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

Judith Pérez-Granado
Institut Hospital del Mar d'Investigacions Mèdiques
Janet Piñero
Institut Hospital del Mar d'Investigacions Mèdiques
Alejandra Medina-Rivera
Universidad Nacional Autónoma de México
Laura I. Furlong (laura.furlong@upf.edu)
Institut Hospital del Mar d'Investigacions Mèdiques

Research Article

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

Posted Date: January 5th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1200645/v1

Abstract

Background: Major Depression is the leading cause of impairment worldwide. The understanding of its molecular underpinnings is key to identifying new potential biomarkers and drug targets to alleviate its burden in 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 using GTEx data and alteration of transcription factor binding sites with pattern matching approaches and chromatin accessibility data.

Results: The fine-mapping of major depression genetic variants uncovered putative causally associated variants whose proximal genes were linked with Major Depression pathophysiology. Four genetic variants altering the expression of 5 genes were found by colocalization analysis, highlighting the role of SLC12A5, involved in chlorine homeostasis in neurons, and MYRF, related with central nervous system myelination and oligodendrocyte differentiation. The transcription factor binding analysis revealed the potential role of the genomic variant rs62259947 in modulating the expression of P4HTM through the alteration of YY1 binding site, altogether regulating hypoxia response.

Conclusions: The combination of GWAS signals, cis-eQTL, transcription factor binding site information and active regulatory regions in the chromatin, enabled the prioritization of putative causal genetic variants in Major Depression. Importantly, our pipeline 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.

Background

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 are only effective in 40% of the patients [2]. Currently, there are no available biomarkers nor 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 molecular underpinnings of the disease and identifying biomarkers or new drug targets. Some authors propose that the efforts should be centred on the interpretation of GWAS signals to identify the causal GVs 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 and 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 towards 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 in 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.

Results

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 ≤ 5e-8, which were the starting point of our analysis. After LD expansion, we obtained a set of 5,723 potentially causal genetic variants (GVs) (Supplementary Figure 1). We annotated these GVs with variant effect predictor (VEP) [9] and Combined Annotation Dependent Depletion (CADD) [10] (Supplementary Figure 2). The identification of probable causal GVs using PICS fine-mapping GWAS data [11] revealed 172 GVs (PICS >10%) structurally linked with the 97 GWAS risk loci (Supplementary Table 1). These GVs are located in different regions of the genome, but most of them 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 (±5,000bps), from now on referred to as pGenes. pGenes were classified based on their expression across tissues based on GTEx gene expression data [12]. Using hierarchical clustering, genes were divided in three roughly equally distributed clusters that seem to correspond to: constitutively, lowly expressed in all tissues, and highly expressed in brain tissues (Supplementary Figure 3).

The pGenes are functionally enriched in GO terms related to nervous system development, neuron differentiation, synaptic signalling, and different cellular components of the neuron such as dendrite,

axon or synapse (Supplementary Table 2); these biological processes and molecular functions are involved in the pathophysiology of MD [13]. pGenes are associated with an abnormal nervous system morphology and physiology according to the Human Phenotype (HP) ontology. Disease enrichment analysis shows an 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) (Supplementary Table 3). pGenes are also associated with comorbid phenotypes and conditions in MD such as smoking behavior, body mass index and duration of sleep [14]. Notably, 37% of pGenes and 42% of GVs have no previous evidence of association to depression or other mental disorders.

Some of the pGenes are associated to processes related to MD pathogenesis such as TLR4, involved in immune response [15]; ESR2, regulator of estrogen response [16]; TCF4 with a role in nervous system development [17]; DCC in charge of axon guidance and neuronal connectivity [18]; PAX5 which interferes in mouse neural stem/progenitor cells proliferation and migration [19, 20]; and CYP7B1 that participates in the metabolism of the neurosteroids DHEA and pregnenolone [21]. Among the potentially pathogenic GVs according to CADD, there are 3 intronic GVs lying in ZNF536, a gene involved in negative regulation of neuron differentiation [22], a relevant process in MD pathogenesis and treatment [23]; rs1021362 lies in SORCS3, a gene previously associated with stress response associated to MD [19, 24]; rs3793577 lies in ELAVL2, whose silencing in animal models is associated with reduced behavioral despair [25]; the remaining GVs have been previously associated with major depression by several PheWAS studies [26].

Major depression causal genetic variants regulate the expression of genes in cis

The 172 fine-mapped GWAS GVs overlap with 13 GTEx PICS GVs (Figure 3), revealing an enrichment of MD causal GVs in eQTLS (p-value = 7.392e-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 [27, 28] and MYRF in central nervous system myelination [29]. In addition, all eQTLs have been previously associated with MD and other mental disorders according to DISGENET plus [30–32] Supplementary Table 4). The eGenes BAG5, SLC12A5 and MYRF show higher expression levels in brain regions according to GTEx (Supplementary Figure 4). Little is known about the function of the long non-coding RNAs RP11-894P9.2 and RP5-1115A15.1.

Table 1
GWAS-eQTL Colocalizing GVs in MD.

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

MD associated GVs colocalizing with eQTLs.

MD associated GVs affect the TFBS in regulatory regions of genes relevant for the disease

The initial set of 5,723 GVs associated with MD was first mapped to transcription factor binding sites (TFBS) using Chip-Seq data from ENCODE. 955 GVs were identified as potentially altering the TFBS of 155 TFs (Figure 4). The GVs' functional impact was assessed with VEP and four 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 (≥2TPM), which left 115 TFs.

Using a pattern matching approach (*variation-scan*) [33], 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 four 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 5). 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 6) [34, 35]. We then looked for evidence of their impact in 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 fulfils 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 [31].

In addition, we also inspected hTFtarget database [36] looking for evidence of mechanistic association between the eGenes, considered as 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).

The GV rs62259947 has been annotated as an eQTL downregulating the expression of P4HTM in 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 [37]. The eGene P4HTM has been associated with the hypoxia-inducible factor HIF1α mediating calcium signalling [38] and its null mutation reduces behavioural despair [39] (Figure 5)

Discussion

Despite the large volume of genetic information available, the pathogenesis and aetiology of MD are not yet fully understood; probably because most GV 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 (specially 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 [40–42] and with MD in particular [43–45]. The regulatory variation analysis pipelines we have designed and 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 [46].

