

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

Curation guidelines for GVs associated with MD

1. Characterisation of MD context

Characterise the context of GV-disease association distinguishing between (Table 2 and Supplementary Table S3):

- **MD:** major depression.
- **Features/Traits:** the study of features and traits associated with MD.
- **Environmental factors:** the consideration of environmental factors in the association study design.
- **Comorbidities:** the study of MD comorbidities
- **No MD-related:** wrongly detected associations where MD is not the topic of study.

Associations in the context of MD, whether or not they consider environmental factors, will resume the curation pipeline.

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 <u>G1287A of the NET Gene Polymorphisms and Brain Volume in Major Depressive Disorder</u> : A Voxel-Based MRI Study. (PMID: 26960194)
Environmental factors (E)	The present study suggests that the combined effect of <u>rs2242446 and rs5569 in the NET gene could modify the response to the negative life events in triggering MD</u> . (PMID: 18779921)
Comorbidities	Association analysis of the <u>5-HT6 receptor polymorphism C267T with depression in patients with Alzheimer's disease</u> . (PMID: 11442897)
No MD-related	From among this cohort, we studied the <u>chloride currents</u> generated by G190S (associated with pronounced <u>transitory depression</u>), F167L (little or no <u>transitory depression</u>), and A531V (variable <u>transitory depression</u>) hClC-1 mutants in transfected HEK293 <u>cells</u> using <u>patch-clamp</u> . (PMID: 23933576)

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

2. Evaluation of publication

Evaluate the publication abstract to detect and remove (Table 3):

- **Reviews:** review-only studies.
- **Incorrect variants:** variants incorrectly captured or harmonised.

We remove the identified associations.

Reviews	<u>In this review</u> , we bridge evidence from neuroimaging, behavioural and clinical studies that have examined the <u>role of COMT variants</u> 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 <u>Ala348Thr (rs1718119)</u> in the depressed and diabetic patient groups. <u>[identified as: rs755302767]</u> (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

Classify the GV-disease association type (Table 4 and Supplementary Table S3):

- **VDA:** strict GV-disease association with no E or TR influence.
- **Environmental:** studies considering environmental factors.
- **Treatment response (TR):** studies considering the response to treatment.

All GVs resume the analysis.

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

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

4. Characterisation of the study type

Characterise the type of association study (Table 5):

- **Preclinical studies:** studies conducted on cell lines/cultures and animal models.
- **Candidate gene studies or CGS:** studies focused on a small set of genes or GVs and generally conducted on a small sample size.

- **GWAS:** studies that evaluate million of GVs in big cohorts.

All GVs resume the analysis.

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 <u>cellular assays</u> ... inversely correlated to the incidence of major depression...(PMID: 18184810)
Candidate gene studies (CGS)	<u>Three SNPs (rs10008257, rs2433320 and rs2452600)</u> were identified in the PDLIM5 gene and <u>genotyped in patients diagnosed with recurrent MDD</u> and in matched control subjects. (PMID: 18197271)
Genome-Wide Association Studies (GWAS)	<u>Genome-wide association study of depression</u> 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

Pass quality controls filters on CGS and GWAS studies:

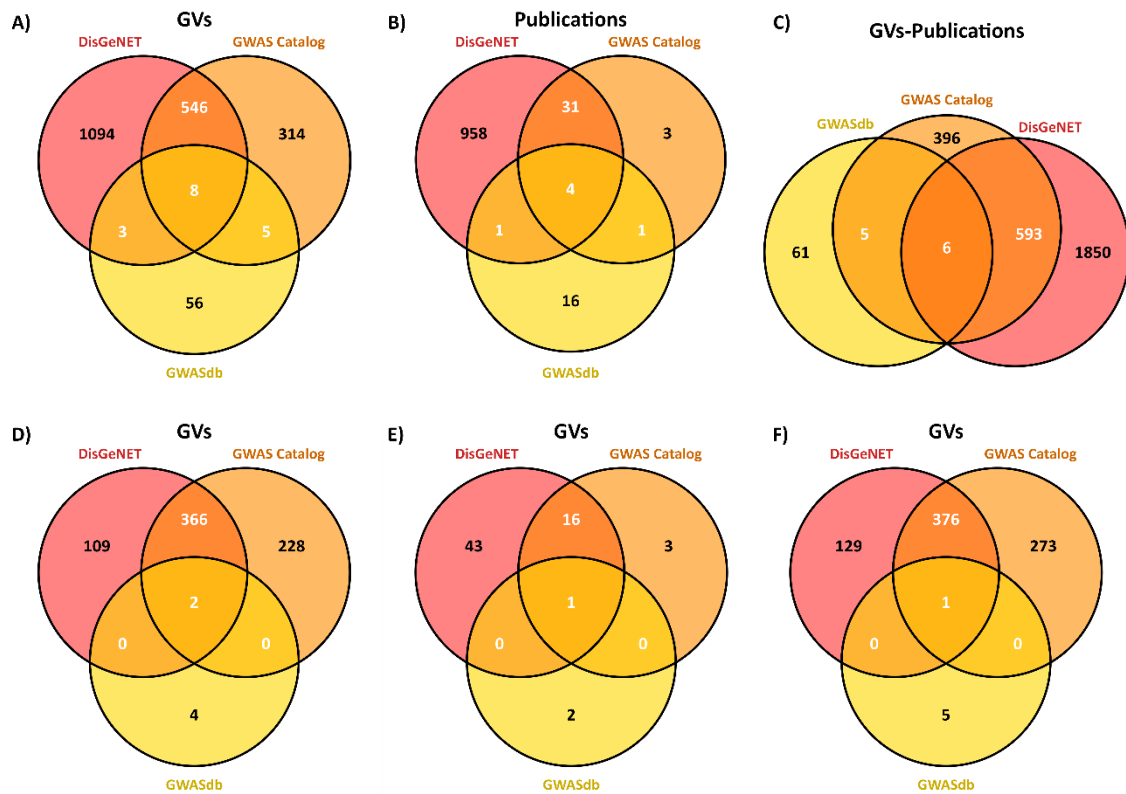
- **Minimal sample size filtering:** sample size cut-off according to the study type and number of GVs under evaluation (Table 6).
- **Significance level filtering:** minimum significance threshold for CGS (0.05) and GWAS (5×10^{-8}).

Studies passing the filters would result in the final curated dataset.

Analysis	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.

Supplementary Figures



Supplementary Figure S1. Sources of original and curated genetic associations. Overlap between the different resources (i.e. DisGeNET, GWAS Catalog or GWASdb) reporting the genetic associations originally extracted (A, B, C) and after curation (C, D, E) divided between A and D) GVs; B and E) Publications; and C and F) GVs-Publications pairs. GV: genetic variant.