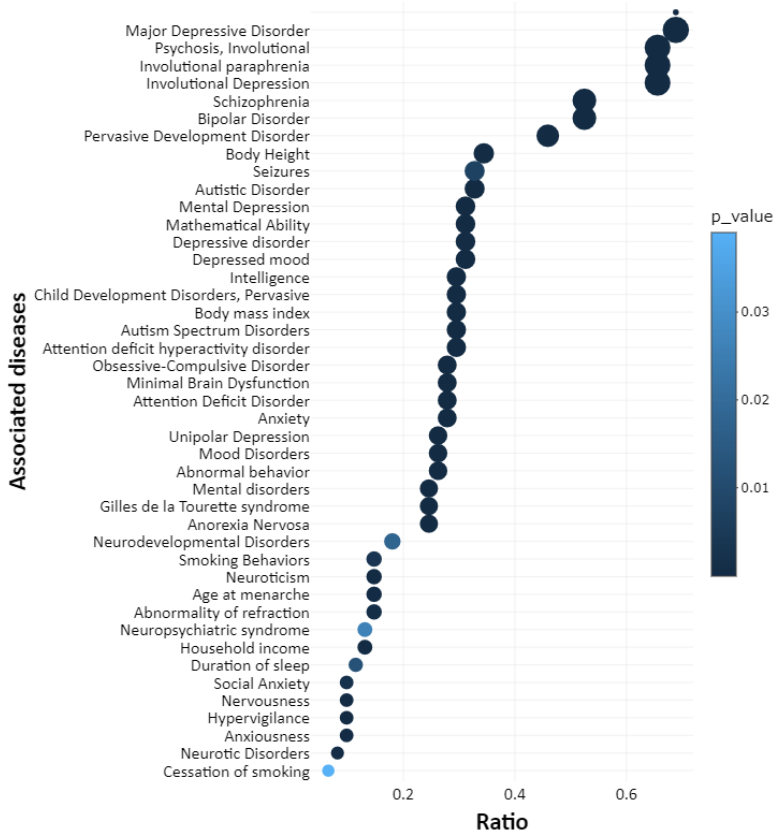
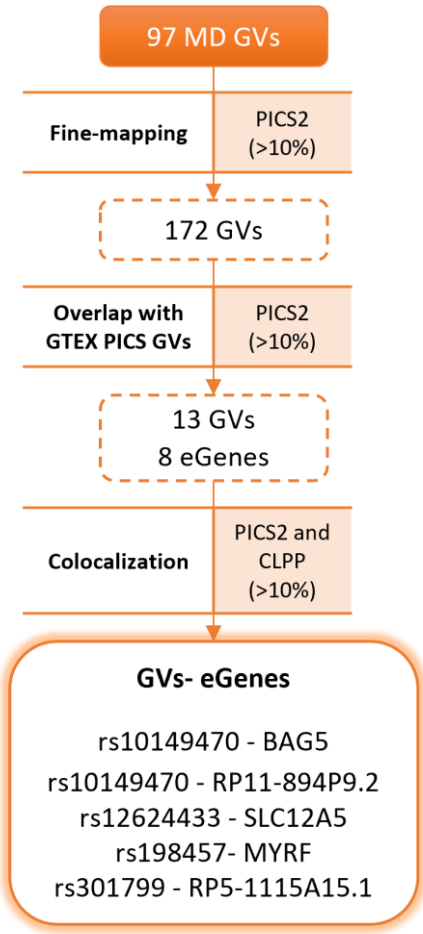


Figure 2. pGenes are associated with mental disorders. Result of the disease enrichment analysis. The ratio corresponds to the number of pGenes associated with each disease relative to all pGenes. Dot size is proportional to the number of pGenes associated with each disease. Gene enrichment analysis was performed using g:Profiler and the DISGENET plus database. pGenes: proximal genes.

Some of the pGenes are associated with processes related to MD pathogenesis, such as TLR4, involved in immune response [33], ESR2, a regulator of estrogen response [34], TCF4, with a role in nervous system development [35], DCC, in charge of axon guidance and neuronal connectivity [36], PAX5, which interferes in mouse neural stem cells proliferation and migration [37,38], and CYP7B1, that participates in the metabolism of the neurosteroids DHEA and pregnenolone [39]. Among the potentially pathogenic GVs, according to CADD, there are 3 intronic GVs lying in ZNF536, a gene involved in the negative regulation of neuron differentiation [40], a relevant process in MD pathogenesis and treatment [41]. rs1021362 lies in SORCS3, a gene previously associated with stress response associated with MD [37,42], rs3793577 lies in ELAVL2, whose silencing in animal models is associated with reduced behavioral despair [43]; the remaining GVs have been previously associated with major depression by several PheWAS studies [15]

3.2. Major Depression Causal Genetic Variants Regulate the Expression of Genes in Cis

The 172 fine-mapped GWAS GVs overlap with 13 GTEx PICS GVs (Scheme 1), revealing an enrichment of MD causal GVs in eQTLS (p-value = 7.392×10^{-10}). The colocalization analysis to identify GVs associated with both MD GWAS and cis-eQTLs resulted in 5 GV–eGenes pairs (i.e., genes whose expression is regulated by these GVs; rs10149470—BAG5, rs10149470—RP11-894P9.2 [ENSG00000258851.1], rs12624433—SLC12A5, rs198457—MYRF, rs301799—RP5-1115A15.1 [ENSG00000232912.5]), with a colocalization probability greater than 10% (Table 1). BAG5 and SLC12A5 are involved in neuron projection [44,45] and MYRF in central nervous system myelination [46]. In addition, all eQTLs have been previously associated with MD and other mental disorders according to DISGENET plus [30,47,48] (Supplementary Table S5). The eGenes BAG5, SLC12A5, and MYRF show higher expression levels in brain regions according to GTEx (Supplementary Figure S3). Little is known about the function of the long non-coding RNAs RP11-894P9.2 and RP5-1115A15.1.



Scheme 1. Fine-mapping and colocalization analysis of MD GVs. MD GWAS GVs have been fine-mapped using PICS and overlapped with GTEx PICS GVs to ultimately perform a colocalization analysis identifying 4 colocalizing GVs affecting the expression of 5 eGenes. MD: major depression; GV: genetic variant; GWAS: genome-wide association studies; PICS: Probabilistic Identification of Causal SNPs; GTEx: Genotype-Tissue Expression; eGenes: genes regulated by expression quantitative trait loci; CLPP: colocalization posterior probability.

3.3. MD Associated GVs Affect the TFBS in Regulatory Regions of Genes Relevant for the Disease

The initial set of 5723 GVs associated with MD was first mapped to transcription factor binding sites (TFBS) using Chip-Seq data from ENCODE. A total of 955 GVs were identified as potentially altering the TFBS of 155 TFs (Scheme 2). The GVs' functional impact was

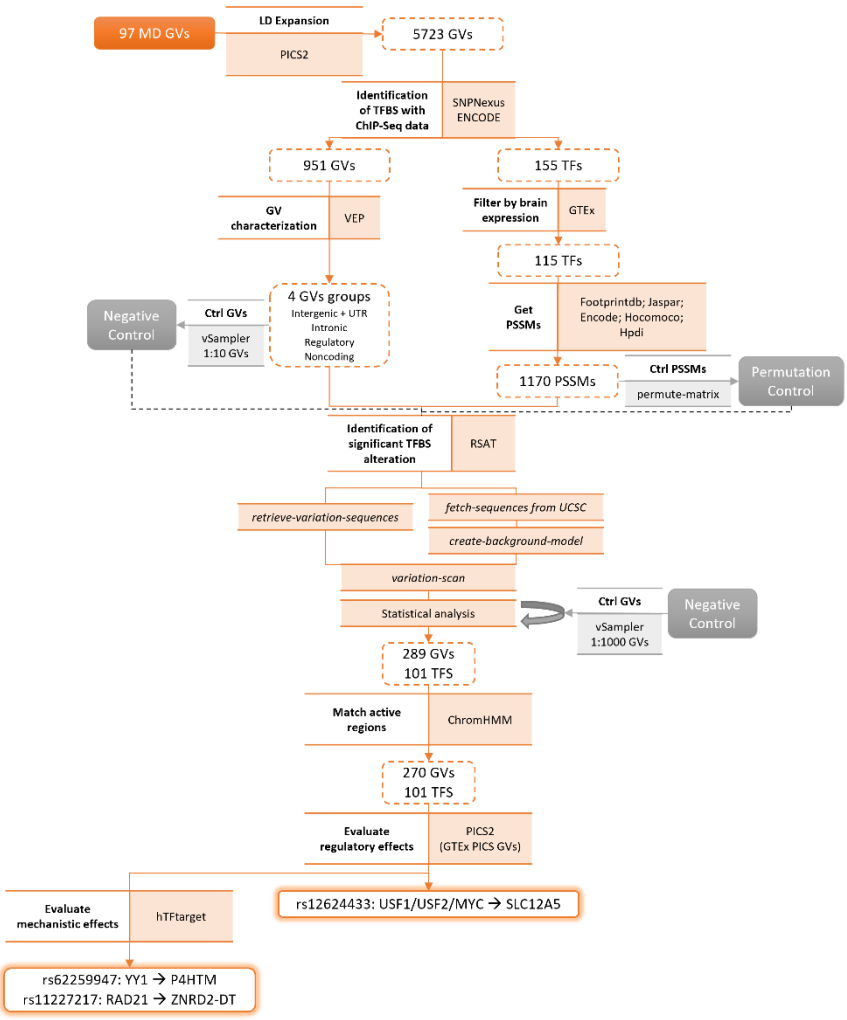
assessed with VEP, and 4 sets were created: (a) intergenic and UTR GVs (333), (b) intronic GVs (562), (c) regulatory GVs (303), and (d) non-coding GVs (389). In addition, we further selected those transcription factors (TFs) that were expressed in brain tissues (≥2 TPM), which left 115 TFs.
Using a pattern matching approach (variation-scan) [49], we identified GVs likely affecting TFBS. As negative controls, we permuted TF motifs and randomly selected variants matching GVs properties (see Methods). Using permuted motifs and randomly selected variants (1:10) as negative controls, we obtained a total of 306 GVs significantly altering the TFBS of 102 TFs (considering the 4 sets together). Ultimately, 289 GVs and 101 TFs passed the statistical analysis using randomly selected variants (1:1000) as negative control. From this final set, 171 GVs are predicted to disrupt the TFBS of 89 TFs, whereas 143 GVs are predicted to create a TFBS for 82 TFs (Supplementary Table S6). Most of these GVs were not characterized as potentially pathogenic by CADD except for 11 GVs (score ≥ 15).

A total of 270 GVs lie in active regulatory regions of the genome of brain-related tissues and cell types according to the epigenome annotation from the ENCODE project based on ChromHMM data (Supplementary Table S7) [25,26]. We then looked for evidence of their impact on gene expression regulation by evaluating their annotation to GTEx eQTLs fine-mapped via PICS. The only GV in this dataset of 270 GVs that also fulfills the criteria of being causal and overlapping GWAS and eQTL PICs in the brain with a probability greater than 10% was rs12624433, which is an eQTL for the gene SLC12A5. This GV is predicted to significantly alter the TFBS of USF1, USF2, and MYC. Both rs12624433 and SLC12A5 have been previously associated with major depression disorder and other mental disorders such as bipolar disorder or schizophrenia [48].

In addition, we also inspected the hTFtarget database [27], looking for evidence of a mechanistic association between the eGenes, considered the targets, and the TFs whose binding site is being altered by the GVs. Focusing on brain regions, we have evidence for two GV-TF-eGene/target associations (rs11227217: RAD21 -> ZNRD2-DT [ENSG00000260233.3]; rs62259947: YY1 -> P4HTM).

GV	eGene	Tissue	PICS Probability GWAS	PICS Probability eQTL	Colocalization Probability
rs10149470	BAG5	Artery Tibial	0.9657	0.633	0.6112881
rs10149470	RP11-894P9.2	Colon Sigmoid	0.9657	0.633	0.6112881
rs10149470	RP11-894P9.2	Esophagus Gastroesophageal Junction	0.9657	0.633	0.6112881
rs10149470	RP11-894P9.2	Esophagus Muscularis	0.9657	0.584	0.5639688
rs10149470	RP11-894P9.2	Artery Aorta	0.9657	0.499	0.4818843
rs10149470	RP11-894P9.2	Breast Mammary Tissue	0.9657	0.4494	0.43398558
rs12624433	SLC12A5	Brain Putamen Basal Ganglia	0.7355	0.303	0.2228565
rs10149470	RP11-894P9.2	Stomach	0.9657	0.1782	0.17208774
rs10149470	RP11-894P9.2	Adipose Subcutaneous	0.9657	0.1621	0.15653997
rs10149470	RP11-894P9.2	Colon Transverse	0.9657	0.1419	0.13703283
rs10149470	RP11-894P9.2	Adipose Visceral Omentum	0.9657	0.1412	0.13635684
rs198457	MYRF	Thyroid	0.9627	0.1258	0.12110766
rs10149470	RP11-894P9.2	Heart Left Ventricle	0.9657	0.1225	0.11829825
rs301799	RP5-1115A15.1	Whole Blood	0.6946	0.1542	0.10710732

Table 1. GWAS-eQTL Colocalizing GVs in MD. MD associated GVs colocalizing with eQTLs. GWAS: genome-wide association studies; eQTL: expression quantitative trait loci; GV: Genetic variant; MD: major depression; eGene: gene regulated by eQTL; PICS: Probabilistic Identification of Causal SNPs.



Scheme 2. Identification of altered TFBS using RSAT. Pipeline followed to identify GVs associated with MD that significantly alter TFBS. Methodologies are referred to in bold and along with them are the resources used. Highlighted in grey are the control datasets. TFBS: transcription factor binding site; RSAT: Regulatory Sequence Analysis Tools; GV: genetic variant; MD: major depression; LD: linkage disequilibrium; ENCODE: Encyclopedia of DNA Elements; VEP: Variant Effect Predictor; Ctrl: control; UTR: untranslated region; TF: transcription factor; GTEx: Genotype-Tissue Expression; PSSM: position weight matrix; PICS: Probabilistic Identification of Causal SNPs.

The GV rs62259947 has been annotated as an eQTL downregulating the expression of P4HTM in the Brain Cerebellar Hemisphere. We propose this effect is likely being mediated by the GV significantly

changing the affinity for YY1 binding (weight difference = 14.98 and p-value ratio = 5058.82) (see Methods), a TF known to participate in gene regulation though looping of the DNA [50]. The eGene P4HTM has been associated with the hypoxia-inducible factor HIF1α mediating calcium signaling [51], and its null mutation reduces behavioral despair [52] (Figure 3).

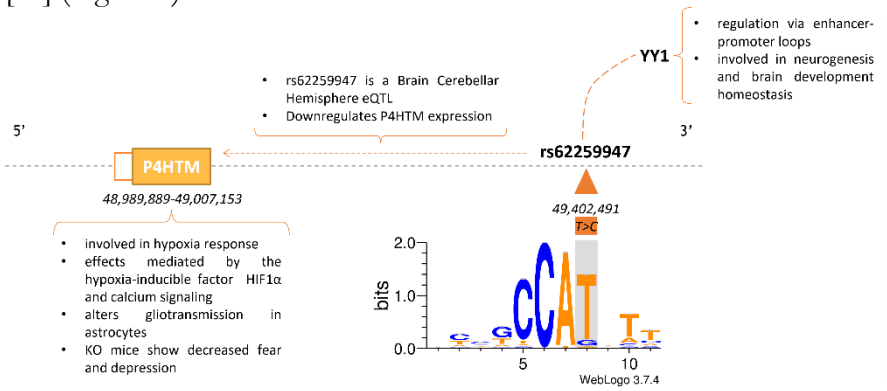


Figure 3. The GV rs62259947 might disrupt the binding of YY1, thus affecting the expression of P4HTM and resulting in behavioral alterations. rs62259947 is an eQTL downregulating the expression of P4HTM and is predicted to disrupt the binding of the TF YY1. This is represented by the sequence logo of the binding site with the nucleotide change highlighted in grey. YY1 is involved in neurogenesis and, in turn, controls the expression of P4HTM, which is mediated by HIF1α regulates calcium signaling and is also associated with behavior. GV: genetic variant; eQTL: expression quantitative trait loci; TF: transcription factor; KO: knockout.

4. Discussion

Despite the large volume of genetic information available, the pathogenesis and etiology of MD are not yet fully understood, probably because most GVs lie in non-coding regions with no obvious direct effect on a gene or function. In this context, leveraging multiple omics data is key for moving forward in the understanding of the influence of genomic variants in MD disease development. On top of that, full-genome summary statistics are not readily available due to study sharing policies (especially for private–public research partnerships) hampering the usage of most post-GWAS data analysis tools. This study aims to unravel the role of MD GVs in genetic regulation by focusing on regulatory variation following two complementary approaches: cis-eQTLs and TF binding alterations. Both are key to identifying potentially causal genes and understanding gene expression regulation [6,8], as reported by supporting evidence for its association with other mental disorders [53,54,55] and with MD in particular [56,57,58]. The

regulatory variation analysis pipelines we have implemented involve fine-mapping, cis-eQTL colocalization, transcription factor binding analysis, and chromatin accessibility data, specially designed to perform well when full-genome summary statistics are not available [59]. These pipelines are in line with other approaches that leverage available omics data, and as such, they could be applied to other complex disorders with a similar genetic architecture and similar data access issues [53,60,61].

Multiple GVs have been associated with MD, most of them characterized as not potentially pathogenic in addition to being common and mostly in non-coding regions of the genome according to CADD and VEP, respectively (Figure 1). The fine-mapping of MD GVs identified 172 causal GVs and 95 pGenes (Supplementary Table S2). The functional enrichment analysis of pGenes stands along with hypotheses of MD pathogenesis such as alterations in neurogenesis and neuroplasticity or the circadian rhythm theory [31]. Additionally, these are also enriched for other phenotypes frequently co-occurring with MD, such as alterations of body mass index or smoking [32]. While most pGenes (63%) and GVs (58%) have previous evidence for association with MD, our study pinpoints novel pGenes and GVs (Supplementary Table S3 and S4). Additionally, existing literature supports the role of pGenes in processes related to MD pathogenesis, such as immune response, nervous system development, response to stress, or behavioral despair.

MD causally associated GVs are those most likely to be causal and functioning as eQTLs and, indeed, proved to be significantly enriched in cis-eQTLs from GTEx, in line with previous findings on MD and other psychiatric disorders [53,62]. The colocalizing eGenes are involved in processes relevant to MD, such as neuron projection [63], and have been associated with MD and related phenotypes according to DISGENET plus [47,48]. BAG5 is constitutively expressed in all tissues, while MYRF and SLC12A5 show higher levels in brain tissues (Supplementary Figure S2). BAG5 has been previously identified as associated with MD [64]. We characterize SLC12A5, involved in chloride homeostasis in neurons, as a pGene, also, and its downregulation has been described as an effect of stress leading to the activation of the hypothalamic–pituitary–adrenal axis, which ultimately can lead to MD-like symptoms [31,65]. However, rs12624433 is an eQTL in the Brain Putamen basal ganglia associated with the upregulation of SLC12A5. Thus, more research is needed to unravel the

exact mechanism by which rs12624433 exerts its role in the regulation of the expression of SLC12A5. This eGene has been described as a potential drug target for mental disorders, but considerations should be taken given its important role in brain development; besides, it is highly influenced by exercise and environmental factors [65]. rs198457 mediates the downregulation of MYRF expression, which plays a role in myelination and oligodendrocyte differentiation [46]. These, in turn, require thyroid hormones for their differentiation and maturation [66]. Furthermore, oligodendrocytes have been stated as crucial for psychological functions likely involved in mental disorders such as MD [67].

