# Unique genetic and risk-factor profiles in multimorbidity clusters of depression-related disease trajectories from a study of 1.2 million subjects

Andras Gezsi[1 *], Sandra Van der Auwera[2,3 *], Hannu Mäkinen[4], Nora Eszlari[5,6], Gabor Hullam[1,5], Tamas Nagy[1,5,6], Sarah Bonk[2], Rubèn González-Colom[7], Xenia Gonda[5,6,8], Linda Garvert[2], Teemu Paajanen[4], Zsofia Gal[5,6], Kevin Kirchner[2], Andras Millinghofer[9], Carsten O. Schmidt[10], Bence Bolgar[1], Josep Roca[7], Isaac Cano[7], Mikko Kuokkanen[4,11,12], Peter Antal[1 #], Gabriella Juhasz[5,6 # +]

[*] These authors contributed equally to the manuscript

[#] These authors jointly supervised the research

[+] Corresponding author: Gabriella Juhasz

[1] Department of Measurement and Information Systems, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111 Budapest, Hungary

[2] Department of Psychiatry and Psychotherapy, University Medicine Greifswald, 17475 Greifswald, Germany.

[3] German Centre for Neurodegenerative Diseases (DZNE), Site Rostock/Greifswald, 17475 Greifswald, Germany.

[4] Department of Public Health and Welfare, Finnish Health and Welfare Institute. Biomedicum 1, Haartmaninkatu 8, 00290 Helsinki.

[5] Department of Pharmacodynamics, Faculty of Pharmacy, Semmelweis University, Nagyvárad tér 4., H-1089 Budapest, Hungary.

[6] NAP3.0-SE Neuropsychopharmacology Research Group, Hungarian Brain Research Program, Semmelweis University, Üllői út 26., H-1085 Budapest, Hungary.

[7] Hospital Clínic de Barcelona. Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS). Universitat de Barcelona. Villarroel 170. Barcelona, 08036. Spain.

[8] Department of Psychiatry and Psychotherapy, Semmelweis University, Gyulai Pál utca 2., H-1085 Budapest, Hungary

[9] Abiomics Europe Ltd., Zólyomi út. 23., H-1118 Budapest, Hungary

[10] Institute for Community Medicine, University Medicine Greifswald, 17475 Greifswald, Germany.

[11] Department of Human Genetics and South Texas Diabetes and Obesity Institute, School of Medicine at University of Texas Rio Grande Valley, Brownsville, TX, United States.

[12] Research Program for Clinical and Molecular Metabolism, Faculty of Medicine, University of Helsinki, Finland

# *Supplementary Material (ethics statement and figures)*

## Content:


- **Ethics statements of participating cohorts** *(p. 3)*

- **Supplementary Figures for results section** *Dynamic Bayesian network analysis reveals seven MDD-related multimorbidity clusters (p. 6)*
  - Figure S1.1. The temporal trajectories of the MDD-associated multimorbidity burden in the discovery cohorts (UKB, THL, CHSS).
  - Figure S1.2. The average onset ages of the consensual diseases according to the seven clusters in the UKB.
  - Figure S2.1-S2.4. Scatter plot of cluster membership variables' coefficients on the onset of all consensual diseases according to UKB versus the other cohorts.
  - Figure S3. Density plot of the age pattern of each cluster using hard cluster assignment for each cohort.
  - Figure S4. Kendall's correlation matrix for MDD-related cluster membership probabilities for each cohort.
  - Figures S5.1-S5.7. Distribution of MDD-related cluster membership for Cluster 7 in UKB data.
  - Figure S6. Count distribution of MDD-related cluster membership for Clusters 1-7 in the individual cohorts.

- **Supplementary Figures for results section** *GWAS analysis of MDD-related multimorbidity clusters in the UKB cohort identifies immune system-related genetic profiles (p. 19)*
  - Figure S7.1-S7.7. GWAS results for MDD-related cluster membership in UKB data for Clusters 1-7.
  - Figure S8. Functional overrepresentation analysis of UKB and FinnGen GWAS.
  - Figure S9. Genetic correlation between clusters and the 86 consensual diseases on which the definitions of the clusters are based.
  - Figure S10. Comparison of the value of coefficient in Cox regression of cluster membership on disease onset versus the genetic correlation between the cluster membership and the disease occurrence in the UKB data.
  - Figure S11.1-S11.7. Functional modules of the human interactome in Cluster 1-7 that are pleiotropic with MDD.

- **Supplementary Figures for results section** *Modifiable risk-factor profiles of MDD-related multimorbidity clusters in the UKB cohort (p. 37)*
    - Figure S12. Complex linear regression models involving available modifiable factors included in the UK Biobank dataset for each cluster.