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 1). 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 [13]. Additionally, these are also enriched for other phenotypes frequently co-occurring with MD like alterations of body mass index or smoking [14]. While most pGenes (63%) and GVs (58%) have previous evidence for association with MD, our study pinpoints novel pGenes and GVs (Supplementary Table 2 and 3). 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 behavioural 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 [40, 47]. The colocalizing eGenes are involved in processes relevant to MD, such as neuron projection [48] and have been associated with MD and related phenotypes according to DISGENET plus [30, 31]. BAG5 is constitutively expressed in all tissues; while MYRF and SLC12A5 show higher levels in brain tissues (Supplementary Figure 3). BAG5 has been previously identified as associated with MD [49]. We characterize SLC12A5, involved in chloride homeostasis in neurons, as a pGene too, and its downregulation has been described as an effect from stress leading to the activation of the hypothalamic-pituitary-adrenal axis which ultimately can lead to MD-like symptoms [50, 51]. However, rs12624433 is an eQTL in 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 on regulation of 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 [50]. rs198457 mediates the downregulation of MYRF expression, which plays a role in myelination and oligodendrocyte differentiation [29]. These, in turn, require thyroid hormones for their differentiation and maturation [52].

Furthermore, oligodendrocytes have been stated as crucial for psychological functions likely involved in mental disorders such as MD [53].

The analysis of TF regulation with RSAT enabled a precise prediction of TF binding alterations. Although TF expression is not highly tissue-specific [7, 54], for this type of analysis it is key to pick meaningful sets of TFs and GVs [55]. 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, 14, 19, 30]. 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 [54]. Thus, further analysis is required to elucidate the impact of these GVs in 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 [56]. USF1 and USF2 generally exert activating effects [57], with USF1 being a risk gene for Alzheimer's disease and relevant for brain cholesterol metabolism involving its transport from astrocytes to neurons [58].

Besides, 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. ZNRD2-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 being clear their exact role in MD pathophysiology, ncRNAs have been described as promising biomarkers and drug targets for MD [59, 60].

Regarding the association rs62259947: YY1 → P4HTM, P4HTM has been related with neurological disorders and social behaviour [38, 39]. It is involved in Ca^{2+} signalling mediated by the hypoxia-inducible factor HIF1α altering astrocytes gliotransmission [38]. Indeed, hypoxia has been associated with mental disorders in general and MD in particular [61–64]. In addition, P4HTM null mutation results in a reduction in fear and depression [39]. In turn, rs62259947 downregulates the expression of P4HTM and changes the binding affinity of YY1 in the Brain Cerebellar Hemisphere. Besides, YY1 regulates transcription by forming loops although its specific role as activator or repressor is not yet fully understood [37]. Furthermore, P4HTM and HIF1α have been reported as potential drug targets for MD [39, 65]. rs11227217 and RAD21 are associated with red blood cell and reticulocyte count respectively by PheWAS [26]. Indeed, red blood cells parameters have been described as altered in patients with mental disorders [66].

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 of MD and mental disorders as well as novel molecular mechanisms underlying MD pathogenesis.

Methods

MD GWAS Dataset and LD expansion

In order to have 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]. The genetic variants (GVs) associated with MD were retrieved from the summary statistics available at GWAS Catalog [67] (Accession Study: GCST007342, note that the full-genome summary statistics for this GWAS were not publicly available; downloaded in December 2020). We filtered the 102 risk loci by significance (p-value $\leq$ 5e-8) and proceeded the analysis with this set. We further fine-mapped MD associated GVs to prioritize the causal ones using the Probabilistic Identification of causal SNPs (PICS) algorithm [11], which takes into consideration Linkage Disequilibrium (LD) structure. Using the PICS2 Data portal, we downloaded the precomputed PICS GVs for this study. This data constituted our full dataset of GVs.

GVs annotation: VEP, CADD and ENCODE

We annotated the full set of GVs with Variant Effect Predictor (VEP) [9] and Combined Annotation Dependent Depletion (CADD) [10]. 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 ±5,000bp distance (from now on pGenes). CADD assesses GVs' potential pathogenicity by evaluating the PHRED-like scaled C-score; the recommended cut-off $\geq$ 15 was set to identify potentially pathogenic variants.

We analysed the GVs with ENCODE [68] to identify those potentially lying in transcription factor binding sites (TFBS). ENCODE data analysis was performed using SNPNexus [69], an online platform that allows a comprehensive annotation of GVs by integrating multiple tools.

Fine-mapping and colocalization of GWAS and cis-eQTLs

PICS2, in addition to GWAS PICS GVs, has precomputed PICS GVs for all GTEx V8 best expression quantitative trait loci (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. 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 CLPP method, which assumes independence of causal probabilities for GWAS and eQTL GVs [70]. The genes regulated by these eQTLs from now on will be referred to as eGenes.

TF binding analysis with RSAT variation tools

We predicted those GVs affecting the TFBS using the 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 retrieves regions of around 100-1,000bp but the actual binding site corresponds to 9-15bp [71, 72]. Thus, we proceeded with the RSAT analysis for a more robust and accurate assessment of the GVs potentially altering the TBFS. 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 get the flanking sequence (30bp per side) of the GVs of interest, using the dbSNP, genomic coordinates, reference and alternative allele.

To assess the TFBS alterations, PSSM matrices for TFs expressed in brain tissues (filtering them using GTEx expression data, ≥2TPM) [12] were retrieved from the following databases: JASPAR [73], ENCODE, HOCOMOCO [74], footprintDB [75] and hPDI [76]. 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 (<1e-3) and p-value ratio between variants (>10). The weight (*W*) represents the binding affinity and the p-value of a score is the probability of observing a score of at least W 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 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 [77] with the following parameters: distance to closest transcription start site (TSS) deviation (±5,000), gene density deviation (±5 in 100kbp), 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 ratio against the distribution of p-value ratios for the given motif in both control datasets. A Wilcox test was undergone to evaluate the results obtained from the controls because normality of p-value ratio distribution couldn't 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 (ie. 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.