The analysis of TF regulation with RSAT enabled a precise prediction of TF binding alterations. Although TF expression is not highly tissue-specific [7,68], for this type of analysis, it is key to pick meaningful sets of TFs and GVs [69]. We focused on TF expressed in brain-related tissues as it has been previously reported that genes involved in depression are highly expressed in brain regions [4,32,37,47]. Our analysis resulted in the prediction of 270 GVs lying in active regulatory regions of the genome of brain-related tissues based on chromatin accessibility data. These GVs alter the binding of 101 TFs, roughly equally distributed as disrupting or creating a binding site. The activating or repressing role of these TFs is hard to interpret since it will always depend on the sequence context and the cofactors involved [68]. Thus, further analysis is required to elucidate the impact of these GVs on gene expression regulation. Our pipeline enabled us to filter and prioritize the large number of candidate GVs by combining different omics data and ultimately propose mechanistic hypotheses.

By using eQTL data, we were able to identify the GV rs12624433, which regulates the expression of SLC12A5. This GV, previously referred to as colocalizing, is predicted to alter the binding of the TFs USF1, USF2, and MYC; these belong to the bHLH family involved in neural development [70]. USF1 and USF2 generally exert activating effects [71], with USF1 being a risk gene for Alzheimer's disease and relevant for brain cholesterol metabolism involving its transport from astrocytes to neurons [72].

Additionally, we found mechanistic evidence for 2 GV-TF-eGene/target associations (rs11227217: RAD21 → ZNRD2-DT; rs62259947: YY1 → P4HTM) when combining pattern matching

results, chromatin accessibility data, GTEx eQTLs PICS, and hTFtarget data. Variant rs11227217 is more than 20 kbp away from ZNRD2-DT, but RAD21 is a member of the cohesion complex, which enables genes and enhancers to interact via loop formation [73,74]. NRD2-DT is a lncRNA, and interestingly, our findings include several lncRNAs in the set of pGenes as well as related with regulatory variations, either colocalizing with cis-eQTLS (RP11-894P9.2 and RP5-1115A15.1) or with mechanistic evidence for its association with gene expression regulation (ZNRD2-DT). Although not their exact role in MD pathophysiology is not clear, ncRNAs have been described as promising biomarkers and drug targets for MD [75,76].

Regarding the association rs62259947: YY1 → P4HTM, P4HTM has been related to neurological disorders and social behavior (Figure 3) [51,52]. It is involved in Ca2+ signaling mediated by the hypoxia-inducible factor HIF1α altering astrocytes gliotransmission [51]. Indeed, hypoxia has been associated with mental disorders in general and MD in particular [77,78,79,80]. In addition, P4HTM null mutation results in a reduction in fear and depression [52]. In turn, rs62259947 downregulates the expression of P4HTM and changes the binding affinity of YY1 in the Brain Cerebellar Hemisphere. Additionally, YY1 regulates transcription by forming loops, although its specific role as activator or repressor is not yet fully understood [50]. Furthermore, P4HTM and HIF1α have been reported as potential drug targets for MD [52,81]. rs11227217 and RAD21 are associated with red blood cell and reticulocyte count, respectively, by PheWAS [15]. Indeed, red blood cell parameters have been described as altered in patients with mental disorders [82].

5. Conclusions

Overall, we have successfully developed and applied a regulatory variation analysis pipeline including fine-mapping, colocalization, TF regulation analysis, and chromatin accessibility data, which overcomes the limitation of the lack of full-genome summary statistics. We have identified causal GVs, pGenes, and eGenes and proposed hypotheses for their role in MD pathogenesis, highlighting the role of chloride homeostasis and myelination. We also found mechanistic evidence involving hypoxia response mediated by altered TF binding. Our findings include GVs and genes supported by the literature on MD and mental disorders, as well as novel molecular mechanisms underlying MD pathogenesis.

Supplementary Materials

The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes13071259/s1, Scheme S1: Study overview; Figure S1: MD lead GVs characterization; Figure S2: Tissue expression of pGenes; Figure S3: Tissue expression of eGenes; Table S1: Summary of resources. Table S2: Causal GVs for MD; Table S3: pGenes functional and disease enrichment analysis; Table S4: Fine-mapped MD causal GVs disease enrichment analysis; Table S5: Colocalizing GWAS-eQTLs association to disease; Table S6: TFBS analysis; Table S7: GVs state annotation.

Author contributions

Conceptualization, J.P., A.M.-R. and L.I.F.; Data curation, J.P.-G.; Formal analysis, J.P.-G.; Funding acquisition, A.M.-R. and L.I.F.; Supervision, J.P., A.M.-R. and L.I.F.; Validation, J.P.-G.; Visualization, J.P.-G.; Writing—original draft, J.P.-G.; Writing—review and editing, J.P.-G., J.P., A.M.-R. and L.I.F. All authors have read and agreed to the published version of the manuscript.

Funding

IMI2-JU resources which are composed of financial contributions from the European Union's Horizon 2020 Research and Innovation Programme and EFPIA [GA: 116030 TransQST and GA: 777365 eTRANSAFE], and the EU H2020 Programme 2014–2020 [GA: 676559 Elixir-Excelerate]; Project 001-P-001647—Valorisation of EGA for Industry and Society funded by the European Regional Development Fund (ERDF) and Generalitat de Catalunya; Agència de Gestió d'Ajuts Universitaris i de Recerca Generalitat de Catalunya [2017SGR00519], and the Institute of Health Carlos III (project IMPaCT-Data, exp. IMP/00019), co-funded by the European Union, European Regional Development Fund (ERDF, "A way to make Europe"). The Research Programme on Biomedical Informatics (GRIB) is a member of the Spanish National Bioinformatics Institute (INB), funded by ISCIII and ERDF (PRB2-ISCIII [PT13/0001/0023, of the PE I + D + i 2013–2016]). The MELIS is a 'Unidad de Excelencia María de Maeztu', funded by the MINECO [MDM-2014-0370]. AMR was supported by CONACYT-FORDECYT-PRONACES grant no. [11311], and Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica–Universidad Nacional Autónoma de México (PAPIIT-UNAM) grant nos. IA203021. JPG was supported by

Instituto de Salud Carlos III-Fondo Social Europeo [FI18/00034]; Instituto de Salud Carlos III [MV20]. This work reflects only the author's view and that the IMI2-JU is not responsible for any use that may be made of the information it contains.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

This study analyzed data generated by other projects, which are publicly available as specified in the Methods and Results sections of this paper and summarized here. The GWAS data is available at the GWAS Catalog repository, Accession Study: GCST007342 and PICS Data Portal, https://pics2.ucsf.edu/Downloads/PICS2-GWAScat-2021-10-29.txt.gz (accessed on 1 November 2021) and https://pics2.ucsf.edu/Downloads/GTEx/ (accessed on 2 November 2021). The GTEx RNA-Seq data can be downloaded from https://www.gtexportal.org/home/datasets (accessed on 18 February 2021) (filename: GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz); ENCODE data can be accessed from the following Accession Numbers: ENCSR674KAN, ENCSR801APH, ENCSR826BFW, ENCSR658SFK, ENCSR082KYZ, ENCSR363VGK, ENCSR738WFF and ENCSR860PXK. The data from JASPAR, ENCODE, HOCOMOCO, footprintDB and hPDI is available at (http://rsat.sb-roscoff.fr/retrieve-matrix_form.cgi (accessed on 24 February, 2 and 4 March 2021), see View matrix descriptions and download full collections); and hTFtarget data can be downloaded from http://bioinfo.life.hust.edu.cn/hTFtarget#!/download (accessed on 15 November 2021) (filename: TF-Target-information.txt). The data generated by the current study as a result of the analysis of the above-mentioned datasets are available at Zenodo (https://doi.org/10.5281/zenodo.6838470) and are also available in the Supplementary Tables File with this manuscript.

Acknowledgments

This work received support from Luis Aguilar, Alejandro León, and Jair García of the Laboratorio Nacional de Visualización Científica Avanzada. We thank Carina Uribe Díaz, and Alejandra Castillo Carbajal for their technical support.

Conflicts of Interest

L.I.F. and J.P. are co-founders and hold shares of Medbioinformatics Solutions SL. J.P.G. and A.M.R. have no competing interests.

Abbreviations

CADD	Combined Annotation Dependent Depletion
CLPP	colocalization posterior probability
eGenes	genes regulated by eQTLs
ENCODE	Encyclopedia of DNA Elements
eQTLs	expression quantitative trait loci
GO	Gene Ontology
GV	genetic variant
GTEx	Genotype-Tissue Expression
GWAS	genome-wide association studies
LD	linkage disequilibrium
MD	major depression pGenes proximal genes
PICS	Probabilistic Identification of Causal SNPs
PSSM	Position-specific scoring matrix or position weight matrix
REAC	Reactome
RSAT	Regulatory Sequence Analysis Tools
TF	transcription factor
TPM	transcripts per million
TFBS	transcription factor binding site
VEP	variant effect predictor

References

1. World Health Organization: Depression. Available online: https://www.who.int/news-room/fact-sheets/detail/depression (accessed on 21 December 2021).
2. Preskorn, S.H. Drug Development in Psychiatry: The Long and Winding Road from Chance Discovery to Rational Development. In Handbook of Experimental Pharmacology; Springer:

Berlin/Heidelberg, Germany, 2018; Volume 250, pp. 307–324. [Google Scholar] [CrossRef]
3. Howard, D.M.; Adams, M.J.; Clarke, T.-K.; Hafferty, J.D.; Gibson, J.; Shirali, M.; Coleman, J.R.I.; Hagenaars, S.P.; Ward, J.; Wigmore, E.M.; et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat. Neurosci. 2019, 22, 343–352. [Google Scholar] [CrossRef] [PubMed][Green Version]
4. Wray, N.R.; Ripke, S.; Mattheisen, M.; Trzaskowski, M.; Byrne, E.M.; Abdellaoui, A.; Adams, M.J.; Agerbo, E.; Air, T.M.; Andlauer, T.M.F.; et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat. Genet. 2018, 50, 668–681. [Google Scholar] [CrossRef] [PubMed][Green Version]
5. Cano-Gamez, E.; Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. Front. Genet. 2020, 11, 424. [Google Scholar] [CrossRef] [PubMed]
6. Umans, B.D.; Battle, A.; Gilad, Y. Where Are the Disease-Associated eQTLs? Trends Genet. 2020, 37, 109–124. [Google Scholar] [CrossRef] [PubMed]
7. Hu, H.; Miao, Y.-R.; Jia, L.-H.; Yu, Q.-Y.; Zhang, Q.; Guo, A.-Y. AnimalTFDB 3.0: A comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res. 2018, 47, D33–D38. [Google Scholar] [CrossRef] [PubMed]
8. Perdomo-Sabogal, Á.; Nowick, K. Genetic Variation in Human Gene Regulatory Factors Uncovers Regulatory Roles in Local Adaptation and Disease. Genome Biol. Evol. 2019, 11, 2178–2193. [Google Scholar] [CrossRef]
9. Buniello, A.; MacArthur, J.A.L.; Cerezo, M.; Harris, L.W.; Hayhurst, J.; Malangone, C.; McMahon, A.; Morales, J.; Mountjoy, E.; Sollis, E.; et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019, 47, D1005–D1012. [Google Scholar] [CrossRef][Green Version]
10. Taylor, K.E.; Ansel, K.M.; Marson, A.; Criswell, L.A.; Farh, K.K.-H. PICS2: Next-generation fine mapping via probabilistic identification of causal SNPs. Bioinformatics 2021, 37, 3004–3007. [Google Scholar] [CrossRef]
11. McLaren, W.; Gil, L.; Hunt, S.E.; Riat, H.S.; Ritchie, G.R.S.; Thormann, A.; Flicek, P.; Cunningham, F. The Ensembl Variant

Effect Predictor. Genome Biol. 2016, 17, 122. [Google Scholar] [CrossRef][Green Version]
12. Rentzsch, P.; Witten, D.; Cooper, G.M.; Shendure, J.; Kircher, M. CADD: Predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019, 47, D886–D894. [Google Scholar] [CrossRef]
13. The ENCODE Project Consortium. An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature 2012, 489, 57–74. [Google Scholar] [CrossRef] [PubMed]
14. Oscanoa, J.; Sivapalan, L.; Gadaleta, E.; Ullah, A.Z.D.; Lemoine, N.R.; Chelala, C. SNPnexus: A web server for functional annotation of human genome sequence variation (2020 update). Nucleic Acids Res. 2020, 48, W185–W192. [Google Scholar] [CrossRef] [PubMed]
15. Ghoussaini, M.; Mountjoy, E.; Carmona, M.; Peat, G.; Schmidt, E.M.; Hercules, A.; Fumis, L.; Miranda, A.; Carvalho-Silva, D.; Buniello, A.; et al. Open Targets Genetics: Systematic identification of trait-associated genes using large-scale genetics and functional genomics. Nucleic Acids Res. 2020, 49, D1311–D1320. [Google Scholar] [CrossRef] [PubMed]
16. Hormozdiari, F.; van de Bunt, M.; Segrè, A.V.; Li, X.; Joo, J.W.J.; Bilow, M.; Sul, J.H.; Sankararaman, S.; Pasaniuc, B.; Eskin, E. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am. J. Hum. Genet. 2016, 99, 1245–1260. [Google Scholar] [CrossRef] [PubMed][Green Version]
17. Landt, S.G.; Marinov, G.K.; Kundaje, A.; Kheradpour, P.; Pauli, F.; Batzoglou, S.; Bernstein, B.E.; Bickel, P.; Brown, J.B.; Cayting, P. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012, 22, 1813–1831. [Google Scholar] [CrossRef] [PubMed][Green Version]
18. Jayaram, N.; Usvyat, D.; Martin, A.C.R. Evaluating tools for transcription factor binding site prediction. BMC Bioinform. 2016, 17, 547. [Google Scholar] [CrossRef] [PubMed][Green Version]
19. GTEx Portal. Available online: https://www.gtexportal.org/home/datasets (accessed on 22 December 2021).
20. Fornes, O.; Castro-Mondragon, J.A.; Khan, A.; Van Der Lee, R.; Zhang, X.; Richmond, P.A.; Modi, B.P.; Correard, S.; Gheorghe, M.; Baranašić, D.; et al. JASPAR 2020: Update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2020, 48, D87–D92. [Google Scholar] [CrossRef] [PubMed]

21. Kulakovskiy, I.V.; Vorontsov, I.E.; Yevshin, I.S.; Sharipov, R.N.; Fedorova, A.D.; Rumynskiy, E.I.; Medvedeva, Y.A.; Magana-Mora, A.; Bajic, V.B.; Papatsenko, D.A.; et al. HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 2017, 46, D252–D259. [Google Scholar] [CrossRef] [PubMed]
22. Sebastian, A.; Contreras-Moreira, B. footprintDB: A database of transcription factors with annotated cis elements and binding interfaces. Bioinformatics 2013, 30, 258–265. [Google Scholar] [CrossRef] [PubMed]
23. Xie, Z.; Hu, S.; Blackshaw, S.; Zhu, H.; Qian, J. hPDI: A database of experimental human protein-DNA interactions. Bioinformatics 2009, 26, 287–289. [Google Scholar] [CrossRef] [PubMed][Green Version]
24. Huang, D.; Wang, Z.; Zhou, Y.; Liang, Q.; Sham, P.C.; Yao, H.; Li, M.J. vSampler: Fast and annotation-based matched variant sampling tool. Bioinformatics 2020, 37, 1915–1917. [Google Scholar] [CrossRef] [PubMed]
25. Ernst, J.; Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. Nat. Protoc. 2017, 12, 2478–2492. [Google Scholar] [CrossRef] [PubMed]
26. Annotation of the non-coding genome. Nature 2015. [CrossRef][Green Version]
27. Zhang, Q.; Liu, W.; Zhang, H.-M.; Xie, G.-Y.; Miao, Y.-R.; Xia, M.; Guo, A.-Y. hTFtarget: A Comprehensive Database for Regulations of Human Transcription Factors and Their Targets. Genom. Proteom. Bioinform. 2020, 18, 120–128. [Google Scholar] [CrossRef] [PubMed]
28. Raudvere, U.; Kolberg, L.; Kuzmin, I.; Arak, T.; Adler, P.; Peterson, H.; Vilo, J. g: Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019, 47, W191–W198. [Google Scholar] [CrossRef] [PubMed][Green Version]
29. Piñero, J.; Ramírez-Anguita, J.M.; Saüch-Pitarch, J.; Ronzano, F.; Centeno, E.; Sanz, F.; Furlong, L.I. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2019, 48, D845–D855. [Google Scholar] [CrossRef] [PubMed][Green Version]
30. Disgenet Plus. Available online: https://beta.disgenetplus.com/ (accessed on 21 December 2021).

31. Shadrina, M.; Bondarenko, E.A.; Slominsky, P.A. Genetics Factors in Major Depression Disease. Front. Psychiatry 2018, 9, 334. [Google Scholar] [CrossRef][Green Version]
32. McIntosh, A.M.; Sullivan, P.F.; Lewis, C.M. Uncovering the Genetic Architecture of Major Depression. Neuron 2019, 102, 91–103. [Google Scholar] [CrossRef]
33. Zhang, K.; Lin, W.; Zhang, J.; Zhao, Y.; Wang, X.; Zhao, M. Effect of Toll-like receptor 4 on depressive-like behaviors induced by chronic social defeat stress. Brain Behav. 2020, 10, e01525. [Google Scholar] [CrossRef]
34. Keyes, K.; Agnew-Blais, J.; Roberts, A.L.; Hamilton, A.; De Vivo, I.; Ranu, H.; Koenen, K. The role of allelic variation in estrogen receptor genes and major depression in the Nurses Health Study. Soc. Psychiatry 2015, 50, 1893–1904. [Google Scholar] [CrossRef][Green Version]
35. Mossakowska-Wójcik, J.; Orzechowska, A.; Talarowska, M.; Szemraj, J.; Galecki, P. The importance of TCF4 gene in the etiology of recurrent depressive disorders. Prog. Neuro-Psychopharmacol. Biol. Psychiatry 2018, 80, 304–308. [Google Scholar] [CrossRef] [PubMed]
36. Berrio, A.T.; Lopez, J.P.; Bagot, R.C.; Nouel, D.; Bo, G.D.; Cuesta, S.; Zhu, L.; Manitt, C.; Eng, C.; Cooper, H.M.; et al. DCC Confers Susceptibility to Depression-like Behaviors in Humans and Mice and Is Regulated by miR-218. Biol. Psychiatry 2016, 81, 306–315. [Google Scholar] [CrossRef] [PubMed][Green Version]
37. Hyde, C.L.; Nagle, M.W.; Tian, C.; Chen, X.; Paciga, S.A.; Wendland, J.R.; Tung, J.Y.; Hinds, D.A.; Perlis, R.H.; Winslow, A.R. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. Nat. Genet. 2016, 48, 1031–1036. [Google Scholar] [CrossRef] [PubMed]
38. Wu, Q.; Tang, W.; Luo, Z.; Li, Y.; Shu, Y.; Yue, Z.; Xiao, B.; Feng, L. DISC1 Regulates the Proliferation and Migration of Mouse Neural Stem/Progenitor Cells through Pax5, Sox2, Dll1 and Neurog2. Front. Cell. Neurosci. 2017, 11, 261. [Google Scholar] [CrossRef] [PubMed][Green Version]
39. Rudzinskas, S.; Hoffman, J.F.; Martinez, P.; Rubinow, D.R.; Schmidt, P.J.; Goldman, D. In vitro model of perimenopausal depression implicates steroid metabolic and proinflammatory genes. Mol. Psychiatry 2020, 26, 3266–3276. [Google Scholar] [CrossRef]

40. Qin, Z.; Ren, F.; Xu, X.; Ren, Y.; Li, H.; Wang, Y.; Zhai, Y.; Chang, Z. ZNF536, a Novel Zinc Finger Protein Specifically Expressed in the Brain, Negatively Regulates Neuron Differentiation by Repressing Retinoic Acid-Induced Gene Transcription. Mol. Cell. Biol. 2009, 29, 3633–3643. [Google Scholar] [CrossRef][Green Version]
41. Laifenfeld, D.; Klein, E.; Ben-Shachar, D. Norepinephrine alters the expression of genes involved in neuronal sprouting and differentiation: Relevance for major depression and antidepressant mechanisms. J. Neurochem. 2002, 83, 1054–1064. [Google Scholar] [CrossRef]
42. Lanshakov, D.A.; Sukhareva, E.V.; Bulygina, V.V.; Bannova, A.V.; Shaburova, E.V.; Kalinina, T.S. Single neonatal dexamethasone administration has long-lasting outcome on depressive-like behaviour, Bdnf, Nt-3, p75ngfr and sorting receptors (SorCS1-3) stress reactive expression. Sci. Rep. 2021, 11, 8092. [Google Scholar] [CrossRef]
43. Sanna, M.D.; Quattrone, A.; Galeotti, N. Antidepressant-like actions by silencing of neuronal ELAV-like RNA-binding proteins HuB and HuC in a model of depression in male mice. Neuropharmacology 2018, 135, 444–454. [Google Scholar] [CrossRef]
44. Beilina, A.; Rudenko, I.N.; Kaganovich, A.; Civiero, L.; Chau, H.; Kalia, S.K.; Kalia, L.V.; Lobbestael, E.; Chia, R.; Ndukwe, K.; et al. Unbiased screen for interactors of leucine-rich repeat kinase 2 supports a common pathway for sporadic and familial Parkinson disease. Proc. Natl. Acad. Sci. USA 2014, 111, 2626–2631. [Google Scholar] [CrossRef][Green Version]
45. Dzhala, V.I.; Talos, D.M.; Sdrulla, D.A.; Brumback, A.; Mathews, G.C.; Benke, T.; Delpire, E.; Jensen, F.E.; Staley, K.J. NKCC1 transporter facilitates seizures in the developing brain. Nat. Med. 2005, 11, 1205–1213. [Google Scholar] [CrossRef] [PubMed]
46. Bujalka, H.; Koenning, M.; Jackson, S.; Perreau, V.M.; Pope, B.; Hay, C.M.; Mitew, S.; Hill, A.F.; Lu, Q.R.; Wegner, M.; et al. MYRF Is a Membrane-Associated Transcription Factor That Autoproteolytically Cleaves to Directly Activate Myelin Genes. PLoS Biol. 2013, 11, e1001625. [Google Scholar] [CrossRef] [PubMed][Green Version]
47. Cross-Disorder Group of the Psychiatric Genomics Consortium. Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms

across Eight Psychiatric Disorders. Cell 2019, 179, 1469–1482.e11. [Google Scholar] [CrossRef] [PubMed][Green Version]
48. Yao, X.; Glessner, J.T.; Li, J.; Qi, X.; Hou, X.; Zhu, C.; Li, X.; March, M.E.; Yang, L.; Mentch, F.D.; et al. Integrative analysis of genome-wide association studies identifies novel loci associated with neuropsychiatric disorders. Transl. Psychiatry 2021, 11, 69. [Google Scholar] [CrossRef] [PubMed]
49. Garcia, W.S.; Rocha-Acevedo, M.; Ramirez-Navarro, L.; Mbouamboua, Y.; Thieffry, D.; Thomas-Chollier, M.; Contreras-Moreira, B.; van Helden, J.; Medina-Rivera, A. RSAT variation-tools: An accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding. Comput. Struct. Biotechnol. J. 2019, 17, 1415–1428. [Google Scholar] [CrossRef] [PubMed]
50. Verheul, T.C.J.; Van Hijfte, L.; Perenthaler, E.; Barakat, T.S. The Why of YY1: Mechanisms of Transcriptional Regulation by Yin Yang 1. Front. Cell Dev. Biol. 2020, 8, 592164. [Google Scholar] [CrossRef]
51. Byts, N.; Sharma, S.; Laurila, J.; Paudel, P.; Miinalainen, I.; Ronkainen, V.-P.; Hinttala, R.; Törnquist, K.; Koivunen, P.; Myllyharju, J. Transmembrane Prolyl 4-Hydroxylase is a Novel Regulator of Calcium Signaling in Astrocytes. ENeuro 2020, 8, 1–23. [Google Scholar] [CrossRef]
52. Leinonen, H.; Koivisto, H.; Lipponen, H.-R.; Matilainen, A.; Salo, A.M.; Dimova, E.Y.; Hämäläinen, E.; Stavén, S.; Miettinen, P.; Myllyharju, J.; et al. Null mutation in P4h-tm leads to decreased fear and anxiety and increased social behavior in mice. Neuropharmacology 2019, 153, 63–72. [Google Scholar] [CrossRef]
53. Bhalala, O.G.; Nath, A.P.; Inouye, M.; Sibley, C.R. UK Brain Expression Consortium Identification of expression quantitative trait loci associated with schizophrenia and affective disorders in normal brain tissue. PLoS Genet. 2018, 14, e1007607. [Google Scholar] [CrossRef]
54. Li, S.; Li, X.; Liu, J.; Huo, Y.; Li, L.; Wang, J.; Luo, X.-J. Functional variants fine-mapping and gene function characterization provide insights into the role of ZNF323 in schizophrenia pathogenesis. Am. J. Med. Genet. Part B Neuropsychiatr. Genet. 2021, 186, 28–39. [Google Scholar] [CrossRef]
55. Roksana, Z. Transcription Factors in Schizophrenia: A Current View of Genetic Aspects. Sci. J. Genet. Gene Ther. 2016, 2, 17–21. [Google Scholar] [CrossRef][Green Version]

56. Li, X.; Su, X.; Liu, J.; Li, H.; Li, M.; Li, W.; Luo, X.-J. Transcriptome-wide association study identifies new susceptibility genes and pathways for depression. Transl. Psychiatry 2021, 11, 306. [Google Scholar] [CrossRef] [PubMed]
57. Zhong, J.; Li, S.; Zeng, W.; Li, X.; Gu, C.; Liu, J.; Luo, X.-J. Integration of GWAS and brain eQTL identifies FLOT1 as a risk gene for major depressive disorder. Neuropsychopharmacology 2019, 44, 1542–1551. [Google Scholar] [CrossRef]
58. Santos-Terra, J.; Deckmann, I.; Fontes-Dutra, M.; Schwingel, G.B.; Bambini-Junior, V.; Gottfried, C. Transcription factors in neurodevelopmental and associated psychiatric disorders: A potential convergence for genetic and environmental risk factors. Int. J. Dev. Neurosci. 2021, 81, 545–578. [Google Scholar] [CrossRef] [PubMed]
59. Burt, C.; Munafò, M. Has GWAS lost its status as a paragon of open science? PLoS Biol. 2021, 19, e3001242. [Google Scholar] [CrossRef] [PubMed]
60. Lee, B.; Yao, X.; Shen, L. Integrative analysis of summary data from GWAS and eQTL studies implicates genes differentially expressed in Alzheimer's disease. BMC Genom. 2022, 23, 414. [Google Scholar] [CrossRef] [PubMed]
61. Brooks-Warburton, J.; Modos, D.; Sudhakar, P.; Madgwick, M.; Thomas, J.P.; Bohar, B.; Fazekas, D.; Zoufir, A.; Kapuy, O.; Szalay-Beko, M.; et al. A systems genomics approach to uncover patient-specific pathogenic pathways and proteins in ulcerative colitis. Nat. Commun. 2022, 13, 2299. [Google Scholar] [CrossRef]
62. O'Brien, H.E.; Hannon, E.; Hill, M.J.; Toste, C.C.; Robertson, M.J.; Morgan, J.E.; McLaughlin, G.; Lewis, C.M.; Schalkwyk, L.C.; Hall, L.S.; et al. Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. Genome Biol. 2018, 19, 194. [Google Scholar] [CrossRef]
63. Hare, B.D.; Duman, R.S. Prefrontal cortex circuits in depression and anxiety: Contribution of discrete neuronal populations and target regions. Mol. Psychiatry 2020, 25, 2742–2758. [Google Scholar] [CrossRef]
64. Amare, A.T.; Vaez, A.; Hsu, Y.-H.; Direk, N.; Kamali, Z.; Howard, D.; McIntosh, A.; Tiemeier, H.; Bültmann, U.; Snieder, H.; et al. Bivariate genome-wide association analyses of the broad depression phenotype combined with major depressive disorder, bipolar disorder or schizophrenia reveal eight novel genetic loci for

depression. Mol. Psychiatry 2019, 25, 1420–1429. [Google Scholar] [CrossRef]
65. Pozzi, D.; Rasile, M.; Corradini, I.; Matteoli, M. Environmental regulation of the chloride transporter KCC2: Switching inflammation off to switch the GABA on? Transl. Psychiatry 2020, 10, 349. [Google Scholar] [CrossRef] [PubMed]
66. Saponaro, F.; Sestito, S.; Runfola, M.; Rapposelli, S.; Chiellini, G. Selective Thyroid Hormone Receptor-Beta (TRβ) Agonists: New Perspectives for the Treatment of Metabolic and Neurodegenerative Disorders. Front. Med. 2020, 7, 331. [Google Scholar] [CrossRef] [PubMed]
67. Zhou, B.; Zhu, Z.; Ransom, B.R.; Tong, X. Oligodendrocyte lineage cells and depression. Mol. Psychiatry 2020, 26, 103–117. [Google Scholar] [CrossRef] [PubMed]
68. Lambert, S.A.; Jolma, A.; Campitelli, L.F.; Das, P.K.; Yin, Y.; Albu, M.; Chen, X.; Taipale, J.; Hughes, T.R.; Weirauch, M.T. The Human Transcription Factors. Cell 2018, 172, 650–665. [Google Scholar] [CrossRef] [PubMed][Green Version]
69. Andersen, M.C.; Engström, P.G.; Lithwick, S.; Arenillas, D.; Eriksson, P.; Lenhard, B.; Wasserman, W.W.; Odeberg, J. In Silico Detection of Sequence Variations Modifying Transcriptional Regulation. PLoS Comput. Biol. 2008, 4, e5. [Google Scholar] [CrossRef] [PubMed][Green Version]
70. Dennis, D.J.; Han, S.; Schuurmans, C. bHLH transcription factors in neural development, disease, and reprogramming. Brain Res. 2019, 1705, 48–65. [Google Scholar] [CrossRef]
71. Rada-Iglesias, A.; Ameur, A.; Kapranov, P.; Enroth, S.; Komorowski, J.; Gingeras, T.R.; Wadelius, C. Whole-genome maps of USF1 and USF2 binding and histone H3 acetylation reveal new aspects of promoter structure and candidate genes for common human disorders. Genome Res. 2008, 18, 380–392. [Google Scholar] [CrossRef][Green Version]
72. Sertbaş, M.; Ülgen, K.; Çakır, T. Systematic analysis of transcription-level effects of neurodegenerative diseases on human brain metabolism by a newly reconstructed brain-specific metabolic network. FEBS Open Bio 2014, 4, 542–553. [Google Scholar] [CrossRef][Green Version]
73. Grubert, F.; Srivas, R.; Spacek, D.V.; Kasowski, M.; Ruiz-Velasco, M.; Sinnott-Armstrong, N.; Greenside, P.; Narasimha, A.; Liu, Q.; Geller, B.; et al. Landscape of cohesin-mediated chromatin loops in

the human genome. Nature 2020, 583, 737–743. [Google Scholar] [CrossRef]
74. Brodie, A.; Azaria, J.R.; Ofran, Y. How far from the SNP may the causative genes be? Nucleic Acids Res. 2016, 44, 6046–6054. [Google Scholar] [CrossRef]
75. Shi, Y.; Wang, Q.; Song, R.; Kong, Y.; Zhang, Z. Non-coding RNAs in depression: Promising diagnostic and therapeutic biomarkers. EBioMedicine 2021, 71, 103569. [Google Scholar] [CrossRef] [PubMed]
76. Żurawek, D.; Turecki, G. The miRNome of Depression. Int. J. Mol. Sci. 2021, 22, 11312. [Google Scholar] [CrossRef] [PubMed]
77. Bian, Z.; Li, H.; Liu, Y.; Cao, Y.; Kang, Y.; Yu, Y.; Zhang, F.; Li, C.; Kang, Y.; Wang, F. The Association Between Hypoxia Improvement and Electroconvulsive Therapy for Major Depressive Disorder. Neuropsychiatr. Dis. Treat. 2021, 17, 2987–2994. [Google Scholar] [CrossRef] [PubMed]
78. Li, G.; Zhao, M.; Cheng, X.; Zhao, T.; Feng, Z.; Zhao, Y.; Fan, M.; Zhu, L. FG-4592 Improves Depressive-Like Behaviors through HIF-1-Mediated Neurogenesis and Synapse Plasticity in Rats. Neurotherapeutics 2019, 17, 664–675. [Google Scholar] [CrossRef] [PubMed]
79. Ding, F.-S.; Cheng, X.; Zhao, T.; Zhao, Y.; Zhang, G.-B.; Wu, H.-T.; Zhu, L.-L.; Wu, K.-W. Intermittent hypoxic preconditioning relieves fear and anxiety behavior in post-traumatic stress model mice. Sheng Li Xue Bao 2019, 71, 537–546. [Google Scholar]
80. Shibata, T.; Yamagata, H.; Uchida, S.; Otsuki, K.; Hobara, T.; Higuchi, F.; Abe, N.; Watanabe, Y. The alteration of hypoxia inducible factor-1 (HIF-1) and its target genes in mood disorder patients. Prog. Neuro-Psychopharmacol. Biol. Psychiatry 2013, 43, 222–229. [Google Scholar] [CrossRef] [PubMed]
81. Kang, I.; Kondo, D.; Kim, J.; Lyoo, I.K.; Yurgelun-Todd, D.; Hwang, J.; Renshaw, P.F. Elevating the level of hypoxia inducible factor may be a new potential target for the treatment of depression. Med. Hypotheses 2020, 146, 110398. [Google Scholar] [CrossRef]
82. Szczepocka, E.; Wysokiński, A. Red Blood Cells Parameters in Patients with Acute Schizophrenia, Unipolar Depression and Bipolar Disorder. Psychiatr. Danub. 2018, 30, 323–330. [Google Scholar] [CrossRef]

3.3. Benchmarking post-GWAS analysis tools: challenges and implications

Despite the numerous tools available for uncovering the role of GVs in disease pathogenesis, there is no guidelines for method selection or validation. In this chapter, we design and apply a workflow using a MD GWAS dataset and eQTL data to compare the outcomes of different fine-mapping and colocalisation tools and their biological implications. We identify fine-mapping as a key step in the post-GWAS analysis, with implications for subsequent steps, which result in different causal GVs and eGenes involved in various biological processes. We also evaluate the assumptions of fine-mapping and colocalisation methods in the context of the results obtained. Finally, we highlight the need for an objective and unbiased assessment of post-GWAS analysis tools to leverage GWAS data to support precision medicine applications.

Pérez-Granado, J., Piñero, J. & Furlong, L. I. Benchmarking post-GWAS analysis tools in major depression: Challenges and implications. *Front Genet* **13**, 2853 (2022).

Benchmarking post-GWAS analysis tools in major depression: Challenges and implications

Judith Pérez-Granado[1], Janet Piñero[1,2] and Laura I. Furlong[1,2*]

6. Research Programme on Biomedical Informatics (GRIB), Hospital Del Mar Medical Research Institute (IMIM), Department of Medicine and Life Sciences (MELIS), Universitat Pompeu Fabra (UPF), Barcelona, Spain
7. MedBioinformatics Solutions SL, Barcelona, Spain

[*]Correspondence: Laura I. Furlong, laura.furlong@upf.edu

Abstract

Our knowledge of complex disorders has increased in the last years thanks to the identification of genetic variants (GVs) significantly associated with disease phenotypes by genome-wide association studies (GWAS). However, we do not understand yet how these GVs functionally impact disease pathogenesis or their underlying biological mechanisms. Among the multiple post-GWAS methods available, fine-mapping and colocalization approaches are commonly used to identify causal GVs, meaning those with a biological effect on the trait, and their functional effects. Despite the variety of post-GWAS tools available, there is no guideline for method eligibility or validity, even though these methods work under different assumptions when accounting for linkage disequilibrium and integrating molecular annotation data. Moreover, there is no benchmarking of the available tools. In this context, we have applied two different fine-mapping and colocalization methods to the same GWAS on major depression (MD) and expression quantitative trait loci (eQTL) datasets. Our goal is to perform a systematic comparison of the results obtained by the different tools. To that end, we have evaluated their results at different levels: fine-mapped and colocalizing GVs, their target genes and tissue specificity according to gene expression information, as well as the biological processes in which they are involved. Our findings highlight the importance of fine-mapping as a key step for subsequent analysis. Notably, the colocalizing variants, altered genes and targeted tissues differed between methods, even regarding their biological implications. This contribution illustrates an important issue in post-GWAS analysis with relevant consequences on the use of GWAS results for elucidation of disease pathobiology, drug target prioritization and biomarker discovery.