Identification of TF active regions with ChromHMM

We used chromatin state annotations from ChromHMM [34, 35], 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 [35]: 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 was considered.

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 hTFtarget database [36] 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. 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.

pGenes, eGenes and GVs characterization

We conducted a gene-set enrichment analysis using the tool g:Profiler via the R package gprofiler2 [78], 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 (WK); 3) miRNA annotations from MIRNA; 4) phenotypic features associated to disease from Human Phenotype (HP), which is mainly focused on rare Mendelian disorders. In addition, we included DISGENET plus [32, 79] association data (v16) in this analysis to evaluate the annotation to complex diseases and phenotypic traits; note that the study by Howard, et al. was removed from these 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), 2) genes regulated by cis-eQTLs (eGenes); and two variant-sets: 1) causal GVs and 2) colocalizing GVs.

Abbreviations

-CADD: Combined Annotation Dependent Depletion

-eGenes: genes regulated by eQTLs

-eQTLs: expression quantitative trait loci

-GO: Gene Ontology

-GV: genetic variant

-GTEx: Genotype-Tissue Expression

-GWAS: genome-wide association studies

-HP: Human Phenotype

-LD: linkage disequilibrium

-MD: major depression

-pGenes: proximal genes

-PICS: probabilistic identification of causal snps

-PSSM: position weight matrix

-REAC: Reactome

-TF: transcription factor

-TPM: transcripts per million

-TFBS: transcription factor binding site

-VEP: variant effect predictor

-W: weight

-WK: WikiPathways

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

This study has 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, and . The GTEx RNA-Seq data can be downloaded from (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 (, see View matrix descriptions & download full collections); and hTFtarget data can be downloaded from (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.5812975) and are also available as the Supplementary tables file with this manuscript.

Competing interests.

LIF and JP are co-founders and hold shares of Medbioinformatics Solutions SL.

JPG and AMR have no competing interests.

Funding

IMI2-JU resources of which are composed of financial contribution 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]. The Research Programme on Biomedical Informatics (GRIB) is a member of the Spanish National Institute (INB), funded by ISCIII and ERDF (PRB2-ISCIII [PT13/0001/0023, of the PE I + D + i 2013–2016]). The DCEXS 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.

Authors' contributions

JPG developed and implemented the analysis pipeline, and wrote the manuscript with the support and guidance of JP, LIF and AMR. All authors reviewed the manuscript. The authors read and approved the final manuscript.

Acknowledgements

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.

References

1. World Health Organization: Depression. https://www.who.int/news-room/fact-sheets/detail/depression. Accessed 10 December 2020.
2. Preskorn SH. Drug Development in Psychiatry: The Long and Winding Road from Chance Discovery to Rational Development. In: Handbook of Experimental Pharmacology. p. 307–24. doi:10.1007/164_2018_169.
3. Howard DM, Adams MJ, Clarke T-K, Hafferty JD, Gibson J, Shirali M, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nature Neuroscience. 2019;22:343–52. doi:10.1101/433367.
4. Wray, Naomi R.Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., … Sullivan PF. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat Genet. 2018;50:668–81. doi:10.1038/s41588-018-0090-3.

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. doi:10.3389/fgene.2020.00424.
6. Umans BD, Battle A, Gilad Y. Where Are the Disease-Associated eQTLs? Trends Genet. 2021;37:109–24. doi:10.1016/j.tig.2020.08.009.
7. Hu H, Miao YR, Jia LH, Yu QY, Zhang Q, Guo AY. AnimalTFDB 3.0: A comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res. 2019;47:D33–D38. doi:10.1093/nar/gky822.
8. Perdomo-Sabogal A, Nowick K. Genetic Variation in Human Gene Regulatory Factors Uncovers Regulatory Roles in Local Adaptation and Disease. Genome Biol. 2019;11:2178–2193. doi:10.1093/gbe/evz131.
9. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17:122. doi:10.1186/s13059-016-0974-4.
10. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47. doi:10.1093/nar/gky1016.
11. Taylor KE, Ansel KM, Marson A, Criswell LA, Farh KK-H. PICS2: next-generation fine mapping via probabilistic identification of causal SNPs. Bioinformatics. 2021;37:3004–7. doi:10.1093/BIOINFORMATICS/BTAB122.
12. GTEx Portal. https://www.gtexportal.org/home/datasets. Accessed 11 January 2021.
13. Shadrina M, Bondarenko EA, Slominsky PA. Genetics Factors in Major Depression Disease. Front Psychiatry. 2018;9:334. doi:10.3389/fpsyt.2018.00334.
14. Mcintosh AM, Sullivan PF, Lewis CM. Uncovering the Genetic Architecture of Major Depression. Neuron. 2019;102:91–103. doi:10.1016/j.neuron.2019.03.022.
15. 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. doi:10.1002/BRB3.1525.
16. Keyes K, Agnew-Blais J, Roberts AL, Hamilton A, De Vivo I, Ranu H, et al. The role of allelic variation in estrogen receptor genes and major depression in the Nurses Health Study. Soc Psychiatry Psychiatr Epidemiol. 2015;50:1893. doi:10.1007/S00127-015-1087-1.
17. Mossakowska-Wójcik J, Orzechowska A, Talarowska M, Szemraj J, Gałecki P. The importance of TCF4 gene in the etiology of recurrent depressive disorders. Prog Neuro-Psychopharmacology Biol Psychiatry. 2018;80:304–8.
18. Torres-Berrío A, Lopez JP, Bagot RC, Nouel D, Dal Bo G, Cuesta S, et al. DCC confers susceptibility to depression-like behaviors in humans and mice and is regulated by miR-218. Biol Psychiatry. 2017;81:306. doi:10.1016/J.BIOPSYCH.2016.08.017.
19. Hyde CL, Nagle MW, Tian C, Chen X, Paciga SA, Wendland JR, et al. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent HHS Public Access Author manuscript. Nat Genet. 2016;48:1031–6.