Keywords: fine-mapping, colocalization, post-GWAS, major depression, eQTLs

1. Introduction

More than 207,400 genetic variants (GVs) have been associated with complex diseases since the introduction of genome-wide association studies (GWAS) (Dehghan, 2018; Buniello et al., 2019). The vast majority of identified GVs lie in non-coding regions of the genome with no clear impact on gene function and disease pathogenesis (Brandes et al., 2022), posing challenges in interpreting the association of the GV with the disease phenotype. Furthermore, these GVs may not be the causal ones but may be in linkage disequilibrium (LD) with the true causal GVs (Visscher et al., 2017; Brandes et al., 2022). We refer to causal GVs to those with a biological impact. A variety of approaches are available to unravel the functional role of GVs identified by GWAS (Kichaev et al., 2014; Amlie-Wolf et al., 2018; Wallace, 2021; Gazal et al., 2022). In addition, there are a plethora of different tools available that serve the same purpose but work with different types of data (e.g., genotype data versus full genome summary statistics), under different assumptions (e.g., one causal GV or more), and with diverse outcomes (e.g., causal GVs or relevant gene-cell type combination) (Cano-Gamez and Trynka, 2020; Adebiyi et al., 2021). There is, however, no guideline for determining which tool is best to use for each approach nor a gold standard for evaluating the validity of the results. Furthermore, in contrast to other areas where benchmarking evaluations of methods are in place, such as for protein structure prediction (Protein Structure Prediction Center, 2020) or disease module identification (Dream Challenges, 2022), among others, methods for GWAS data analysis have not been objectively benchmarked. Selecting the right tool is critical in post-GWAS analysis, to properly unravel the functional mechanisms by which the GVs lead to disease, and where different performances can lead to different results (Wen et al., 2017; Rüeger et al., 2018; LaPierre et al., 2021).

There is an absence of a benchmark dataset to assess the performance of post-GWAS analysis tools. Therefore, we propose a systematic and objective comparison of the results obtained by different tools when applied to the same datasets. We designed a fine-mapping and colocalization workflow with different tools running alternatively. Fine-mapping analysis identifies the causal GVs and is a necessary step in most post-GWAS analyses. We used the tools Probabilistic

Identification of Causal SNPs (PICS) (Taylor et al., 2021) and TORUS (Wen, 2016) as alternative tools for fine-mapping. Colocalization methods pinpoint the GVs causally associated with a phenotype and a molecular trait of interest, such as expression or methylation. We focused our analysis on expression quantitative trait loci (eQTL), to identify GVs with an effect on the expression of genes, from now on referred to as eGenes. We applied two methods for colocalization analysis: the Colocalization Posterior Probability (CLPP) approach (Hormozdiari et al., 2016) and the Fast Enrichment Estimation Aided Colocalization analysis (fastENLOC) (Pividori et al., 2020; Hukku et al., 2021). We applied the fine-mapping and colocalization workflow to the same GWAS on major depression (MD) and eQTL datasets (Figure 1). The results obtained with each combination of tools were evaluated in terms of fine-mapped and colocalizing GVs, the retrieved eGenes, the tissues in which this regulation of gene expression might take place, as well as the biological processes in which these genes are involved.

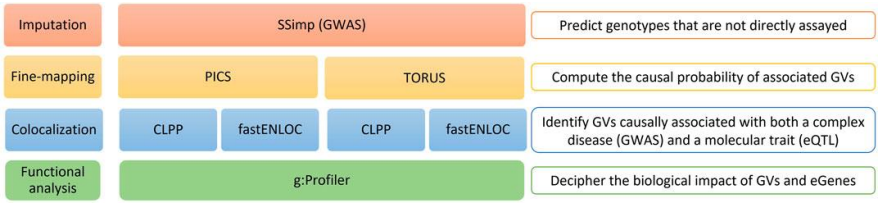


Figure 1. Overview of the study workflow. Schematic representation of the entire analysis workflow: 1) SSimp Imputation; 2) Alternative fine-mapping with PICS and TORUS; 3) Alternative colocalization analysis with CLPP and fastENLOC to both PICS and TORUS fine-mapping results; and 4) Functional analysis of the GVs and eGenes obtained at the end of the workflow. SSimp, Summary Statistics Imputation software; GWAS, genome-wide association studies; PICS, Probabilistic Identification of Causal SNPs; CLPP, Colocalization Posterior Probability; fastENLOC, Fast Enrichment Estimation Aided Colocalization Analysis; GVs, genetic variants; eGenes, genes regulated by expression quantitative trait loci.

The results of the workflow reveal divergence across tools, pinpointing a relevant issue in post-GWAS analysis derived from the lack of method benchmarking. Our findings demonstrate how critical is the fine-mapping step to subsequent analysis and how colocalization outcomes are in turn highly impacted by the assumptions of each tool. As a consequence, the causal GVs and eGenes identified are different and are involved in different biological processes. Overall, given the lack of agreement among tools, we highlight the need for an objective and unbiased assessment of post-GWAS analysis methods and tools to properly leverage GWAS data.

2. Materials and methods

Among the plethora of available methods for post-GWAS analysis and which have been reviewed elsewhere (Cano-Gamez and Trynka, 2020; Adebiyi et al., 2021), we focused on fine-mapping and colocalization. Then, we conducted a tool selection based on: workability with full-genome summary statistics, documentation quality, software maturity and developer support availability.

The workflow we describe in this manuscript applies alternative tools for post-GWAS analysis to compare their outcomes (Figure 1). We begin with an imputation step, followed by a fine-mapping and colocalization analysis using two different tools for each of these processes, and finish with a functional analysis of the results obtained using different tools and databases. We present below a more detailed explanation of each step.

2.1. GWAS dataset and imputation

We have selected the latest genome-wide association study (GWAS) on major depression (MD) with publicly available full-genome summary statistics (GCST005902) (Howard et al., 2018). This GWAS evaluated 7,666,894 genetic variants (GVs) in 322,580 European participants (113,769 cases and 208,811 controls). We used the harmonized version of this GWAS dataset. This implies the genomic position is reported against the latest genome build (GRCh38) and the orientation is checked by flipping the effect allele (ie., the allele that confers the risk, which is not always the minor allele) and other alleles whenever appropriate. The beta and 95% confidence interval is also inverted accordingly [downloaded from: http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST005001-GCST006000/GCST005902/harmonised/29662059-GCST005902-EFO_0003761.h.tsv.gz].

The Genotype-Tissue Expression (GTEx) expression quantitative trait loci (eQTL) dataset contains single-tissue cis-eQTL data with eGene, meaning genes regulated by eQTL, and significant variant-gene associations for 49 tissues [downloaded from: https://www.gtexportal.org/home/datasets] (Genotype-Tissue Expression, 2017).

To match the GTEx eQTL panel, we imputed not genotyped GVs in the MD dataset with the Summary Statistics Imputation software (SSimp) (Rüeger et al., 2018). The parameters we used were GWAS full-genome summary statistics GVs with their matching z-scores, reference and effect alleles along with the European 1,000 genomes linkage disequilibrium (LD) reference panel [downloaded from: http://hgdownload.cse.ucsc.edu/gbdb/hg19/1000Genomes/phase3/]. We computed the z-scores by dividing GVs' effect size, understood as the effect of the risk allele relative to the reference allele, over the standard error (Shi, 2017). We then assessed the imputation quality returned by the SSimp software using the r2. pred parameter, which ranges between 0 -bad quality- and 1 -good quality-. Note that we only considered single nucleotide polymorphisms (SNPs) for this analysis.

2.2. Fine-mapping with PICS and TORUS

Before applying the Probabilistic Identification of Causal SNPs (PICS) and TORUS fine-mapping tools, we matched GWAS GVs and eQTLs to their corresponding LD blocks using the European 1,000 Genomes LD reference panel (Berisa and Pickrell, 2016).

We run PICS by programmatically accessing its web application form. We used LD-based PICS (https://pics2.ucsf.edu/pics2-LD.html), which performs LD expansion and fine-mapping. In brief, PICS takes the most significant GV per association locus along with its associated p-value, performs LD expansion and then computes the probabilities by performing empirical permutations per GV. For GWAS, we submitted the data and obtained the computed PICS probabilities for the input GVs and those in LD, from now on linked GVs. As for eQTLs, we downloaded precomputed LD-based PICS for all GTEx best eQTLs per gene per tissue type [downloaded from: https://pics2.ucsf.edu/Downloads/GTEx/].

We executed TORUS software package using the parameters "-load_zval -dump_pip". TORUS accepts full-genome summary statistics data, meaning all GVs analysed in the study, and their associated z-scores. Then it computes the causal probabilities using an expectation-maximization algorithm which assumes there is only one causal GV per locus. We obtained these probabilities for all GWAS GVs and GTEx eQTLs (v8) [downloaded from: https://storage.googleapis.com/gtex_analysis_v8/single_tissue_qtl_data/GTEx_Analysis_v8_eQTL.tar], per tissue.

The major histocompatibility complex region (chr6: 28,510,120–33,480,577, GRCh38) was excluded from all datasets for the analysis due to the complex LD structure of GVs, which may lead to inaccurate results (Ghoussaini et al., 2020).

2.3. Colocalization analysis via CLPP and fastENLOC

For the colocalization analysis, we implemented two different approaches in the workflow: the Colocalization Posterior Probability (CLPP) approach and the Fast Enrichment Estimation Aided Colocalization Analysis (fastENLOC). We applied these two methods to the fine-mapping results obtained with both PICS and TORUS to identify the genes regulated by causal GVs, also known as eGenes. Both tools consider that there can be more than one causal GV per association locus. CLPP assumes independence between GWAS and eQTL data while fastENLOC does not and computes the enrichment of GWAS on eQTL data using an embedded function. In addition, fastENLOC not only computes SNP colocalization probabilities (SCP) but also regional colocalization probabilities (RCP) to overcome the inability to narrow down to a single causal SNP, common to all tools. Please note that the tools were run following the guidelines and parameters recommended by the authors. We conducted a CLPP approach by computing the product of PICS probabilities for GWAS and eQTL overlapping linked GVs. Based on previous experience in post-GWAS data analysis, we narrowed down the results to the most likely causal GVs (Farh et al., 2015; Ghoussaini et al., 2020; Pérez-Granado et al., 2022) by filtering GWAS GVs and eQTLs PICS probabilities, as well as their product by >10%. We run fastENLOC with fine-mapped GWAS GVs and eQTLs per tissue using the following parameters: default shrinkage 1) and total variants (7,666,894). We filtered the results by RCP >0.5 and SCP >0.001 (Wen et al., 2017).

2.4. Proximal genes

Common gene mapping practices involve looking at the GVs' overlapping or nearest downstream and upstream genes, also known as proximal genes or pGenes. We retrieved this genetic information using Ensembl via SNPnexus (Oscanoa et al., 2020).

We first identified pGenes associated with GVs from PICS and TORUS fine-mapping results and performed gene-set enrichment analysis on both sets.

Then, for each fine-mapping and colocalization combination of tools, we obtained the pGenes to which the GVs mapped and compared them to the corresponding set of eGenes. We also evaluated each pGenes-eGenes set for their association with disease and performed a gene-set enrichment analysis.

2.5. Functional analysis

For the evaluation of association to disease, we followed two different approaches. When evaluating GVs from fine-mapping results, we used variant association data from DISGENET plus (Piñero et al., 2019; DISGENET plus, 2022). Note that the GWAS under evaluation (Howard et al., 2018) and a meta-analysis that it is a part of (Howard et al., 2019) were removed from DISGENET plus datasets to avoid circularity. As for genes, we used the R package disgenetplus2r (disgenetplus2r, 2022), which contains gene-disease association data, and considered Medical Subject Headings (MeSH) disease classes system for disease grouping.

We performed the gene-set enrichment analysis using g:Profiler via the R package gprofiler2 (Raudvere et al., 2019) and the following databases: 1) Gene Ontology (GO) biological processes, molecular functions and cellular processes; 2) Reactome and WikiPathways pathways; 3) miRNA annotations; 4) Human Phenotype Ontology, which focuses on rare Mendelian disorders, and has phenotypic features associated with disease; and 5) DISGENET plus, which has genes' association data to disease and phenotypic traits (v19). The whole set of known human genes was used as domain scope for the analysis and electronic GO annotations were not considered. Furthermore, to make the functional enrichment analysis more meaningful, we filtered the terms by their specificity using their term size (<1,500 genes), which corresponds to the number of genes associated with that term.

In addition, we applied a guilt-by-association approach to overcome the lack of functional information for some genes and assign the function of better-characterized neighbours in the interactome. Thus, we used molecular interaction data from IntAct (Orchard et al., 2014) clustered with MONET (Tomasoni et al., 2020) to evaluate whether different eGenes retrieved from the workflow could belong to the same cluster and thus affect the same molecular pathway. We performed a gene-set enrichment analysis of the retrieved clusters filtering by an eGene-cluster genes ratio of 1:50.

We evaluated the fine-mapping and colocalization results at different levels: the tissue specificity, colocalizing causal GVs, their target genes (eGenes) and their biological implications. We examined the results individually and then compared them across tools, with classic approaches (pGenes) and with the results reported in the original publication.

3. Results

This study evaluates and compares the outcomes of different fine-mapping and colocalization tools (Figure 1). To accomplish this, we have run our analysis using the same genome-wide association study (GWAS) on major depression (MD) and expression quantitative trait loci (eQTL) datasets. In addition, and in line with our goal, we address the results of each analytical step individually before getting into their biological implications. The workflow begins with an imputation phase (SSimp) to predict the genotypes not directly assayed in the original GWAS. Then, a fine-mapping step with Probabilistic Identification of Causal SNPs (PICS) and TORUS to identify the most likely causal genetic variants (GVs), meaning those likely to have a biological effect on the trait, and compute their causal probabilities. Next, a colocalization analysis using the Colocalization Posterior Probability (CLPP) approach and the fast enrichment estimation aided colocalization (fastENLOC) software, to identify the GVs causally associated with both MD and a change in expression of a target gene. Finally, the functional analysis, leveraging a diversity of databases, aims to decipher the impact of the identified GVs and eGenes, meaning genes regulated by eQTLs.

3.1. GWAS dataset imputation

The original genome-wide association study (GWAS) consisted of 7,624,931 harmonised genetic variants (GVs) and after imputation to predict missing Genotype-Tissue expression (GTEx) eQTLs, we obtained 7,947,219 GVs (ie. a total of 554,824 imputed GVs). The estimated imputation quality provided by SSimp was generally good for all chromosomes (r2. pred >0.8) except for chromosome 17.

3.2. Fine-mapping with PICS and TORUS

We run linkage disequilibrium (LD)-based PICS by inputting the most significant GWAS GVs per LD block (1,707 GVs) along with their p-values (Figure 2). After PICS LD expansion and fine-mapping, we

obtained 54,649 GVs with their corresponding PICS probabilities. As for the GTEx eQTLs, we downloaded the precomputed LD-based PICS per tissue from the data portal. In parallel, we computed the z-scores for all GWAS GVs and GTEx eQTLs and along with the LD block specification, we used them as input for TORUS.

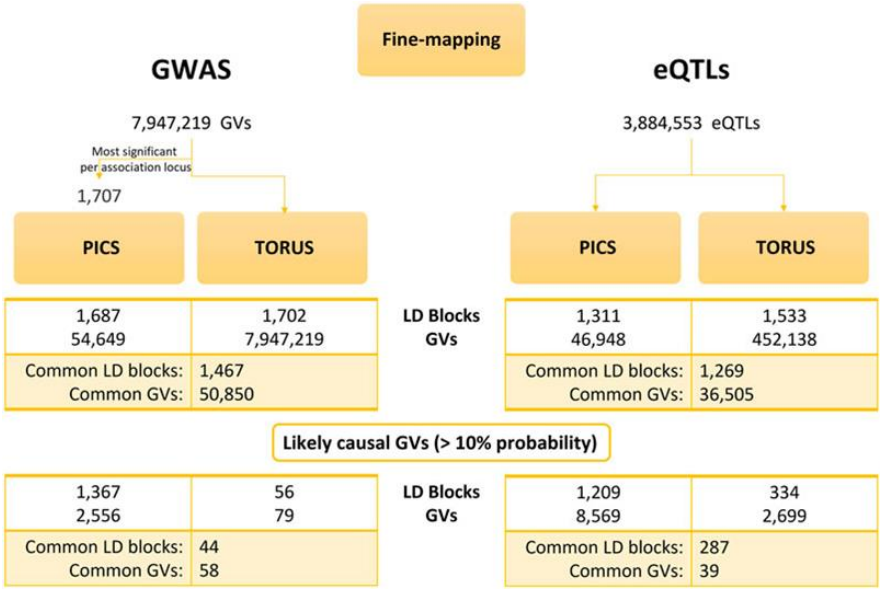


Figure 2. Results of PICS and TORUS fine-mapping analysis. Comparison of PICS and TORUS fine-mapping outcomes at GV and LD block level for both GWAS and eQTL datasets. PICS, Probabilistic Identification of Casual SNPs; GVs, genetic variants; LD, linkage disequilibrium; GWAS, genome-wide association studies; eQTLs, quantitative trait loci.

We compared PICS and TORUS initial fine-mapping results (Supplementary Figure S1) and then filtered GVs by a probability >10% to keep the most likely causal GVs. Because each tool has its own assumptions and different GVs could be identified, but these may be in LD, the comparison was done considering the probabilities per LD block. In addition, we examined the distribution of PICS and TORUS sum of probabilities for all LD blocks with likely causal GVs (GWAS: 1,367 and 56, respectively; GTEx: 1,209 and 334, respectively) (Supplementary Figure S1A) as well as the common ones (44 and 287, respectively) (Figure 3A). PICS probabilities for GWAS GVs are biased towards higher values in all cases, with 74% of GWAS LD blocks having a probability greater than 50%. Meanwhile, TORUS probability distribution is skewed towards lower values with only 21% of LD blocks

surpassing the 50% probability. Regarding GTEx eQTLs, PICS and TORUS results generally follow a more similar distribution with probabilities biased towards higher values, especially when only GVs with probabilities greater than 10% are considered (Figure 3B and Supplementary Figure S1B; note that we focused on Brain Frontal Cortex region because it is relevant to MD and for illustrative purposes).

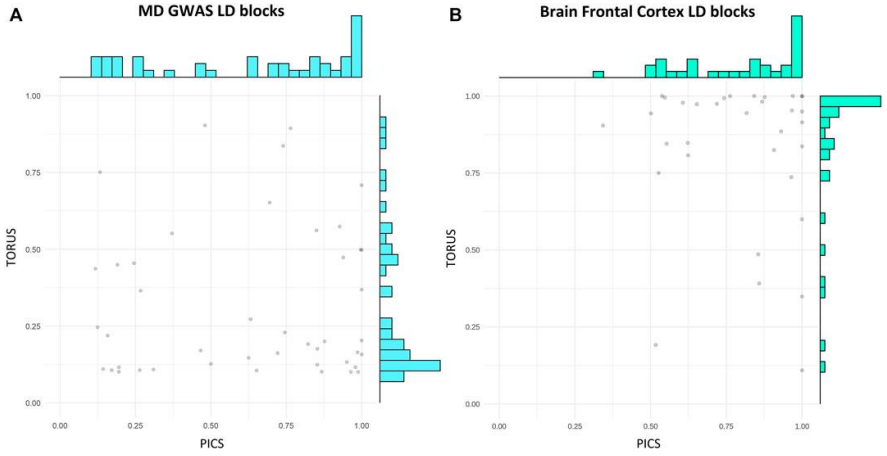


Figure 3. PICS and TORUS fine-mapping probabilities have different distributions. Scatter plot and distribution of PICS and TORUS probabilities for LD blocks containing GVs with PICS and TORUS probabilities >10%. (A) MD GWAS and (B) Brain Frontal Cortex LD block. PICS, Probabilistic Identification of Causal SNPs; LD, linkage disequilibrium; GVs, genetic variants; MD, major depression; GWAS, genome-wide association studies.

The analysis of PICS and TORUS most likely causal GVs (probability >10%) revealed that both sets are enriched in GVs associated with MD, bipolar disorder and other psychiatric disorders (Supplementary Tables S1, S2). PICS causal GVs are also enriched in metabolic-related traits such as triglycerides measurement.

Additionally, we applied classic gene-mapping approaches to PICS and TORUS fine-mapping results, yielding 1,277 and 1,248 proximal genes or pGenes, respectively. Both sets were enriched in genes associated with neurogenesis as well as neuron differentiation and development (Supplementary Tables S3, S4).

3.3. Colocalization analysis via CLPP and fastENLOC

The colocalization results from CLPP approach using PICS fine-mapping results yielded 44 GVs and 43 genes regulated by eQTLS, also

known as eGenes, affecting 28 tissues (Supplementary Table S5), whereas no results were obtained when using TORUS causal GVs. In parallel, fastENLOC applied to causal GVs identified by PICS resulted in 24 GVs and 17 eGenes across 13 tissues (Supplementary Table S6), while when applied on TORUS probabilities yielded 10 GVs and 3 eGenes in 2 tissues (Supplementary Table S7).

When comparing methods, the use of different colocalization tools after fine-mapping with PICS yields the most similar results. When using PICS, all tissues and eGenes identified by fastENLOC are also obtained by CLPP, with differences found at the GV level, and CLPP retrieving additional eGenes compared to fastENLOC (Supplementary Table S8 and Figure 4). Meanwhile, when comparing the use of PICS or TORUS fine-mapping probabilities followed by fastENLOC, we only identified one common tissue but with different eGenes and GVs. Similarly, PICS+CLPP and TORUS+fastENLOC yielded common findings only at the tissue level. Among the tissues with causal GVs and eGenes retrieved when using PICS and either colocalization tools, we can find diverse brain regions like the frontal cortex or hypothalamus.

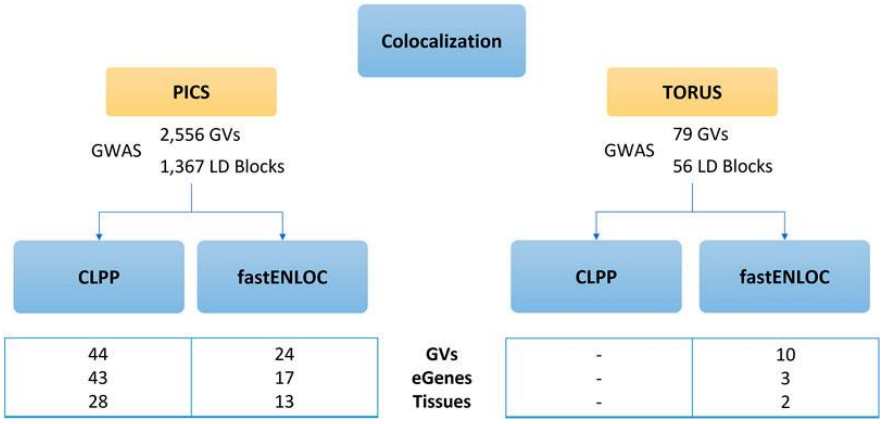


Figure 4. Results of CLPP and fastENLOC colocalization analysis. Comparison of CLPP and fastENLOC colocalization outcomes according to the prior fine-mapping tool used. CLPP, Colocalization Posterior Probability; fastENLOC, Fast Enrichment Estimation Aided Colocalization Analysis; GVs, genetic variants; eGenes, genes regulated by an expression quantitative trait loci.

3.4. Proximal genes and functional analysis

We compared eGenes from fine-mapping and colocalization workflow to pGenes from PICS and TORUS fine-mapping results. Only 3 genes

overlapped between pGenes from PICS fine-mapping and PICS+CLPP eGenes (KTN1, PXMP4 and ESYT2) and one with PICS+fastENLOC (KTN1). There was no overlap with pGenes when comparing to TORUS. The eGenes from PICS+CLPP are enriched in their association with miRNAs and the eGenes from PICS+fastENLOC in RNA Polymerase I Promoter Escape (Supplementary Table S9). Considering all eGenes together (46), these are functionally enriched in terms related to transcription factor regulation and miRNA. We also assessed the distribution of the eGenes in a clusterized human interactome. The three sets of eGenes (i.e., PICS+CLPP, PICS+fastENLOC and TORUS+fastENLOC) belonged to different clusters, except for eGenes shared across tools results (ie. 17 shared eGenes which are located in 10 clusters). Some of these clusters were associated with transcription factor regulation, inflammation or neurogenesis (Supplementary Tables S10, S11). No clusters identified for TORUS+fastENLOC passed the functional analysis filters, that is a ratio of eGenes over cluster genes higher than 1:50 and enriched term size <1,500 genes.

When we applied traditional gene mapping approaches to the GVs that were found to regulate the expression of those eGenes, we discovered a total of 74 pGenes. The vast majority of eGenes identified do not match pGenes, which holds true across all workflows (Table 1). In addition, most matches derive from GVs lying in an intronic region of the genome (Supplementary Table S12). Nonetheless, all sets of pGenes and eGenes are associated with mental disorders, behaviour and behaviour mechanisms as well as psychological phenomena and processes and nervous system disease (Supplementary Figures S2, S3). Additionally, pGenes are enriched in GO terms associated with diverse signalling pathways (Supplementary Table S13).

Furthermore, we compared the results obtained with the original publication where 14 GVs and 7 pGenes were reported. The latter are functionally associated with synapsis (Supplementary Table S13) and 6 of them have a prior association with mental disorders (Supplementary Figure S4). Only 5 fine-mapped GVs from PICS and 7 GVs from TORUS overlapped with the GVs reported in the original publication, and 1 pGene (SGIP1), which is in both sets of fine-mapped pGenes. However, none of the GVs and pGenes obtained by colocalization with any combination of tools evaluated in our pipeline overlapped with the GVs and pGenes reported in the original publication.

	eGenes	pGenes	Matches
PICS + CLPP	43	64	10
PICS + fastENLOC	17	31	3
TORUS + fastENLOC	3	8	0

Table 1. Identified eGenes differ from classic gene mapping (pGenes). The number of eGenes, also known as genes regulated by eQTLs, retrieved from the fine-mapping and colocalization analysis; the number of pGenes or proximal genes, that is overlapping or nearest upstream and downstream genes; and matches between eGenes and pGenes. The information is shown for each combination of fine-mapping and colocalization tools used. eGenes: genes regulated by expression quantitative trait loci; pGenes: proximal genes; PICS: Probabilistic Identification of Causal SNPs; CLPP: Colocalization Posterior Probability; fastENLOC: fast enrichment estimation aided colocalization analysis.

4. Discussion

Currently, there are a plethora of strategies available for post-GWAS analysis (Cano-Gamez and Trynka, 2020; Adebiyi et al., 2021). Here, we have focused on two main approaches: fine-mapping, which aims to identify the likely causal GVs, and colocalization, aimed at identifying which genes are regulated by the GVs at the expression level (eGenes). Furthermore, while many tools address the same goal, there is no standard set of causal GVs that have been experimentally validated for benchmarking to determine and compare which one is the most adequate (Brandes et al., 2022). Thus, we have designed an evaluation exercise to assess the outcome of different fine-mapping and colocalization tools using the same MD GWAS and eQTL dataset. To the best of our knowledge, no study goes beyond the comparison of the different tool's assumptions and thus the evaluation of the biological implications of their findings (Wen et al., 2017; Cano-Gamez and Trynka, 2020).

Our main premise throughout this analysis has been to use each tool as it was intended by following developers' recommendations and guidelines as closely as possible. This way we could get the most out of them and compare their optimised outcomes. Furthermore, one of the primary reasons behind the tools' selection was their ability to work with full-genome summary statistics instead of individual genotype data, which can be difficult to obtain due to privacy concerns. Other criteria

for tool selection included the quality of documentation, the maturity of the software, and the availability of developer support.

Prior to post-GWAS analysis, the imputation process using SSimp yielded very good quality results except for chromosome 17. One possible explanation is that SSimp provides hg19 1000Genomes phase3 as the reference panel. This version of the genome has some gaps, most of which are found in telomeres and centromeres, having a strong impact on chromosome 17 (Rashid-Kolvear et al., 2007; Genome Reference Consortium, 2022). We then proceeded with the fine-mapping and colocalization workflow, keeping the previously mentioned issue in mind when evaluating their results.

Fine-mapping is significantly influenced by LD patterns and the used tools, PICS and TORUS, which work under different assumptions (see Methods). Therefore, to have comparable results we considered the probabilities obtained at the LD block level, because the most likely causal GVs may differ or not be discernible due to high LD between GVs. In addition, to account for the difference in the number of GVs which could be driving the observed inverse distribution of probabilities between tools (Supplementary Figure S1), only the most likely causal GVs were considered in the comparison (Figure 2). In general, TORUS retrieves GVs with lower probabilities compared to PICS. This could be explained by the algorithm's conservative nature and its assumption of one causal GV per association locus, with probabilities biased towards zero when the locus contains multiple causal GVs (Wen, 2016). Indeed, the one causal GV assumption has been debated, with multiple GVs acting together resulting in a more reasonable theory (Burgess, 2022). Nonetheless, both PICS and TORUS most likely causal GVs are enriched in their association with MD, bipolar disorder and other psychiatric disorders (Supplementary Tables S1, S2). This suggests that both fine-mapping approaches identify likely causal GVs associated with MD. GVs fine-mapped by PICS are also enriched in diseases and traits usually comorbid with MD such as alcohol consumption (Gémes et al., 2019) and metabolic traits like serum total cholesterol measurement (Gold et al., 2020). Classic gene mapping of PICS and TORUS fine-mapping results (2,556 GVs and 79 GVs respectively, common- 58 GVs) (Figure 2), yielded 1,277 and 1,248 pGenes, respectively, with all TORUS pGenes included in PICS. These results could be explained because, compared to TORUS, PICS computes higher probability values and may retrieve more than one likely causal

GV per locus. But provided the set probability threshold, some of these GVs may be in LD and therefore mapping to the same genes. Both sets of pGenes are enriched in genes associated with neurogenesis, highly affected in MD (Li Z. et al., 2021). All in all, fine-mapping is a critical step in post-GWAS analysis, with high divergence observed between different methods, particularly at the level of GVs and their associated probabilities, which will highly impact subsequent colocalization analysis.

CLPP and fastENLOC colocalization approaches were applied to both fine-mapping results from PICS and TORUS. Following the same logic, given that TORUS computed lower probability values, PICS yielded more colocalization findings (Supplementary Tables S5–S7). Furthermore, we have similar results under CLPP assumption of independence between GWAS and eQTLs compared to fastENLOC built-in function to compute their enrichment, with fastENLOC being more stringent as previously described (Hukku et al., 2021). Interestingly, when focusing on a single tissue, the results do not match at the GV level but do so at the eGene level (Supplementary Table S8). This suggests that there might be different GVs that have an effect on the expression of the same eGenes. It also highlights the importance of the identification of eGenes to determine how GVs may ultimately impact the disease phenotype.

The overlap between eGenes and pGenes from PICS and TORUS fine-mapping was very small, with 3 genes in total. Among them, KTN1 has also been associated with MD (Dall'Aglio et al., 2021) and ESYT2 is involved in neurodevelopmental pathways and may be associated with suicidal behaviour trends in MDD although more research is needed (Calabrò et al., 2018). eGenes from PICS+CLPP were functionally enriched with miRNAs. These have been recently reported as relevant in MD pathogenesis and treatment (Dwivedi, 2014). Specifically, hsa-miR-23a-3p has repeatedly been associated with duloxetine treatment response assessment in MD (Kim et al., 2019). Moreover, the eGene GMPPB identified from TORUS+fastENLOC has already been associated with MD pathogenesis in proteome-wide association studies (Wingo et al., 2021). GMPPB is involved in glycosylation, which has been reported as relevant and even hypothesized as a potential biomarker for MD (Yamagata and Nakagawa, 2020). Considering all eGenes together, they are enriched in their association with transcription factor regulation (Supplementary Table S9), which has

already been related to MD (Zhong et al., 2019; Li X. et al., 2021; Pérez-Granado et al., 2022). The mapping of eGenes to protein interaction clusters indicated that the three sets of genes (i.e., PICS+CLPP, PICS+fasENLOC and TORUS+fastENLOC) belonged to distinct clusters and are thus likely to be involved in different biological processes. Nevertheless, PICS+CLPP and PICS+fastENLOC associated sets of clusters were enriched with genes associated with processes involving TF regulation as well as inflammation or neurogenesis (Supplementary Tables S10, S11). All these processes are associated with MD pathogenesis (Shadrina et al., 2018; Zhong et al., 2019; Li X. et al., 2021, Li et al., 2021 Z.; Pérez-Granado et al., 2022). In general, the identified eGenes are poorly characterized yet the cluster analysis shades some light on their potential molecular associations.

Fine-mapping and colocalization analysis successfully identified eGenes associated with mental disorders (Supplementary Figures S2–S4) that differed from the set of pGenes, particularly when focusing on non-coding regions of the genome (Table 1 and Supplementary Table S12). Accordingly, pGenes are enriched in their association with pathways that have been reported as disrupted in MD such as MAPK (Wang et al., 2020), ErbB (Ledonne and Mercuri, 2020), PI3K/AKT (Matsuda et al., 2019) and ERK (Wang and Mao, 2019) signalling pathways (Supplementary Table S13); as well as MD potential causes like stress or inflammation (Shadrina et al., 2018; Li Z. et al., 2021). When comparing the results from our workflow to the original manuscript, there were only matches when considering the fine-mapped PICS and TORUS results but not after colocalization analysis. The common pGene between the three datasets was SGIP1, which has been involved in mood regulation (Dvorakova et al., 2021).

Brain regions are of particular interest in MD and as such, we focused the evaluation of our results on them. The brain frontal cortex, hypothalamus, pituitary and brain cerebellar hemisphere have common findings between PICS and both colocalization tools. MD and myclonus-dystonia are usually comorbid, and their association has typically been studied in relation to SGCE mutation and its potential pleiotropic effect (Peall et al., 2013; Kim et al., 2017; Cazurro-Gutiérrez et al., 2021). However, whether SGCE plays a role in MD manifestation has been debated. On the one hand, animal studies have shown that knocking out this gene causes myoclonus, motor coordination deficits, and depression-like behaviour (Cazurro-Gutiérrez et al., 2021) which is

consistent with the lower expression levels reported by GTEx. On the other hand, a similar frequency of MD has been reported in SCGE mutated and wild-type myoclonus dystonia patients (Kim et al., 2017). Focusing on the hypothalamus, one of the most common causes of MD is stress, which affects the hypothalamic-pituitary-adrenal axis by increasing glucocorticoid levels (Karger et al., 2018; Oliva et al., 2018). These have an impact on various signalling pathways, including the Wnt pathway, in which FZD5 plays a role, and neurogenesis (Karger et al., 2018). However, the changes in gene expression caused by rs77678807 reported by GTEx are the inverse of what we would expect (Genotype-Tissue Expression, 2017). PCOLCE2 is highly expressed in the pituitary and there is evidence of reduced levels in depression-like behaviours in mice (Yamawaki et al., 2018), consistent with rs9757063 effect. Indeed, it has already been associated with psychiatric disorders by GWAS studies (Martínez-Magaña et al., 2021). However, how exactly they play a role in MD pathogenesis is still unknown. Little is known about the eGenes and GVs identified in the brain cerebellar hemisphere. Additionally, in the brain frontal cortex and hypothalamus, two different lncRNAs have been identified, LINC01159 and RP11-838N2.5 respectively. Even though little is known about them, lncRNAs seem to play a relevant role in MD pathogenesis and therapeutics (Shi et al., 2021; Hao et al., 2022). PICS+CLPP identified rs1480432 as upregulating the expression of DTNA, which is associated with neurogenesis and underregulating the maturation and stability of postsynaptic density (Chen et al., 2022). MAO B has been found to be overexpressed in postmortem brain tissue from MD patients, while DTNA is found to be underexpressed in MAO B knockout mice. The colon is another tissue whose associations with MD have produced intriguing results. ACTL8 is both associated with the microbiome composition and MD, but it is still unclear whether and/or which role the gut microbiome may have in a person's susceptibility to MD (Martins-Silva et al., 2021).

In general, both classic gene mapping approaches and colocalization analysis identified genes associated with MD or associated relevant processes. Colocalization analysis can provide insights about the effect of GVs located in non-coding regions of the genome, pinpointing the genes they regulate and the relevant tissues. As it has previously been reported the closest gene may not always be the causal one (Brodie et al., 2016; Zhu et al., 2016). These results would need further evaluation with other types of functional genomics data and ultimately

experimental validation to verify the role of these regulatory mechanisms in disease pathogenesis (Dehghan, 2018).

Our goal was to illustrate the impact of the lack of standards on the selection of the most adequate post-GWAS analysis method using a fine-mapping and colocalization workflow that compared different tools. The results revealed a high divergence between fine-mapping methods due to their assumptions, which in turn highly impacted the next steps. TORUS one causal variant assumption may tip the balance in favour of PICS considering fine-mapping and posterior analytical steps. Colocalization results seem to diverge in the amount of GVs and eGenes identified, with fastENLOC being more stringent by considering the enrichment of GWAS on eQTLs. All in all, despite the potential of combining GWAS data with molecular profiling datasets to guide in the interpretation of the functional impact of GVs located in non-coding regions of the genome, the results of our analysis revealed shortcomings related with the analytic tools. We propose that objective evaluation and benchmarking of post-GWAS analysis tools is required in order to fully leverage GWAS data for precision medicine and drug R&D applications.

Data availability statement

The original contributions presented in the study are included in the article Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

JP-G, JP, and LF designed the analysis. JP-G implemented the different analytical tools and wrote the manuscript with the support and guidance of JP and LF. All authors reviewed the manuscript. The authors read and approved the final manuscript.

Funding

IMI2-JU resources which are composed of financial contributions from the European Union's Horizon 2020 Research and Innovation Programme and EFPIA (GA: 116030 TransQST and GA: 777365 eTRANSAFE), and the EU H2020 Programme 2014–2020 (GA: 676559 Elixir-Excelerate); Project 001-P-001647—Valorisation of EGA for Industry and Society funded by the European Regional

Development Fund (ERDF) and Generalitat de Catalunya; Agència de Gestió d'Ajuts Universitaris i de Recerca Generalitat de Catalunya (2017SGR00519), and the Institute of Health Carlos III (project IMPaCT-Data, exp. IMP/00019), co-funded by the European Union, European Regional Development Fund (ERDF, "A way to make Europe"). The Research Programme on Biomedical Informatics (GRIB) is a member of the Spanish National Bioinformatics Institute (INB), funded by ISCIII and ERDF (PRB2-ISCIII (PT13/0001/0023, of the PE I + D + i 2013–2016)). The MELIS is a 'Unidad de Excelencia María de Maeztu', funded by the MINECO (MDM-2014-0370). JP-G was supported by Instituto de Salud Carlos III-Fondo Social Europeo (FI18/00034). This statement is a requirement from our funding agencies and therefore has to be included in the Funding section.