20. Wu Q, Tang W, Luo Z, Li Y, Shu Y, Yue Z, et al. DISC1 regulates the proliferation and migration of mouse neural stem/progenitor cells through Pax5, Sox2, Dll1 and Neurog2. Front Cell Neurosci. 2017;11:261.
21. Rudzinskas S, Hoffman JF, Martinez P, Rubinow DR, Schmidt PJ, Goldman D. In vitro model of perimenopausal depression implicates steroid metabolic and proinflammatory genes. Mol Psychiatry 2020 267. 2020;26:3266–76. doi:10.1038/s41380-020-00860-x.
22. Qin Z, Ren F, Xu X, Ren Y, Li H, Wang Y, et al. 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. doi:10.1128/MCB.00362-09.
23. 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;83:1054–64. doi:10.1046/J.1471-4159.2002.01215.X.
24. Lanshakov DA, Sukhareva E V., Bulygina V V., Bannova A V., Shaburova E V., Kalinina TS. 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 Reports 2021 111. 2021;11:1–13. doi:10.1038/s41598-021-87652-7.
25. Sanna MD, 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;135:444–54.
26. Ghoussaini M, Mountjoy E, Carmona M, Peat G, Schmidt EM, Hercules 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. doi:10.1093/nar/gkaa840.
27. Beilina A, Rudenko IN, Kaganovich A, Civiero L, Chau H, Kalia SK, 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 U S A. 2014;111:2626–31. doi:10.1073/PNAS.1318306111/-/DCSUPPLEMENTAL.
28. Dzhala VI, Talos DM, Sdrulla DA, Brumback AC, Mathews GC, Benke TA, et al. NKCC1 transporter facilitates seizures in the developing brain. Nat Med 2005 1111. 2005;11:1205–13. doi:10.1038/nm1301.
29. Bujalka H, Koenning M, Jackson S, Perreau VM, Pope B, Hay CM, et al. MYRF Is a Membrane-Associated Transcription Factor That Autoproteolytically Cleaves to Directly Activate Myelin Genes. PLoS Biol. 2013;11:1001625. doi:10.1371/JOURNAL.PBIO.1001625.
30. Group of the Psychiatric Genomics Consortium C-D, Lee PH, Anttila V, Won H, Feng Y-CA, Rosenthal J, et al. Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. Cell. 2019;179:1469-1482.e11. doi:10.1016/j.cell.2019.11.020.
31. Yao X, Glessner JT, Li J, Qi X, Hou X, Zhu C, et al. Integrative analysis of genome-wide association studies identifies novel loci associated with neuropsychiatric disorders. Transl Psychiatry. 2021;11:69. doi:10.1038/S41398-020-01195-5.

32. DISGENET Plus. https://beta.disgenetplus.com/. Accessed 20 September 2021.
33. Santana-Garcia W, Rocha-Acevedo M, Ramirez-Navarro L, Mbouamboua Y, Thieffry D, Thomas-Chollier M, et al. RSAT variation-tools: An accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding. Comput Struct Biotechnol J. 2019;17:1415–28. doi:10.1016/j.csbj.2019.09.009.
34. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc. 2017;12:2478–92.
35. 1. Annotation of the non-coding genome. Nature. 2015. doi:10.1038/nature14309.
36. Zhang Q, Liu W, Zhang HM, Xie GY, Miao YR, Xia M, et al. hTFtarget: A Comprehensive Database for Regulations of Human Transcription Factors and Their Targets. Genomics, Proteomics Bioinforma. 2020;18:120–8.
37. Verheul TCJ, van Hijfte L, Perenthaler E, Barakat TS. The Why of YY1: Mechanisms of Transcriptional Regulation by Yin Yang 1. Front Cell Dev Biol. 2020;8:592164. doi:10.3389/FCELL.2020.592164/BIBTEX.
38. Byts N, Sharma S, Laurila J, Paudel P, Miinalainen I, Ronkainen VP, et al. Transmembrane Prolyl 4-Hydroxylase is a Novel Regulator of Calcium Signaling in Astrocytes. eNeuro. 2021;8:1–23. doi:10.1523/ENEURO.0253-20.2020.
39. Leinonen H, Koivisto H, Lipponen HR, Matilainen A, Salo AM, Dimova EY, et al. Null mutation in P4h-tm leads to decreased fear and anxiety and increased social behavior in mice. Neuropharmacology. 2019;153:63–72.
40. Bhalala OG, Nath AP, Inouye M, Sibley CR. Identification of expression quantitative trait loci associated with schizophrenia and affective disorders in normal brain tissue. PLOS Genet. 2018;14:e1007607. doi:10.1371/JOURNAL.PGEN.1007607.
41. Li S, Li X, Liu J, Huo Y, Li L, Wang J, et al. 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.
42. Roksana Z. Transcription Factors in Schizophrenia: A Current View of Genetic Aspects. Sci J Genet Gene Ther. 2016;2:017–21.
43. Li X, Su X, Liu J, Li H, Li M, Li W, et al. Transcriptome-wide association study identifies new susceptibility genes and pathways for depression. Transl Psychiatry. 2021;11:306. doi:10.1038/S41398-021-01411-W.
44. Zhong J, Li S, Zeng W, Li X, Gu C, Liu J, et al. Integration of GWAS and brain eQTL identifies FLOT1 as a risk gene for major depressive disorder. Neuropsychopharmacol 2019 449. 2019;44:1542–51. doi:10.1038/s41386-019-0345-4.
45. Santos-Terra J, Deckmann I, Fontes-Dutra M, Schwingel GB, 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;81:545–78. doi:10.1002/JDN.10141.