Conflict of interest

Competing interest reported. LF and JP are co-founders and hold shares of Medbioinformatics Solutions SL.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1006903/full#supplementary-material

Abbreviations

CLPP, Colocalization Posterior Probability; eGenes, genes regulated by eQTLs; eQTL, expression quantitative trait loci; fastENLOC, Fast Enrichment Estimation Aided Colocalization Analysis; GTEx, Genotype-Tissue Expression; GV, genetic variant; GWAS, Genome-Wide Association Studies; LD, linkage disequilibrium; MD, major depression; pGenes, proximal genes; PICS, Probabilistic Identification of Causal SNPs; SNPs, single nucleotide polymorphisms; SSimp, Summary Statistics Imputation software.

References

Adebiyi, E., Adam, Y., Samtal, C., Brandenburg, J. T., and Falola, O. (2021). Performing post-genome-wide association study analysis: Overview, challenges and recommendations. F1000Res. 10, 1002. doi:10.12688/f1000research.53962.1

Amlie-Wolf, A., Tang, M., Mlynarski, E. E., Kuksa, P. P., Valladares, O., Katanic, Z., et al. (2018). INFERNO: inferring the molecular mechanisms of noncoding genetic variants. Nucleic Acids Res. 46, 8740–8753. doi:10.1093/NAR/GKY686

Berisa, T., and Pickrell, J. K. (2016). Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics 32, 283–285. doi:10. 1093/BIOINFORMATICS/BTV546

Brandes, N., Weissbrod, O., and Linial, M. (2022). Open problems in human trait genetics. Genome Biol. 23, 131. doi:10.1186/s13059-022-02697-9

Brodie, A., Azaria, J. R., and Ofran, Y. (2016). How far from the SNP may the causative genes be? Nucleic Acids Res. 44, 6046–6054. doi:10.1093/NAR/GKW500

Buniello, A., Macarthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 47, D1005–D1012. doi:10.1093/NAR/GKY1120

Burgess, D. J. (2022). Fine-mapping causal variants — Why finding 'the one' can be futile. Nat. Rev. Genet. 23, 261. doi:10.1038/s41576-022-00484-7

Calabrò, M., Mandelli, L., Crisafulli, C., Lee, S. J., Jun, T. Y., Wang, S. M., et al. (2018). Neuroplasticity, neurotransmission and brain-related genes in major depression and bipolar disorder: Focus on treatment outcomes in an asiatic sample. Adv. Ther. 35, 1656–1670. doi:10.1007/s12325-018-0781-2

Cano-Gamez, E., and Trynka, G. (2020). From GWAS to function: Using functional genomics to identify the mechanisms underlying complex diseases. Front. Genet. 11, 424. doi:10.3389/fgene.2020.00424

Cazurro-Gutiérrez, A., Marcé-Grau, A., Correa-Vela, M., Salazar, A., Vanegas, M. I., Macaya, A., et al. (2021). ε-Sarcoglycan: Unraveling the

myoclonus-dystonia gene. Mol. Neurobiol. 58, 3938–3952. doi:10.1007/ s12035-021-02391-0

Chen, K., Palagashvili, T., Hsu, W., Chen, Y., Tabakoff, B., Hong, F., et al. (2022). Brain injury and inflammation genes common to a number of neurological diseases and the genes involved in the Genesis of GABAnergic neurons are altered in monoamine oxidase B knockout mice. Brain Res. 1774, 147724. doi:10.1016/J. BRAINRES.2021.147724

Dall'Aglio, L., Lewis, C. M., and Pain, O. (2021). Delineating the genetic component of gene expression in major depression. Biol. Psychiatry 89, 627–636. doi:10.1016/J.BIOPSYCH.2020.09.010

Dehghan, A. (2018). Genome-wide association studies. Methods Mol. Biol. 1793, 37–49. doi:10.1007/978-1-4939-7868-7_4

DISGENET plus (2022). DISGENET plus. Available at: https://beta.disgenetplus. com/ (Accessed December 21, 2021).

disgenetplus2r (2022). disgenetplus2r: An R package to explore the molecular underpinnings of human diseases. Available at: https://medbio.gitlab.io/ disgenetplus2r/ (Accessed December 21, 2022).

Dream Challenges (2022). DREAM Challenges use crowd-sourcing to solve complex biomedical research questions. Available at: https://dreamchallenges. org/ (Accessed May 26, 2022).

Dvorakova, M., Kubik-Zahorodna, A., Straiker, A., Sedlacek, R., Hajkova, A., Mackie, K., et al. (2021). SGIP1 is involved in regulation of emotionality, mood, and nociception and modulates in vivo signalling of cannabinoid CB1 receptors. Br. J. Pharmacol. 178, 1588–1604. doi:10.1111/BPH.15383

Dwivedi, Y. (2014). Emerging role of microRNAs in major depressive disorder: diagnosis and therapeutic implications. Dialogues Clin. Neurosci. 16, 43–61. doi:10. 31887/DCNS.2014.16.1/YDWIVEDI

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., et al. (2015). Genetic and epigenetic fine-mapping of causal autoimmune disease variants. Nature 518, 337–343. doi:10.1038/NATURE13835

Gazal, S., Weissbrod, O., Hormozdiari, F., Dey, K. K., Nasser, J., Jagadeesh, K. A., et al. (2022). Combining SNP-to-gene linking

strategies to identify disease genes and assess disease omnigenicity. Nat. Genet. 54, 827–836. doi:10.1038/s41588-022- 01087-y

Gémes, K., Forsell, Y., Janszky, I., László, K. D., Lundin, A., Ponce De Leon, A., et al. (2019). Moderate alcohol consumption and depression – A longitudinal population-based study in Sweden. Acta Psychiatr. Scand. 139, 526–535. doi:10. 1111/ACPS.13034

Genome Reference Consortium (2022). The genome reference Consortium. Available at: https://www.ncbi.nlm.nih.gov/grc (Accessed April 30, 2022).

Genotype-Tissue Expression (2017). GTEx portal. Available at: https://www. gtexportal.org/home/datasets (Accessed February 18, 2021).

Ghoussaini, M., Mountjoy, E., Carmona, M., Peat, G., Schmidt, E. M., Hercules, A., et al. (2020). Open targets genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. Nucleic Acids Res. 49, D1311–D1320. doi:10.1093/nar/gkaa840

Gold, S. M., Köhler-Forsberg, O., Moss-Morris, R., Mehnert, A., Miranda, J. J., Bullinger, M., et al. (2020). Comorbid depression in medical diseases. Nat. Rev. Dis. Prim. 6, 69–22. doi:10.1038/s41572-020-0200-2

Hao, W. Z., Chen, Q., Wang, L., Tao, G., Gan, H., Deng, L. J., et al. (2022). Emerging roles of long non-coding RNA in depression. Prog. Neuropsychopharmacol. Biol. Psychiatry 115, 110515. doi:10.1016/J.PNPBP.2022.110515

Hormozdiari, F., van de Bunt, M., Segrè, A. v., Li, X., Joo, J. W. J., Bilow, M., et al. (2016). Colocalization of GWAS and eQTL signals detects target genes. Am. J. Hum. Genet. 99, 1245–1260. doi:10.1016/J.AJHG.2016.10.003

Howard, D. M., Adams, M. J., Shirali, M., Clarke, T.-K., Marioni, R. E., Davies, G., et al. (2018). Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways. Nat. Commun. 9, 1470. doi:10.1038/s41467-018-03819-3

Howard, D. M., Adams, M. J., Clarke, T.-K., Hafferty, J. D., Gibson, J., Shirali, M., et al. (2019). Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the

prefrontal brain regions. Nat. Neurosci. 22, 343–352. doi:10.1038/s41593-018-0326-7

Hukku, A., Pividori, M., Luca, F., Pique-Regi, R., Im, H. K., and Wen, X. (2021). Probabilistic colocalization of genetic variants from complex and molecular traits: Promise and limitations. Am. J. Hum. Genet. 108, 25–35. doi:10.1016/j.ajhg.2020.11.012

Karger, S., Parhar, I. S., Teo, C. H., and Soga, T. (2018). Brain beta-catenin signalling during stress and depression. Neurosignals. 26, 31–42. doi:10.1159/ 000487764

Kichaev, G., Yang, W. Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., et al. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. 10, e1004722. doi:10.1371/JOURNAL.PGEN. 1004722

Kim, J. Y., Lee, W. W., Shin, C. W., Kim, H. J., Park, S. S., Chung, S. J., et al. (2017).

Psychiatric symptoms in myoclonus-dystonia syndrome are just concomitant features regardless of the SGCE gene mutation. Park. Relat. Disord. 42, 73–77. doi:10.1016/J.PARKRELDIS.2017.06.014

Kim, H. K., Tyryshkin, K., Elmi, N., Dharsee, M., Evans, K. R., Good, J., et al. (2019). Plasma microRNA expression levels and their targeted pathways in patients with major depressive disorder who are responsive to duloxetine treatment. J. Psychiatr. Res. 110, 38–44. doi:10.1016/J.JPSYCHIRES.2018. 12.007

LaPierre, N., Taraszka, K., Huang, H., He, R., Hormozdiari, F., and Eskin, E. (2021). Identifying causal variants by fine mapping across multiple studies. PLoS Genet. 17, e1009733. doi:10.1371/JOURNAL.PGEN.1009733

Ledonne, A., and Mercuri, N. B. (2020). On the modulatory roles of neuregulins/ ErbB signaling on synaptic plasticity. Int. J. Mol. Sci. 21, 275. doi:10.3390/ IJMS21010275

Li, X., Su, X., Liu, J., Li, H., Li, M., Li, W., et al. (2021a). Transcriptome-wide association study identifies new susceptibility genes and pathways for depression. Transl. Psychiatry 11, 306. doi:10.1038/S41398-021-01411-W

Li, Z., Ruan, M., Chen, J., and Fang, Y. (2021b). Major depressive disorder: Advances in neuroscience research and translational

applications. Neurosci. Bull. 37, 863–880. doi:10.1007/s12264-021-00638-3

Martínez-Magaña, J. J., Genis-Mendoza, A. D., Villatoro Velázquez, J. A., Bustos- Gamiño, M., Juárez-Rojop, I. E., Tovilla-Zarate, C. A., et al. (2021). Genome-wide association study of psychiatric and substance use comorbidity in Mexican individuals. Sci. Rep. 11, 6771. doi:10.1038/s41598-021-85881-4

Martins-Silva, T., Salatino-Oliveira, A., Genro, J. P., Meyer, F. D. T., Li, Y., Rohde,

L. A., et al. (2021). Host genetics influences the relationship between the gut microbiome and psychiatric disorders. Prog. Neuropsychopharmacol. Biol. Psychiatry 106, 110153. doi:10.1016/J.PNPBP.2020.110153

Matsuda, S., Ikeda, Y., Murakami, M., Nakagawa, Y., Tsuji, A., and Kitagishi, Y. (2019). Roles of PI3K/AKT/GSK3 pathway involved in psychiatric illnesses. Diseases 7, 22. doi:10.3390/DISEASES7010022

Oliva, C. A., Montecinos-Oliva, C., and Inestrosa, N. C. (2018). Wnt signaling in the central nervous system: New insights in Health and disease. Prog. Mol. Biol. Transl. Sci. 153, 81–130. doi:10.1016/BS.PMBTS.2017.11.018

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., et al. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 42, D358–D363. doi:10.1093/ NAR/GKT1115

Oscanoa, J., Sivapalan, L., Gadaleta, E., Dayem Ullah, A. Z., Lemoine, N. R., and Chelala, C. (2020). SNPnexus: A web server for functional annotation of human genome sequence variation (2020 update). Nucleic Acids Res. 48, W185–W192. doi:10.1093/NAR/GKAA420

Peall, K. J., Smith, D. J., Kurian, M. A., Wardle, M., Waite, A. J., Hedderly, T., et al. (2013). SGCE mutations cause psychiatric disorders: clinical and genetic characterization. Brain 136, 294–303. doi:10.1093/BRAIN/AWS308

Pérez-Granado, J., Piñero, J., Medina-Rivera, A., and Furlong, L. I. (2022). Functional genomics analysis to disentangle the role of genetic variants in major depression. Genes (Basel) 13, 1259. doi:10.3390/GENES13071259

Piñero, J., Piñero, P., Manuel Ramírez-Anguita, J., Sä Uch-Pitarch, J., Ronzano, F., Centeno, E., et al. (2019). The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 48, D845–D855. doi:10.1093/nar/gkz1021

Pividori, M., Rajagopal, P. S., Barbeira, A., Liang, Y., Melia, O., Bastarache, L., et al. (2020). PhenomeXcan: Mapping the genome to the phenome through the transcriptome. Sci. Adv. 6, eaba2083. doi:10.1126/SCIADV.ABA2083

Protein Structure Prediction Center (2020). Available at: https://predictioncenter. org/ (Accessed May 26, 2022).

Rashid-Kolvear, F., Pintilie, M., and Done, S. J. (2007). Telomere length on chromosome 17q shortens more than global telomere length in the development of breast cancer. Neoplasia 9, 265–270. doi:10.1593/NEO.07106

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et al. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 47, W191–W198. doi:10.1093/nar/ gkz369

Rüeger, S., McDaid, A., and Kutalik, Z. (2018). Evaluation and application of summary statistic imputation to discover new height-associated loci. PLoS Genet. 14, e1007371. doi:10.1371/JOURNAL.PGEN.1007371

Shadrina, M., Bondarenko, E. A., and Slominsky, P. A. (2018). Genetics factors in major depression disease. Front. Psychiatry 9, 334. doi:10.3389/fpsyt.2018.00334

Shi, Y., Wang, Q., Song, R., Kong, Y., and Zhang, Z. (2021). Non-coding RNAs in depression: Promising diagnostic and therapeutic biomarkers. EBioMedicine 71, 103569. doi:10.1016/J.EBIOM.2021.103569

Shi, H. (2017). Tips for formatting A lot of GWAS summary association statistics data. Available at: https://huwenboshi.github.io/data%20management/2017/11/23/ tips-for-formatting-gwas-summary-stats.html (Accessed March 30, 2022).

Taylor, K. E., Ansel, K. M., Marson, A., Criswell, L. A., and Farh, K. K.-H. (2021). PICS2: next-generation fine mapping via probabilistic

identification of causal SNPs. Bioinformatics 37, 3004–3007. doi:10.1093/BIOINFORMATICS/BTAB122

Tomasoni, M., Gómez, S., Crawford, J., Zhang, W., Choobdar, S., Marbach, D., et al. (2020). MONET: a toolbox integrating top-performing methods for network modularization. Bioinformatics 36, 3920–3921. doi:10.1093/BIOINFORMATICS/ BTAA236

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 Years of GWAS discovery: Biology, function, and translation. Am. J. Hum. Genet. 101, 5–22. doi:10.1016/j.ajhg.2017.06.005

Wallace, C. (2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. PLoS Genet. 17, e1009440. doi:10.1371/JOURNAL. PGEN.1009440

Wang, J. Q., and Mao, L. (2019). The ERK pathway: Molecular mechanisms and treatment of depression. Mol. Neurobiol. 56, 6197–6205. doi:10.1007/S12035-019- 1524-3

Wang, X. L., Yuan, K., Zhang, W., Li, S. X., Gao, G. F., and Lu, L. (2020).

Regulation of circadian genes by the MAPK pathway: Implications for rapid antidepressant action. Neurosci. Bull. 36, 66–76. doi:10.1007/s12264-019-00358-9

Wen, X., Pique-Regi, R., and Luca, F. (2017). Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLoS Genet. 13, e1006646. doi:10.1371/JOURNAL.PGEN. 1006646

Wen, X. (2016). Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control. Ann. Appl. Stat. 10, 1619–1638. doi:10. 1214/16-AOAS952

Wingo, T. S., Liu, Y., Gerasimov, E. S., Gockley, J., Logsdon, B. A., Duong, D. M., et al. (2021). Brain proteome-wide association study implicates novel proteins in depression pathogenesis. Nat. Neurosci. 24, 810–817. doi:10.1038/s41593-021- 00832-6

Yamagata, H., and Nakagawa, S. (2020). Glycosylation and depression — a review. Trends Glycosci. Glycotechnol. 32, 157–160. doi:10.4052/tigg.2002.1E

Yamawaki, Y., Yoshioka, N., Nozaki, K., Ito, H., Oda, K., Harada, K., et al. (2018). Sodium butyrate abolishes lipopolysaccharide-induced

depression-like behaviors and hippocampal microglial activation in mice. Brain Res. 1680, 13–38. doi:10.1016/ J.BRAINRES.2017.12.004

Zhong, J., Li, S., Zeng, W., Li, X., Gu, C., Liu, J., et al. (2019). Integration of GWAS and brain eQTL identifies FLOT1 as a risk gene for major depressive disorder. Neuropsychopharmacology 44, 1542–1551. doi:10.1038/s41386-019-0345-4

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. 48, 481–487. doi:10.1038/ng.3

4. DISCUSSION

4.1 Available genetic association data on MD

Current status of curation guidelines

Advances in genetic association studies have encouraged the development of curation guidelines and the use of standard terminologies to ensure a systematic and accurate data annotation process, as well as resource interoperability[87,89,93]. However, curation criteria for complex diseases in general, and MD in particular, do not consider multiple evidences or further assess study type and design. Instead, their association evaluation remains at the individual publication level. Therefore, the first objective of this thesis was to develop curation guidelines for evaluating the validity of MD genetic association data. The second one was to create a database of GVs associated with MD following the developed guidelines.

Compared to Mendelian and rare disorders, which have gene-disease and GV-disease guidelines for association validity, complex diseases pose additional challenges to the creation of such guidelines. Multiple GVs contribute to disease risk, most of which are found in non-coding regions of the genome and have no clear link to a target gene or regulatory mechanism[95,96]. Thus, guidelines designed for Mendelian or rare disorders cannot be directly translated into complex diseases, and such features should be considered.

The current understanding of MD is based on a variety of genetic association studies, highlighting CGS and GWAS, which have identified GVs and genes significantly associated with MD[8,76,139]. These data are scattered across multiple resources and in the literature. On top of that, the complexity of MD has led to its study from various angles, such as the environment or treatment response aspect. Overall, there is a critical need for the follow-up and validation of such findings to better understand their role in disease pathogenesis.

Proposal of curation guidelines

A curated knowledge base is required to promote an accurate posterior analysis. In this thesis, we have developed curation guidelines for assessing MD-associated GVs combining different resources and evidences from diverse association and study types; we have evaluated

their contexts and conducted a quality control. Ultimately, we have created a knowledge resource that takes into account all MD aspects mentioned above (**Chapter 3.1**).