46. Burt C, Munafò M. Has GWAS lost its status as a paragon of open science? PLoS Biol. 2021;19. doi:10.1371/JOURNAL.PBIO.3001242.
47. O'Brien HE, Hannon E, Hill MJ, Toste CC, Robertson MJ, Morgan JE, et al. Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders 06 Biological Sciences 0604 Genetics 11 Medical and Health Sciences 1109 Neurosciences. Genome Biol. 2018;19:1–13. doi:10.1186/S13059-018-1567-1/FIGURES/4.
48. Hare BD, Duman RS. Prefrontal cortex circuits in depression and anxiety: contribution of discrete neuronal populations and target regions. Mol Psychiatry. 2020;25:2742. doi:10.1038/S41380-020-0685-9.
49. Amare AT, Vaez A, Hsu YH, Direk N, Kamali Z, Howard DM, 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. 2020;25:1420. doi:10.1038/S41380-018-0336-6.
50. 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 101. 2020;10:1–11. doi:10.1038/s41398-020-01027-6.
51. Shadrina M, Bondarenko EA, Slominsky PA. Genetics Factors in Major Depression Disease. Front Psychiatry. 2018;9:334. doi:10.3389/fpsyt.2018.00334.
52. 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.
53. Zhou B, Zhu Z, Ransom BR, Tong X. Oligodendrocyte lineage cells and depression. Mol Psychiatry 2020 261. 2020;26:103–17. doi:10.1038/s41380-020-00930-0.
54. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. Cell. 2018;172:650–65.
55. Andersen MC, Engström PG, Lithwick S, Arenillas D, Eriksson P, Lenhard B, et al. In silico detection of sequence variations modifying transcriptional regulation. PLoS Comput Biol. 2008;4:0043–54. doi:10.1371/journal.pcbi.0040005.
56. Dennis DJ, Han S, Schuurmans C. bHLH transcription factors in neural development, disease, and reprogramming. Brain Res. 2019;1705:48–65.
57. Rada-Iglesias A, Ameur A, Kapranov P, Enroth S, Komorowski J, Gingeras TR, et al. 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;18:380–92. doi:10.1101/GR.6880908.
58. Sertbaş M, Ülgen K, Çakir 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. doi:10.1016/J.FOB.2014.05.006.

59. Shi Y, Wang Q, Song R, Kong Y, Zhang Z. Non-coding RNAs in depression: Promising diagnostic and therapeutic biomarkers. EBioMedicine. 2021;71:103569. doi:10.1016/J.EBIOM.2021.103569.
60. Żurawek D, Turecki G. The miRNome of Depression. Int J Mol Sci. 2021;22:11312. doi:10.3390/IJMS222111312.
61. Bian Z, Li H, Liu Y, Cao Y, Kang Y, Yu Y, et al. The Association Between Hypoxia Improvement and Electroconvulsive Therapy for Major Depressive Disorder. Neuropsychiatr Dis Treat. 2021;17:2987–94. doi:10.2147/NDT.S318919.
62. Li G, Zhao M, Cheng X, Zhao T, Feng Z, Zhao Y, et al. FG-4592 Improves Depressive-Like Behaviors through HIF-1-Mediated Neurogenesis and Synapse Plasticity in Rats. Neurotherapeutics. 2020;17:664. doi:10.1007/S13311-019-00807-3.
63. Ding F-S, Cheng X, Zhao T, Zhao Y, Zhang G-B, Wu H-T, et al. Intermittent hypoxic preconditioning relieves fear and anxiety behavior in post-traumatic stress model mice. Sheng li xue bao: [Acta physiologica Sinica]. 2019;71(4):537–546.
64. Shibata T, Yamagata H, Uchida S, Otsuki K, Hobara T, Higuchi F, et al. The alteration of hypoxia inducible factor-1 (HIF-1) and its target genes in mood disorder patients. Prog Neuro-Psychopharmacology Biol Psychiatry. 2013;43:222–9.
65. Kang I, Kondo D, Kim J, Lyoo IK, Yurgelun-Todd D, Hwang J, et al. Elevating the level of hypoxia inducible factor may be a new potential target for the treatment of depression. Med Hypotheses. 2021;146:110398.
66. Szczepocka E, Wysokiński A. Red Blood Cells Parameters in Patients with Acute Schizophrenia, Unipolar Depression and Bipolar Disorder. Psychiatr Danub. 2018;30:323–30. doi:10.24869/PSYD.2018.323.
67. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47:D1005–12. doi:10.1093/nar/gky1120.
68. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74. doi:10.1038/nature11247.
69. Oscanoa J, Sivapalan L, Gadaleta E, Dayem Ullah AZ, Lemoine NR, Chelala C. SNPnexus: A web server for functional annotation of human genome sequence variation (2020 update). Nucleic Acids Res. 2020;48:W185–W192. doi:10.1093/NAR/GKAA420.
70. Hormozdiari F, van de Bunt M, Segrè A V., Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet. 2016;99:1245–60. doi:10.1016/J.AJHG.2016.10.003/ATTACHMENT/1004BE78-A72A-4117-B65B-FA553FEA5C74/MMC1.PDF.
71. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012;22:1813–1831. doi:10.1101/gr.136184.111.

72. Jayaram N, Usvyat D, Martin AC. Evaluating tools for transcription factor binding site prediction. BMC Bioinformatics. 2016;17:547. doi:10.1186/s12859-016-1298-9.
73. Fornes O, Castro-Mondragon JA, Khan A, Van Der Lee R, Zhang X, Richmond PA, et al. JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. Nucleic Acids Res. 2020;48:D87–D92. doi:10.1093/nar/gkz1001.
74. Kulakovskiy I V., Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, 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. 2018;46:D252–9. doi:10.1093/nar/gkx1106.
75. Sebastian A, Contreras-Moreira B. footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. Bioinformatics. 2014;30:258–65. doi:10.1093/bioinformatics/btt663.
76. Xie Z, Hu S, Blackshaw S, Zhu H, Qian J. hPDI: A database of experimental human protein-DNA interactions. Bioinformatics. 2010;26:287–9. doi:10.1093/bioinformatics/btp631.
77. Huang D, Wang Z, Zhou Y, Liang Q, Sham PC, Yao H, et al. vSampler: fast and annotation-based matched variant sampling tool. Bioinformatics. 2020. doi:10.1093/bioinformatics/btaa883.
78. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47:191–8. doi:10.1093/nar/gkz369.
79. Piñero J, Piñero P, Manuel Ramírez-Anguita J, Sä Uch-Pitarch J, Ronzano F, Centeno E, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2019;48:845–55. doi:10.1093/nar/gkz1021.

Figures

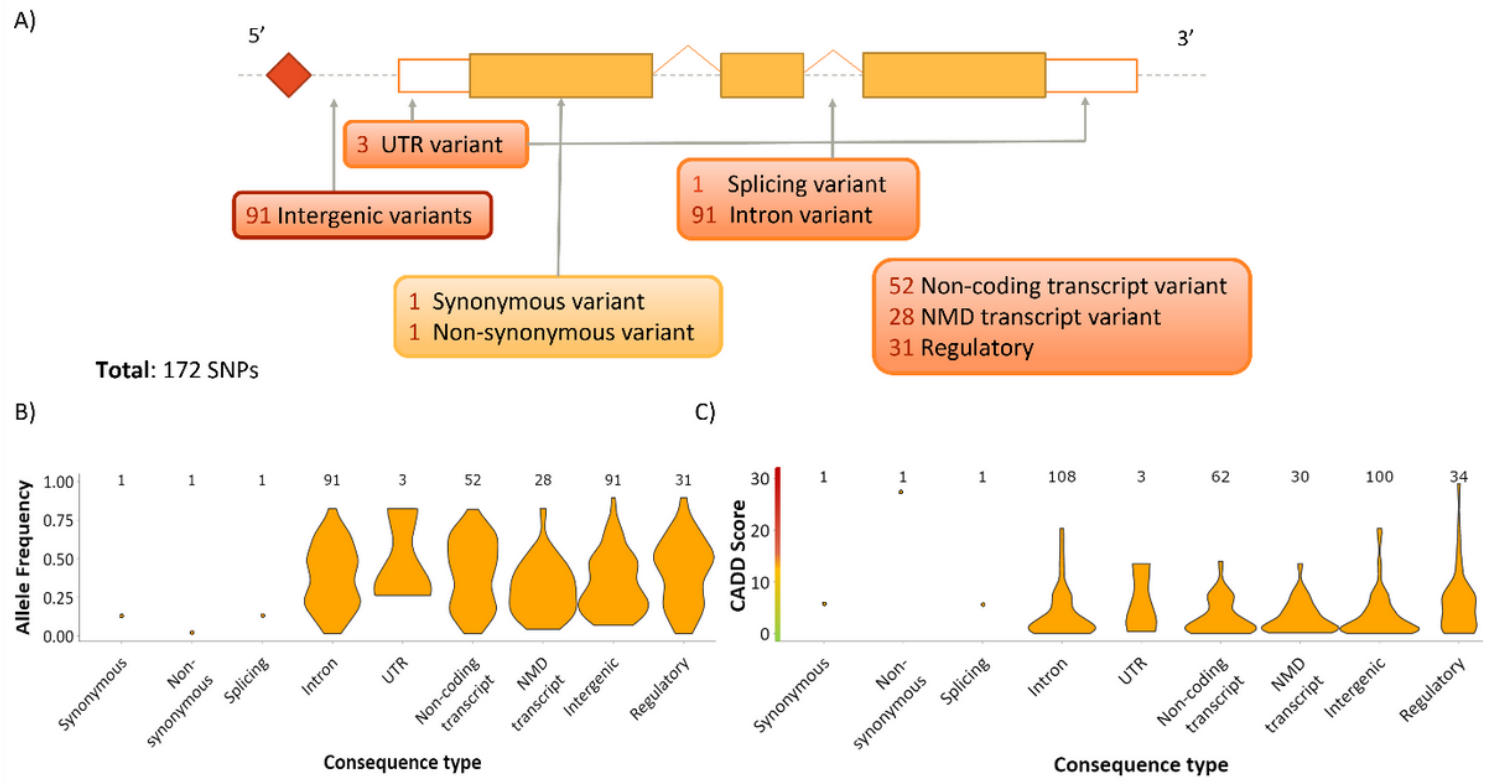

Figure 1

MD GVs are mostly non-coding, common and potentially not pathogenic. A) GVs distributed along the genome according to its 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. NMD: Nonsense Mediated Decay.