In the absence of curation guidelines that evaluate genetic association data for MD, we developed ours based on current literature, other diseases' curation guidelines, and an accurate inspection of collected association data. The four main steps of the curation guidelines address a critical aspect of MD genetic association data, given its complexity and heterogeneity. In this thesis, we have focused on the particular case of MD. However, we think an adapted version of the developed curation guidelines could be used for other complex diseases with a comparable genetic architecture.

The outcomes and limitations of the curation process

The conducted curation process has been a semi-automatic approach which enabled the creation of an expert-curated database. Aided by bioinformatic tools for data collection and curation, we have extracted and evaluated more than 2000 publications that resulted in 709 GVs and 65 publications. For that, we believe that a comprehensive manual curation process is required for an appropriate evaluation of the evidences and to identify false associations.

Data from different types of genetic association studies (i.e., VDA, TR and E) and from diverse nature (i.e., studies in human and animal models) were collected by inspecting different resources. The small overlap between types of association studies can be due to different GVs affecting different aspects of the disease. Regarding study type, whereas GWAS repositories only contain GWAS data, TM can capture associations from CGS, GWAS and preclinical models. Given the small overlap between resources and study types, data integration from various resources resulted in a more representative dataset. The small overlap between CGS and GWAS, combined with the fact that many CGS were discarded during the quality control step, highlights CGS's lack of replicative power attributed to the small sample size typically used in such designs. Indeed, we could observe that no CGS conducted before 2008 passed our curation criteria. In addition, CGS are hypothesis-driven and generally evaluate GVs in coding regions of the genome. In contrast, GWAS are hypothesis-free and identify GVs mostly in non-coding regions, with potential regulatory roles in gene expression. Nonetheless, it should be considered that the GVs

identified by GWAS may not be the ones with a role in disease pathogenesis, but these may be in LD with the true causal ones. All in all, the principles underlying CGS and GWAS, as well as their limitations, led to a small overlap between the genes these GVs mapped to.

The criticism against CGS is based on the reported large effects from GVs identified using small sample sizes and their lack of replicative power[43,44]. In addition, CGS focus on a single or small set of GVs and genes, while MD is a complex disease in which many GVs, each with a minor effect, are expected to play a role. Nonetheless, we found CGS evidences passing the curation pipeline quality control. And, although it has been argued that CGS should be abandoned[58], if their study design is properly performed, these could still aid in understanding MD pathogenesis. Specifically, it has been proposed that best-informed decisions could be made by considering all available information, especially in the case of psychiatric disorders[140].

Cellular and animal models of MD are very diverse and mimic different aspects of the disease phenotype with no one-fits-all model[78,80]. As a result, their quality control may include the characterisation and evaluation of factors other than sample size and significance. For instance, how the model was generated, how it resembles MD phenotype and which tests are employed. Furthermore, the compromise reached in the sample size cut-off for GWAS and CGS could be further fine-tuned depending on the scope and number of GVs expected to be identified, as well as their effects[57,70,141]. Future work in defining the quality control criteria for cellular and animal models will enable the objective assessment of evidences provided by these type of studies.

The GVs in the curated dataset were mostly lying in non-coding regions of the genome with no clear effect on MD pathogenesis[95]. These findings require a posterior functional analysis to elucidate the causal GVs and understand how these influence disease risk; this is emphasised by the single GV that overlapped across study types. These results along with the GVs' minor effects have led to some criticism towards GWAS clinical utility[142]. However, in other complex diseases, GVs identified by GWAS are being applied for disease risk assessment, disease classification, therapeutic development, and drug selection, offering some promise[142]. To promote an accurate diagnosis and

accelerate treatment options, larger sample sizes and more specific phenotyping, as well as advances in posterior functional analyses of these GVs, will be critical.

4.2. Limitations, challenges and future steps in the post-GWAS era

Post-GWAS studies: available data and analyses

Multiple genetic association studies on MD have been carried out; however, despite this wealth of information, there is still much to learn about MD pathophysiological mechanisms. Therefore, the third objective of the thesis was to uncover potential regulatory mechanisms by which GVs associated with MD may contribute to disease pathogenesis. Having considered that objective, the fourth one came naturally: to benchmark post-GWAS analysis tools by systematically evaluating their performance and selecting the most suitable ones for the interpretation of GWAS findings on MD.

Thanks to the advances in the different omics fields (i.e., genomics, transcriptomics, proteomics, metabolomics and epigenomics), we can reach molecular-level characterisation of disease mechanisms. As a result, integrating omics data with GWAS findings has become key to advance in our understanding of complex diseases pathogenesis and, thus, MD[143].

Despite a large amount of genetic association data available, privacy constraints relating to the exploitation of individual data limit the use of full genotype data. Recent efforts to increase sample sizes have encouraged private and public sector collaborations, which has led to stringent data-sharing policies regarding full genome summary statistics (i.e., association results of all GVs tested). As a result, several post-GWAS analysis tools have adapted to using summary statistics (i.e., the most significant GVs) as an alternative when full genome summary statistics are unavailable[144,145].

Numerous post-GWAS analysis tools are available for analysing GWAS findings for various purposes. Although different tools with the same goal exist (e.g., for fine-mapping or for colocalization), their data requirements, assumptions, and results may differ[104,146,147]. There is currently no protocol or standard procedure to address which tool

would be the best in each case, and maybe their combination would be the best approach as it has been suggested in other contexts[129]. Some method comparisons have been made, but these have remained at the mathematical level rather than delving into the biological implications of their findings[104,147]. Additionally, there is no benchmark dataset to assist in the evaluation of the performance of post-GWAS tools. A set of manually curated genes with moderate to high confidence evidence of their functional role has been proposed to aid in prioritising causal genes at GWAS loci[148]. But, to reduce potential biases on this type of datasets, high-quality gold-standard GWAS datasets that represent a wide range of molecular mechanisms and genetic architectures are required.

In this context, we have conducted diverse functional genomics analyses, including fine-mapping, colocalization and transcription factor binding site analysis to determine the role of GVs in MD (**Chapters 3.2 and 3.3**). We have also developed and implemented a workflow that compares the outcomes of different post-GWAS analysis tools, focused on fine-mapping and colocalization, and their biological implications (**Chapter 3.3**).

Finding the causal ones

Fine-mapping methods help unravel which are the most likely causal GVs[144]. These do not necessarily correspond to the most significant GWAS findings but may be in LD, with LD patterns significantly influencing this type of analysis. First, since these patterns are population specific, the LD reference panel and the GWAS dataset population should match. Another layer of complexity is added when GVs are in high LD to distinguish between individual GVs' effects[149].

The common assumption that only a single causal GV exists per loci has been questioned, and it has been argued that the phenotype may be driven by the interaction of multiple GVs within the same locus[149]. Differences resulting from either considering it or not were demonstrated by comparing the fine-mapping tools TORUS[150] and Probabilistic Identification of Causal SNPs (PICS)[144]. Given TORUS one causal variant assumption, its application yielded probabilities biased towards zero when multiple causal GVs were present in that locus. In contrast, PICS retrieved more potentially causal GVs with higher estimated causal probabilities. As a result, fine-mapping proved to be a critical step, especially since it is often required before

performing other post-GWAS analyses. Nonetheless, both sets of GVs were enriched in their association with MD or its comorbidities, and the genes they mapped to were associated with neurogenesis, suggesting that the results of both tools were valid.

Translating into functional elements

In contrast to fine-mapping methods' comparison, colocalization using Colocalization Posterior Probability (CLPP)[151] and Fast Enrichment Estimation Aided Colocalization Analysis (fastENLOC)[152,153] yielded more similar results; especially when PICS outcomes, which had higher probability values, were considered. FastENLOC may be more stringent because it computes the enrichment of GWAS on eQTLs rather than assuming their independence[153]. Although colocalizing GVs (i.e., GV causal for both the disease and the molecular trait) may not match, when looking at the genes these GVs regulate, the overlap is higher, highlighting the relevance of identifying the genes impacting disease phenotype.

The extension of the performed comparisons would be of interest. We did apply the developed benchmarking workflow to a GWAS dataset on height[154], which also has a complex genetic architecture, and obtained similar results. Nevertheless, evaluating additional post-GWAS analyses with different assumptions and goals would be interesting in order to further explore their strengths and weaknesses. It would be interesting to undergo colocalization and TWAS analyses to compare the resulting sets of genes being functionally regulated by the causal GVs. Indeed, it has been proposed that their combination could enhance the identification of biologically relevant genes[117]. Likewise, colocalization and Mendelian randomisation have also been proposed as complementary approaches. While in colocalization the GVs may be associated with both the disease and the molecular trait, and causal effects may occur or not; in Mendelian randomisation, the GVs would be directly associated with risk and outcome only via the risk[155].

Overall, the lack of objective standards for method selection, evaluation and benchmarking challenges the potential use of GWAS data for downstream application in drug research and precision medicine. The plethora of tools available for the same purpose diverge not only in findings (i.e., GVs or genes) but also in their biological impacts, ultimately influencing subsequent functional analysis steps. Despite

that, they all contribute to a better understanding of the functional impact of GVs in non-coding regions of the genome.

Leveraging different data types increases the capability to filter and prioritise relevant GVs and genes. Regarding transcription factor binding analysis, we have combined data from: ChIP-seq, gene expression, TFs' position specific-score matrices, epigenome annotation, eQTLs and TF-target pairs; to identify evidence for regulatory and mechanistic effects. Further experimental validation would be of interest to determine the role of the identified TF whose binding site is being altered, i.e., activator or repressor, and consequently support the GV's impact on gene expression.

The diverse functional analysis, reinforced with genomic annotation data, enabled the identification of GVs potentially causal of MD, their proximal genes, as well as genes these GVs could be regulating or whose TFBS are altering. Besides their identification and the proposal of a mechanistic hypothesis for the latter, their characterisation via gene-set and variant-set enrichment analyses revealed their relevance in MD. The identified functional elements were enriched in their association with brain networks, neurogenesis, synapsis, and neurons. Additionally, pathways disrupted in MD or associated with its triggers, such as stress or inflammation, were also identified. As a result, further functional evaluation and, eventually, experimental validation of such findings would be required to determine the exact regulatory mechanism by which these GVs ultimately impact MD pathogenesis[84,156,157].

5. CONCLUSIONS

- We developed curation guidelines for assessing genetic association data for the specific case of MD, considering its genetic architecture and the diversity of methodologies used to study it.

- Implementing such guidelines proved critical for developing an expert-curated database of 709 GVs associated with MD.

- The curation guidelines could be applied to other complex disorders with similar genetic architecture.

- We implemented bioinformatic pipelines to evaluate GWAS findings, leveraging a variety of post-GWAS techniques and omics data, enabling the prioritisation of GVs and the formulation of mechanistic hypotheses by which these impact disease pathogenesis.

- Our pipelines are able to exploit either summary statistics or full genome summary statistics, adapting to their availability.

- The use of fine-mapping techniques to identify potential causal GVs and posterior colocalization analysis to identify the genes regulated by these GVs revealed the importance of chloride homeostasis and myelination in the pathobiology of MD.

- The integration of GWAS with diverse genomic annotation data (including ChIP-seq, gene expression, TFs' position specific-score matrices, chromatin accessibility data, eQTLs and TF-target pairs) revealed a potential role of hypoxia response in MD mediated by altered TF binding.

- The benchmarking of post-GWAS analysis tools highlighted important differences in the results obtained with different tools at both genetic and biological levels, which may impact downstream analytical steps and, ultimately, biological interpretation. Nonetheless, it revealed that different tools could lead to biologically plausible findings.

- The benchmarking workflow has proven the need for community-accepted guidelines for the selection and objective evaluation of the most suitable post-GWAS analysis methods among all the available ones to fully leverage GWAS data for drug research and precision medicine.

6. LIST OF PUBLICATIONS

6.1. Articles

1. Pérez-Granado, J., Piñero, J., & Furlong, L.I. Benchmarking post-GWAS analysis tools in major depression: Challenges and implications. *Frontiers in Genetics*, **13**, 2853 (2022). https://doi.org/10.3389/fgene.2022.1006903
2. Pérez-Granado, J., Piñero, J., Medina-Rivera, A., & Furlong, L. I. Functional Genomics Analysis to Disentangle the Role of Genetic Variants in Major Depression. *Genes (Basel)*, **13**, 1259 (2022). https://doi.org/10.3390/GENES13071259
3. Trincado, J. L., Reixachs-Solé, M., Pérez-Granado, J., Fugmann, T., Sanz, F., Yokota, J., & Eyras, E. ISOTOPE: ISOform-guided prediction of epiTOPEs in cancer. *PLOS Computational Biology*, **17**, e1009411 (2021). https://doi.org/10.1371/journal.pcbi.1009411
4. Pérez-Granado, J., Piñero, J., & Furlong, L. I. ResMarkerDB: a database of biomarkers of response to antibody therapy in breast and colorectal cancer. *Database*, baz060 (2019). https://doi.org/10.1093/database/baz060

6.2. Oral communications

1. BioScience-App Day | ResMarkerDB. BioScience-App Courses. 2021-07-16. https://www.youtube.com/watch?v=pO3Y6Smb-Rc
2. Aprende ggplot2. RLadies Querétaro. 2021-03-20. https://www.youtube.com/watch?v=3Vm-WNIv-qg
3. The genetics of depression. Barcelona Science Slam 2020; 2020-02-22. https://www.youtube.com/watch?v=1IkfsDtjy0c
4. Exploring the genetic architecture of Major Depression: Low agreement between Text Mining and GWAS results. 4th European Conference on Translational Bioinformatics; 2019-12-12.
5. ResMarkerDB: un base de datos de biomarcadores de respuesta a anticuerpos monoclonales en cáncer mama y cáncer

colorrectal. XXII Congreso Nacional de Informática de la Salud; 2019-03-07.

6. Terapia en càncer basada en biomarcadors. Biennal Ciutat i Ciència; 2019-02-12. https://www.youtube.com/watch?v=9UPRSbpi0d4
7. ResMarkerDB: a database of biomarkers of response to antibody therapy in breast and colorectal cancer. 7th Jornada d'Investigadors Predoctorals Interdisciplinària; 2019-02-04.

6.3. Posters

1. Pérez Granado J, Piñero J and Furlong LI. Exploring the genetic architecture of Major Depression: Low agreement between Text Mining and GWAS results. F1000Research 2019, 8(ISCB Comm J):1984 (poster) (https://doi.org/10.7490/f1000research.1117696.1)

   Presented at:

   - Advances in Computational Biology Conference 2019-Fostering collaboration among women scientists; 2019-12-29

2. Pérez Granado J, Piñero J and Furlong LI. ResMarkerDB: a database of biomarkers of response to antibody therapy in breast and colorectal cancer. F1000Research 2019, 8:1543 (poster) (https://doi.org/10.7490/f1000research.1117434.1)

   Presented at:

   - Advances in Computational Biology Conference 2019-Fostering collaboration among women scientists; 2019-12-29 - Best poster award
   - 4th European Conference on Translational Bioinformatics; 2019-12-12
   - Intelligent Systems for Molecular Biology (ISMB)-European Conference on Computational Biology (18th Annual Conference) ISMB/ECCB 2019; 2019-07-25
   - VI Jornada de Bioinformàtica i Genòmica; 2018-12-20

7. BIBLIOGRAPHY

1. Wittchen, H. U. *et al.* The size and burden of mental disorders and other disorders of the brain in Europe 2010. *European Neuropsychopharmacology* **21**, 655–679 (2011).
2. Kessler, R. C., Chiu, W. T., Demler, O. & Walters, E. E. Prevalence, Severity, and Comorbidity of 12-Month DSM-IV Disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry* **62**, 617 (2005).
3. Qin, X. *et al.* Prevalence and rates of recognition of depressive disorders in internal medicine outpatient departments of 23 general hospitals in Shenyang, China. *J Affect Disord* **110**, 46–54 (2008).
4. World Health Organization. *The global burden of disease: 2004 update.* https://apps.who.int/iris/handle/10665/43942 (2008).
5. Sullivan, P. F., Neale, M. C. & Kendler, K. S. Genetic epidemiology of major depression: Review and meta-analysis. *American Journal of Psychiatry* **157**, 1552–1562 (2000).
6. Li, Z., Ruan, M., Chen, J. & Fang, Y. Major Depressive Disorder: Advances in Neuroscience Research and Translational Applications. *Neurosci Bull* **37**, 863–880 (2021).
7. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet* **50**, 668–681 (2018).
8. Howard, D. M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat Neurosci* **22**, 343–352 (2019).
9. Kendall, K. M. *et al.* The role of rare copy number variants in depression. *bioRxiv* 378307 (2018) doi:10.1101/378307.
10. Zhang, X. *et al.* Genome-wide Burden of Rare Short Deletions Is Enriched in Major Depressive Disorder in Four Cohorts. *Biol Psychiatry* **85**, 1065–1073 (2019).
11. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. (American Psychiatric Association, 2013). doi:10.1176/appi.books.9780890425596.

12. World Health Organization. *The ICD-10 classification of mental and behavioural disorders: Diagnostic criteria for research*. (World Health Organization, 1993).
13. Unal-Aydin, P., Aydin, O. & Arslan, A. Genetic Architecture of Depression: Where Do We Stand Now? *Adv Exp Med Biol* **1305**, 203–230 (2021).
14. Mei, L. *et al.* Overlapping common genetic architecture between major depressive disorders and anxiety and stress-related disorders. *Prog Neuropsychopharmacol Biol Psychiatry* **113**, 110450 (2022).
15. Gold, S. M. *et al.* Comorbid depression in medical diseases. *Nat Rev Dis Primers* **6**, 1–22 (2020).
16. Kang, H.-J. *et al.* Sex differences in the genetic architecture of depression. *Sci Rep* **10**, 9927 (2020).
17. Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat Genet* **48**, 624–633 (2016).
18. Berton, O. & Nestler, E. J. New approaches to antidepressant drug discovery: beyond monoamines. *Nat Rev Neurosci* **7**, 137–151 (2006).
19. Yuan, Z., Chen, Z., Xue, M., Zhang, J. & Leng, L. Application of antidepressants in depression: A systematic review and meta-analysis. *Journal of Clinical Neuroscience* **80**, 169–181 (2020).
20. Aydin, O., Unal Aydin, P. & Arslan, A. Development of Neuroimaging-Based Biomarkers in Psychiatry. *Adv Exp Med Biol* **1192**, 159–195 (2019).
21. Hillhouse, T. M. & Porter, J. H. A brief history of the development of antidepressant drugs: From monoamines to glutamate. *Exp Clin Psychopharmacol* **23**, 1 (2015).
22. Arslan, A. *Application of neuroimaging in the diagnosis and treatment of depression*. *Understanding Depression: Clinical Manifestations, Diagnosis and Treatment* vol. 2 (Springer Singapore, 2018).
23. Krishnan, V. & Nestler, E. J. The molecular neurobiology of depression. *Nature* **455**, 894–902 (2008).
24. Kraft, J. B., Slager, S. L., McGrath, P. J. & Hamilton, S. P. Sequence Analysis of the Serotonin Transporter and Associations with Antidepressant Response. *Biol Psychiatry* **58**, 374–381 (2005).