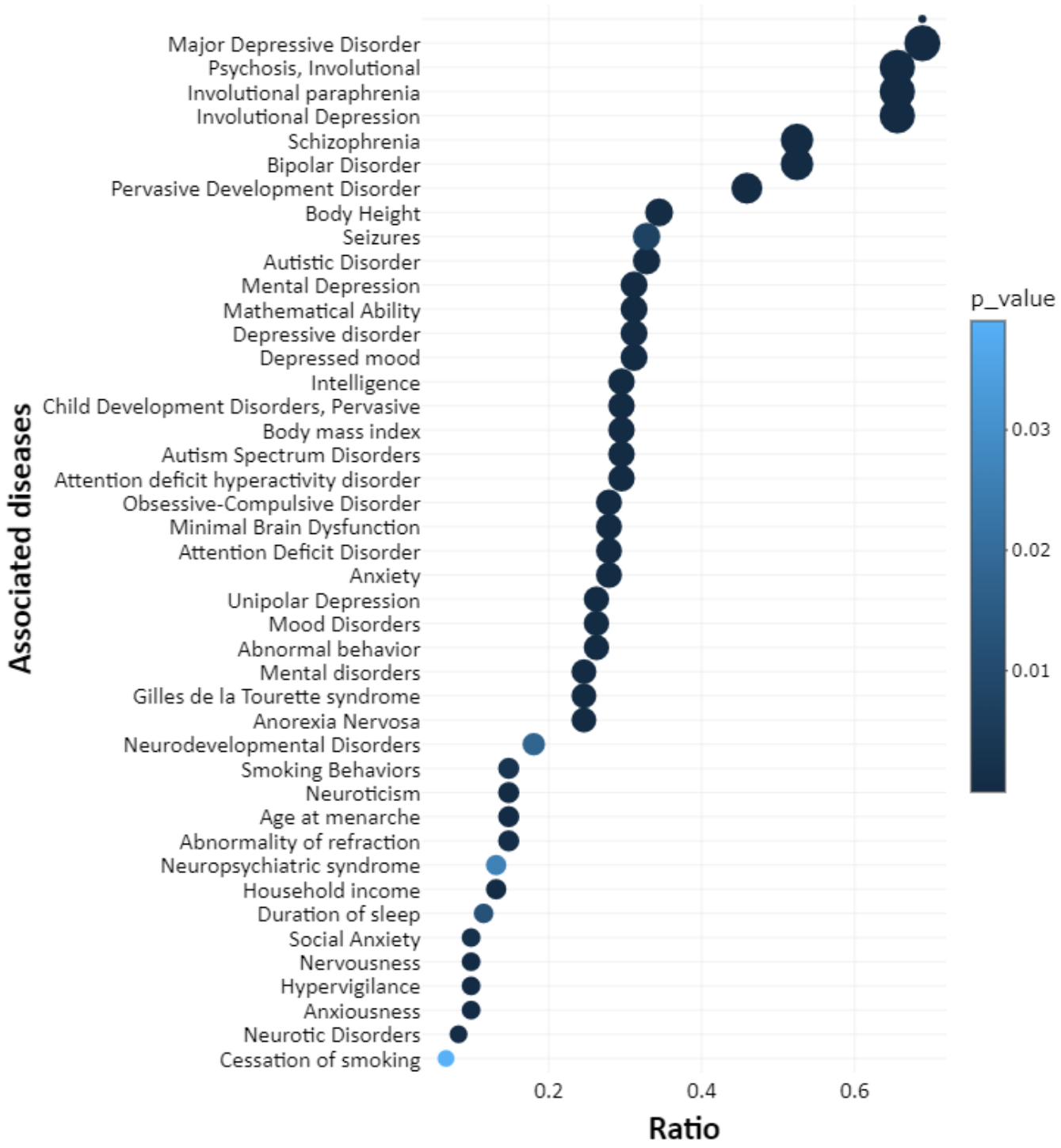

Figure 2

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.

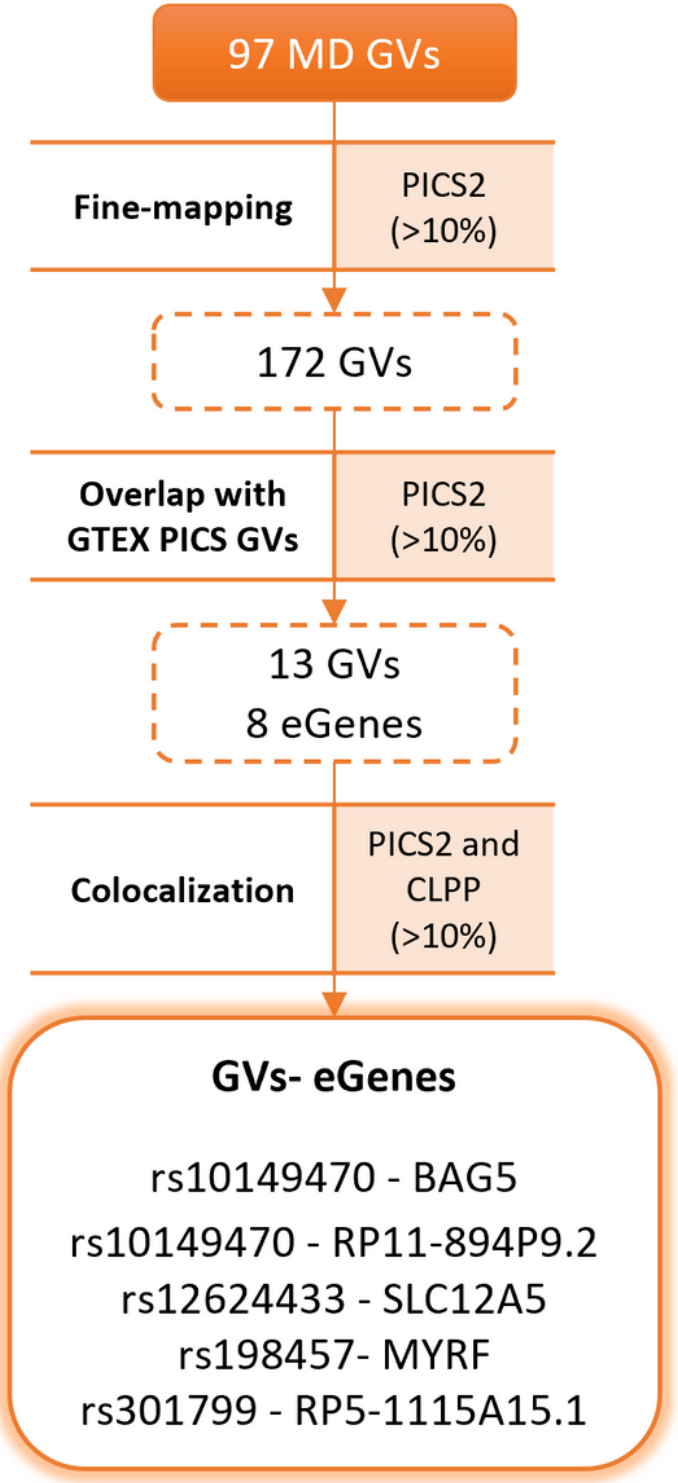

Figure 3

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.