25. Schildkraut, J. J. The catecholamine hypothesis of affective disorders: a review of supporting evidence. *Am J Psychiatry* **122**, 509–522 (1965).
26. Rush, A. J. *et al.* Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report. *American Journal of Psychiatry* **163**, 1905–1917 (2006).
27. Bao, A. M., Meynen, G. & Swaab, D. F. The stress system in depression and neurodegeneration: Focus on the human hypothalamus. *Brain Res Rev* **57**, 531–553 (2008).
28. Holsboer, F. & Barden, N. Antidepressants and Hypothalamic-Pituitary-Adrenocortical Regulation. *Endocr Rev* **17**, 187–205 (1996).
29. Pariante, C. M. & Lightman, S. L. The HPA axis in major depression: classical theories and new developments. *Trends Neurosci* **31**, 464–468 (2008).
30. Nandam, L. S., Brazel, M., Zhou, M. & Jhaveri, D. J. Cortisol and Major Depressive Disorder—Translating Findings From Humans to Animal Models and Back. *Front Psychiatry* **10**, 974 (2020).
31. Menke, A. Is the HPA axis as target for depression outdated, or is there a new hope? *Front Psychiatry* **10**, 101 (2019).
32. Kadriu, B. *et al.* Glutamatergic Neurotransmission: Pathway to Developing Novel Rapid-Acting Antidepressant Treatments. *International Journal of Neuropsychopharmacology* **22**, 119–135 (2019).
33. Abdallah, C. G., Sanacora, G., Duman, R. S. & Krystal, J. H. The Neurobiology of Depression, Ketamine and Rapid-Acting Antidepressants: Is it Glutamate Inhibition or Activation? *Pharmacol Ther* **190**, 148 (2018).
34. Lener, M. S. *et al.* Glutamate and Gamma-Aminobutyric Acid Systems in the Pathophysiology of Major Depression and Antidepressant Response to Ketamine. *Biol Psychiatry* **81**, 886–897 (2017).
35. Duman, R. S. & Li, N. A neurotrophic hypothesis of depression: role of synaptogenesis in the actions of NMDA receptor antagonists. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**, 2475–2484 (2012).

36. Kojima, M., Matsui, K. & Mizui, T. BDNF pro-peptide: physiological mechanisms and implications for depression. *Cell Tissue Res* **377**, 73–79 (2019).
37. Jeon, S. W. & Kim, Y. K. The role of neuroinflammation and neurovascular dysfunction in major depressive disorder. *J Inflamm Res* **11**, 179 (2018).
38. Jeon, S. W. & Kim, Y. K. The role of neuroinflammation and neurovascular dysfunction in major depressive disorder. *J Inflamm Res* **11**, 179 (2018).
39. Lindqvist, D. *et al.* Oxidative stress, inflammation and treatment response in major depression. *Psychoneuroendocrinology* **76**, 197–205 (2017).
40. Xie, X. *et al.* Nicotinamide mononucleotide ameliorates the depression-like behaviors and is associated with attenuating the disruption of mitochondrial bioenergetics in depressed mice. *J Affect Disord* **263**, 166–174 (2020).
41. Silva, S., Bicker, J., Falcão, A. & Fortuna, A. Antidepressants and Circadian Rhythm: Exploring Their Bidirectional Interaction for the Treatment of Depression. *Pharmaceutics* **13**, 1975 (2021).
42. Winter, G., Hart, R. A., Charlesworth, R. P. G. & Sharpley, C. F. Gut microbiome and depression: What we know and what we need to know. *Rev Neurosci* **29**, 629–643 (2018).
43. Border, R. *et al.* No support for historical candidate gene or candidate gene-by-interaction hypotheses for major depression across multiple large samples. *American Journal of Psychiatry* **176**, 376–387 (2019).
44. Mcintosh, A. M., Sullivan, P. F. & Lewis, C. M. Uncovering the Genetic Architecture of Major Depression. *Neuron* **102**, 91–103 (2019).
45. Duncan, L. E., Ostacher, M. & Ballon, J. How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. *Neuropsychopharmacology* **44**, 1518–1523 (2019).
46. Shadrina, M., Bondarenko, E. A. & Slominsky, P. A. Genetics Factors in Major Depression Disease. *Front Psychiatry* **9**, 334 (2018).
47. Caspi, A. *et al.* Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. *Science (1979)* **301**, 386–389 (2003).

48. Oo, K. Z., Aung, Y. K., Jenkins, M. A. & Win, A. K. Associations of 5HTTLPR polymorphism with major depressive disorder and alcohol dependence: A systematic review and meta-analysis. *Australian & New Zealand Journal of Psychiatry* **50**, 842–857 (2016).
49. Culverhouse, R. C. *et al.* Collaborative meta-analysis finds no evidence of a strong interaction between stress and 5-HTTLPR genotype contributing to the development of depression. *Mol Psychiatry* **23**, 133–142 (2018).
50. Massat, I. *et al.* COMT and age at onset in mood disorders: a replication and extension study. *Neurosci Lett* **498**, 218–221 (2011).
51. Comasco, E. *et al.* Postpartum depression symptoms: a case-control study on monoaminergic functional polymorphisms and environmental stressors. *Psychiatr Genet* **21**, 19–28 (2011).
52. Potter, G. G. *et al.* The COMT Val158Met polymorphism and cognition in depressed and nondepressed older adults. *Int J Geriatr Psychiatry* **24**, 1127–1133 (2009).
53. Frisch, A. *et al.* Association of unipolar major depressive disorder with genes of the serotonergic and dopaminergic pathways. *Mol Psychiatry* **4**, 389–392 (1999).
54. Ferreira Fratelli, C. *et al.* BDNF Genetic Variant and Its Genotypic Fluctuation in Major Depressive Disorder. *Behavioural neurology* **2021**, 7117613 (2021).
55. Sullivan, P. F. How good were candidate gene guesses in schizophrenia genetics? *Biol Psychiatry* **82**, 696 (2017).
56. Bosker, F. J. *et al.* Poor replication of candidate genes for major depressive disorder using genome-wide association data. *Mol Psychiatry* **16**, 516–532 (2011).
57. Zondervan, K. T. & Cardon, L. R. Designing candidate gene and genome-wide case-control association studies. *Nat Protoc* **2**, 2492–2500 (2007).
58. Gordon, J. A. *Towards a genomic psychiatry: recommendations of the genomics workgroup of the NAMHC. DirectorVs Messages published online March* vol. 29 https://www.nimh.nih.gov/about/director/messages/2018/towards-a-genomic-psychiatry-recommendations-of-the-genomics-workgroup-of-the-namhc (2018).
59. Dunn, E. C., Wang, M.-J. & Perlis, R. H. A Summary of Recent Updates on the Genetic Determinants of

Depression. in *Major Depressive Disorder* 1–27 (Elsevier, 2020). doi:10.1016/B978-0-323-58131-8.00001-X.
60. Adebiyi, E., Adam, Y., Samtal, C., Brandenburg, J. tristan & Falola, O. Performing post-genome-wide association study analysis: overview, challenges and recommendations. *F1000Res* **10**, 1002 (2021).
61. Pasaniuc, B. *et al.* Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**, 2906–2914 (2014).
62. Zhou, W. *et al.* Large-scale whole-exome sequencing association study identifies FOXH1 gene and sphingolipid metabolism pathway influencing major depressive disorder. *CNS Neurosci Ther* **27**, 1425–1428 (2021).
63. Curtis, D. Analysis of 200 000 exome-sequenced UK Biobank subjects fails to identify genes influencing probability of developing a mood disorder resulting in psychiatric referral. *Psychiatr Genet* **31**, 194–198 (2021).
64. Borczyk, M., Piechota, M., Rodriguez Parkitna, J. & Korostynski, M. Prospects for personalization of depression treatment with genome sequencing. *Br J Pharmacol* **179**, 4220–4232 (2022).
65. Zuk, O. *et al.* Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences* **111**, E455–E464 (2014).
66. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat Rev Genet* **20**, 467–484 (2019).
67. Sullivan, P. F. *et al.* Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo. *Mol Psychiatry* **14**, 359–375 (2009).
68. Flint, J. & Kendler, K. S. The Genetics of Major Depression. *Neuron* **81**, 484–503 (2014).
69. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics* **101**, 5–22 (2017).
70. Sullivan, P. F. The Psychiatric GWAS Consortium: Big Science Comes to Psychiatry. *Neuron* **68**, 182–186 (2010).
71. Levinson, D. F. *et al.* Genetic studies of major depressive disorder: Why are there no genome-wide association study findings and what can we do about it? *Biol Psychiatry* **76**, 510–512 (2014).

72. Evangelou, E. & Ioannidis, J. P. A. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet* **14**, 379–389 (2013).
73. Sullivan, P. F. *et al.* Psychiatric Genomics: An Update and an Agenda. *Am J Psychiatry* **175**, 15 (2018).
74. Cai, N. *et al.* Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588–591 (2015).
75. Cai, N. *et al.* Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nat Genet* **52**, 437–447 (2020).
76. Levey, D. F. *et al.* Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. *Nat Neurosci* **24**, 954–963 (2021).
77. Sullivan, P. F. Spurious Genetic Associations. *Biol Psychiatry* **61**, 1121–1126 (2007).
78. Jantas, D. Cell-Based Systems of Depression: An Overview. in *Herbal Medicine in Depression* (ed. Grosso, C.) 75–117 (Springer International Publishing, 2016). doi:10.1007/978-3-319-14021-6_3.
79. Colpo, G. D. & Teixeira, A. L. Induced Pluripotent Stem Cells (iPSCs) Technology: Potential Targets for Depression. *Adv Exp Med Biol* **1305**, 493–501 (2021).
80. Nestler, E. J. & Hyman, S. E. Animal models of neuropsychiatric disorders. *Nat Neurosci* **13**, 1161–1169 (2010).
81. Planchez, B., Surget, A. & Belzung, C. Animal models of major depression: drawbacks and challenges. *J Neural Transm* **126**, 1383–1408 (2019).
82. Becker, M., Pinhasov, A. & Ornoy, A. Animal Models of Depression: What Can They Teach Us about the Human Disease? *Diagnostics* **11**, 123 (2021).
83. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005–D1012 (2019).
84. Alsheikh, A. J. *et al.* The landscape of GWAS validation: systematic review identifying 309 validated non-coding variants across 130 human diseases. *BMC Med Genomics* **15**, 1–21 (2022).

85. Lappalainen, T. & MacArthur, D. G. From variant to function in human disease genetics. *Science (1979)* **373**, 1464–1468 (2021).
86. Krier, J. B., Kalia, S. S. & Green, R. C. Genomic sequencing in clinical practice: applications, challenges, and opportunities. *Dialogues Clin Neurosci* **18**, 299 (2016).
87. The GenCC Home Page. *https://thegencc.org/* https://thegencc.org/ (2020).
88. Strande, N. T. *et al.* Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. *Am J Hum Genet* **100**, 895–906 (2017).
89. Thaxton, C. *et al.* Utilizing ClinGen gene-disease validity and dosage sensitivity curations to inform variant classification. *Hum Mutat* **43**, 1031–1040 (2022).
90. Preston, C. G. *et al.* ClinGen Variant Curation Interface: a variant classification platform for the application of evidence criteria from ACMG/AMP guidelines. *Genome Med* **14**, 1–12 (2022).
91. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**, 405–424 (2015).
92. Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Res* **47**, D1038–D1043 (2019).
93. Pavan, S. *et al.* Clinical Practice Guidelines for Rare Diseases: The Orphanet Database. *PLoS One* **12**, e0170365 (2017).
94. Martin, A. R. *et al.* PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet* **51**, 1560–1565 (2019).
95. Zeng, J. *et al.* Widespread signatures of natural selection across human complex traits and functional genomic categories. *Nat Commun* **12**, 1164 (2021).
96. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).

97. Griffith, M. *et al.* CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet* **49**, 170–174 (2017).
98. Tamborero, D. *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med* **10**, 1–8 (2018).
99. Li, M. J. *et al.* GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res* **40**, D1047–D1054 (2012).
100. Beck, T., Shorter, T. & Brookes, A. J. GWAS Central: a comprehensive resource for the discovery and comparison of genotype and phenotype data from genome-wide association studies. *Nucleic Acids Res* **48**, D933–D940 (2020).
101. Landrum, M. J. & Kattman, B. L. ClinVar at five years: Delivering on the promise. *Hum Mutat* **39**, 1623–1630 (2018).
102. Gutierrez-Sacristan, A. *et al.* PsyGeNET: a knowledge platform on psychiatric disorders and their genes. *Bioinformatics* **31**, 3075 (2015).
103. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature 2009 461:7265* **461**, 747–753 (2009).
104. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front Genet* **11**, 424 (2020).
105. Stram, D. O. Tag SNP selection for association studies. *Genet Epidemiol* **27**, 365–374 (2004).
106. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* **19**, 491–504 (2018).
107. Slatkin, M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* **9**, 477–485 (2008).
108. Pereira, L., Mutesa, L., Tindana, P. & Ramsay, M. African genetic diversity and adaptation inform a precision medicine agenda. *Nat Rev Genet* **22**, 284–306 (2021).
109. LaPierre, N. *et al.* Identifying causal variants by fine mapping across multiple studies. *PLoS Genet* **17**, e1009733 (2021).

110. Nicolae, D. L. *et al.* Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLoS Genet* **6**, e1000888 (2010).
111. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends in Genetics* **37**, 109–124 (2021).
112. Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Smith, G. D. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Stat Med* **27**, 1133–1163 (2008).
113. Davies, N. M., Holmes, M. v. & Davey Smith, G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ* **362**, 601 (2018).
114. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc* **15**, 2759–2772 (2020).
115. Hu, X. *et al.* Integrating Autoimmune Risk Loci with Gene-Expression Data Identifies Specific Pathogenic Immune Cell Subsets. *The American Journal of Human Genetics* **89**, 496–506 (2011).
116. Li, B. & Ritchie, M. D. From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries. *Front Genet* **12**, 1502 (2021).
117. Hukku, A., Sampson, M. G., Luca, F., Pique-Regi, R. & Wen, X. Analyzing and reconciling colocalization and transcriptome-wide association studies from the perspective of inferential reproducibility. *The American Journal of Human Genetics* **109**, 825–837 (2022).
118. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res* **22**, 1748–1759 (2012).
119. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res* **48**, D882–D889 (2020).
120. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–30 (2015).
121. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).

122. Genotype-Tissue Expression. GTEx Portal. https://www.gtexportal.org/home/datasets (2017).
123. Liu, X., Li, Y. I. & Pritchard, J. K. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* **177**, 1022-1034.e6 (2019).
124. Degtyareva, A. O., Antontseva, E. v. & Merkulova, T. I. Regulatory SNPs: Altered Transcription Factor Binding Sites Implicated in Complex Traits and Diseases. *Int J Mol Sci* **22**, 6454 (2021).
125. Jayaram, N., Usvyat, D. & Martin, A. C. Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics* **17**, 547 (2016).
126. Stead, J. A., Keen, J. N. & McDowall, K. J. The identification of nucleic acid-interacting proteins using a simple proteomics-based approach that directly incorporates the electrophoretic mobility shift essay. *Molecular and Cellular Proteomics* **5**, 1697–1702 (2006).
127. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886–D894 (2019).
128. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812–3814 (2003).
129. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Non-synonymous and Splice Site SNVs. *Hum Mutat* **37**, 235 (2016).
130. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763 (2015).
131. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *The American Journal of Human Genetics* **102**, 717–730 (2018).
132. Ghazi, A. R. *et al.* Design tools for MPRA experiments. *Bioinformatics* **34**, 2682–2683 (2018).
133. Corradin, O. *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res* **24**, 1–13 (2014).

134. Gaj, T., Sirk, S. J., Shui, S. & Liu, J. Genome-Editing Technologies: Principles and Applications. *Cold Spring Harb Perspect Biol* **8**, a023754 (2016).
135. Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nat Rev Mol Cell Biol* **16**, 245–257 (2015).
136. Pal, K., Forcato, M. & Ferrari, F. Hi-C analysis: from data generation to integration. *Biophys Rev* **11**, 67–78 (2019).
137. Marti-Renom, M. A. *et al.* Challenges and guidelines toward 4D nucleome data and model standards. *Nat Genet* **50**, 1352–1358 (2018).
138. García-Campos, M. A., Espinal-Enríquez, J. & Hernández-Lemus, E. Pathway analysis: State of the art. *Front Physiol* **6**, 383 (2015).
139. Oliveira, S. *et al.* Impact of genetic variations in ADORA2A gene on depression and symptoms: a cross-sectional population-based study. *Purinergic Signal* **15**, 37–44 (2019).
140. Harrison, P. J., Mould, A. & Tunbridge, E. M. New drug targets in psychiatry: Neurobiological considerations in the genomics era. *Neurosci Biobehav Rev* **139**, 104763 (2022).
141. Gauderman, W. J. Sample Size Requirements for Association Studies of Gene-Gene Interaction. *Am J Epidemiol* **155**, 478–484 (2002).
142. Manolio, T. A. Bringing genome-wide association findings into clinical use. *Nat Rev Genet* **14**, 549–558 (2013).
143. Akiyama, M. Multi-omics study for interpretation of genome-wide association study. *J Hum Genet* **66**, 3–10 (2021).
144. Taylor, K. E., Ansel, K. M., Marson, A., Criswell, L. A. & Farh, K. K.-H. PICS2: next-generation fine mapping via probabilistic identification of causal SNPs. *Bioinformatics* **37**, 3004–3007 (2021).
145. Rüeger, S., McDaid, A. & Kutalik, Z. Evaluation and application of summary statistic imputation to discover new height-associated loci. *PLoS Genet* **14**, e1007371 (2018).
146. Pérez-Granado, J., Piñero, J. & Furlong, L. I. Benchmarking post-GWAS analysis tools in major

depression: Challenges and implications. *Front Genet* **13**, 2853 (2022).
147. Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genet* **13**, e1006646 (2017).
148. Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat Genet* **53**, 1527–1533 (2021).
149. Burgess, D. J. Fine-mapping causal variants — why finding 'the one' can be futile. *Nat Rev Genet* **23**, 261–261 (2022).
150. Wen, X. Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control. *Ann Appl Stat* **10**, 1619–1638 (2016).
151. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am J Hum Genet* **99**, 1245–1260 (2016).
152. Pividori, M. *et al.* PhenomeXcan: Mapping the genome to the phenome through the transcriptome. *Sci Adv* **6**, (2020).
153. Hukku, A. *et al.* Probabilistic colocalization of genetic variants from complex and molecular traits: promise and limitations. *Am J Hum Genet* **108**, 25–35 (2021).
154. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173–1186 (2014).
155. Zuber, V. *et al.* Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. *Am J Hum Genet* **109**, 767–782 (2022).
156. Dehghan, A. Genome-wide association studies. *Methods in Molecular Biology* **1793**, 37–49 (2018).
157. Minton, K. Decoding noncoding variation in dementia. *Nat Rev Genet* **23**, 649–649 (2022).