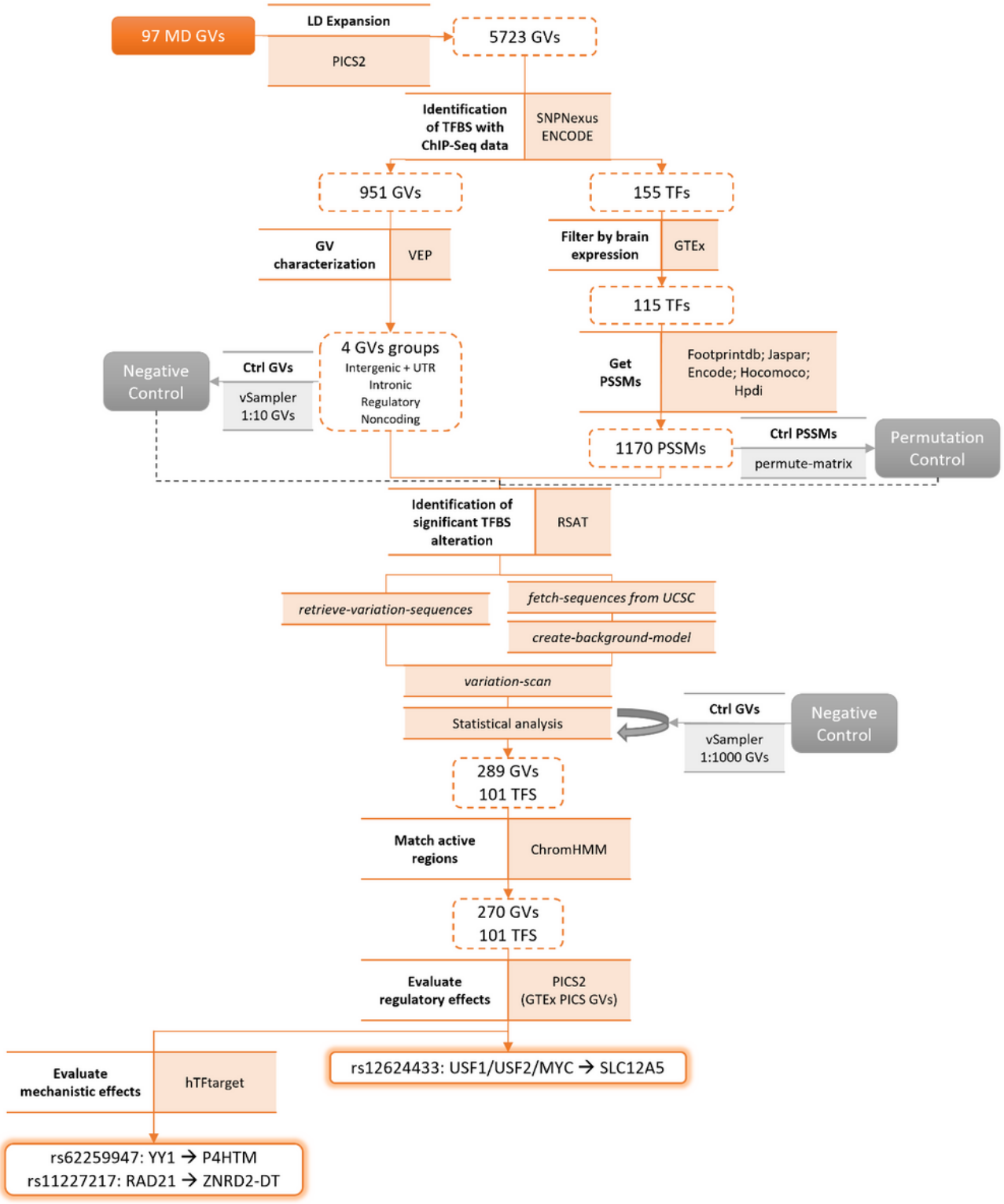

Figure 4

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.

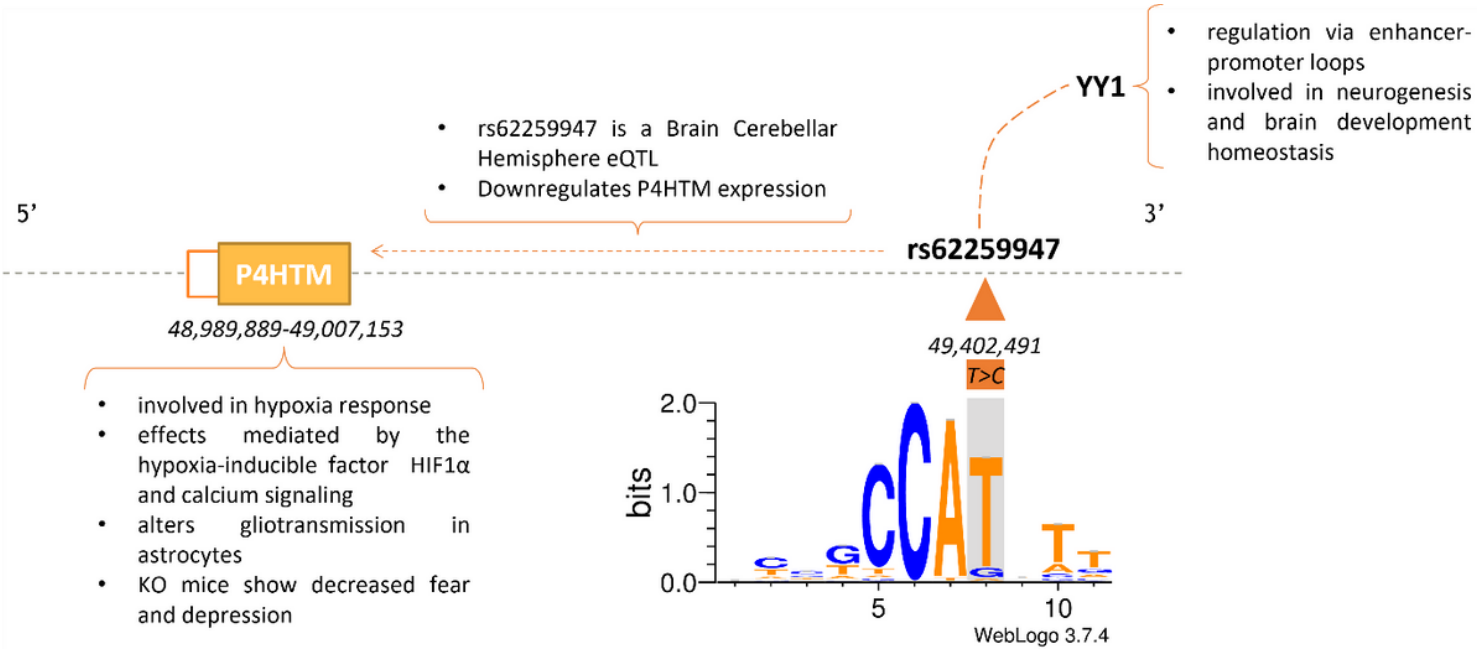

Figure 5

The GV rs62259947 might disrupt the binding of YY1, thus affecting the expression of P4HTM and resulting in behavioural 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 mediated by HIF1α regulates calcium signalling and is also associated with behaviour.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- PerezGranadoSupplementaryfile.docx
- PerezGranadoSupplementarytables16.xlsx