- **Supplementary Figures for results section** *Validation of MDD-related multimorbidity profiles at the genetic and modifiable risk-factor levels (p. 38)*
    - Figure S13. Pearson correlation pattern of GWAS derived beta values between MDD-related clusters.
    - Figure S14. Results for polygenic risk score (PRS) analysis in THL.
    - Figure S15.1-S15.7. GWAS results for MDD-related cluster membership in FinnGen data.
    - Figure S16. Scatter plot of gene-level aggregated *p*-values in UKB versus FinnGen.
    - Figure S17. Complex linear regression models involving available modifiable factors included in the THL data for each cluster.
    - Figure S18. Four measures that show the performance of the clustering method using only a fraction of all consensual disease variables from UKB.
    - Figure S19. Pearson correlation matrix for polygenic risk scores (PRS) in SHIP.
    - Figure S20. Simple linear regression models involving modifiable risk factors included in the SHIP dataset for each cluster.

## 1. Ethics statements:

**a. UK Biobank Ethics statement:** Under the application number *1602* we used data from the UK Biobank (UKB) database resource which includes medical and phenotypic data of recruited participants on the NHS patient registers of people aged 40–69 years (DOI: 10.1371/journal.pone.0075362). Ethical approval was given by the National Research Ethics Service Committee North West–Haydock (DOI: 10.1038/s41467-018-03242-8), and all participants gave

written informed consent. All procedures were in accordance with the Declaration of Helsinki.

b. **Finngen/THL DF 10 Ethics statement:** Patients and control subjects in FinnGen provided informed consent for biobank research, based on the Finnish Biobank Act. Alternatively, separate research cohorts, collected prior the Finnish Biobank Act came into effect (in September 2013) and start of FinnGen (August 2017), were collected based on study-specific consents and later transferred to the Finnish biobanks after approval by Fimea (Finnish Medicines Agency), the National Supervisory Authority for Welfare and Health. Recruitment protocols followed the biobank protocols approved by Fimea. The Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS) statement number for the FinnGen study is Nr HUS/990/2017.

The FinnGen study is approved by Finnish Institute for Health and Welfare (permit numbers: THL/2031/6.02.00/2017, THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018, THL/283/6.02.00/2019, THL/1721/5.05.00/2019 and THL/1524/5.05.00/2020), Digital and population data service agency (permit numbers: VRK43431/2017-3, VRK/6909/2018-3, VRK/4415/2019-3), the Social Insurance Institution (permit numbers: KELA58/522/2017, KELA 131/522/2018, KELA 70/522/2019, KELA 98/522/2019, KELA 134/522/2019, KELA 138/522/2019, KELA 2/522/2020, KELA 16/522/2020), Findata permit numbers THL/2364/14.02/2020, THL/4055/14.06.00/2020, THL/3433/14.06.00/2020, THL/4432/14.06/2020, THL/5189/14.06/2020, THL/5894/14.06.00/2020, THL/6619/14.06.00/2020, THL/209/14.06.00/2021, THL/688/14.06.00/2021, THL/1284/14.06.00/2021, THL/1965/14.06.00/2021, THL/5546/14.02.00/2020, THL/2658/14.06.00/2021, THL/4235/14.06.00/2021, Statistics Finland (permit numbers: TK-53-1041-17 and TK/143/07.03.00/2020 (earlier TK-53-90-20) TK/1735/07.03.00/2021, TK/3112/07.03.00/2021) and Finnish Registry for Kidney Diseases permission/extract from the meeting minutes on 4$^{th}$July 2019.

The Biobank Access Decisions for FinnGen samples and data utilized in FinnGen Data Freeze 10 include: THL Biobank BB2017_55, BB2017_111,

BB2018_19, BB_2018_34, BB_2018_67, BB2018_71, BB2019_7, BB2019_8, BB2019_26, BB2020_1, BB2021_65, Finnish Red Cross Blood Service Biobank 7.12.2017, Helsinki Biobank HUS/359/2017, HUS/248/2020, HUS/150/2022 § 12, §13, §14, §15, §16, §17, §18, and §23, Auria Biobank AB17-5154 and amendment #1 (August 17 2020) and amendments BB_2021-0140, BB_2021-0156 (August 26 2021, Feb 2 2022), BB_2021-0169, BB_2021-0179, BB_2021-0161, AB20-5926 and amendment #1 (April 23 2020)and it´s modification (Sep 22 2021), Biobank Borealis of Northern Finland_2017_1013, 2021_5010, 2021_5018, 2021_5015, 2021_5023, 2021_5017, 2022_6001, Biobank of Eastern Finland 1186/2018 and amendment 22 § /2020, 53§/2021, 13§/2022, 14§/2022, 15§/2022, Finnish Clinical Biobank Tampere MH0004 and amendments (21.02.2020 & 06.10.2020), §8/2021, §9/2022, §10/2022, §12/2022, §20/2022, §21/2022, §22/2022, §23/2022, Central Finland Biobank 1-2017, and Terveystalo Biobank STB 2018001 and amendment 25th Aug 2020, Finnish Hematological Registry and Clinical Biobank decision 18th June 2021, Arctic biobank P0844: ARC_2021_1001.

c. **CHSS Ethics statement:** The Ethics Committee for Human Research at Hospital Clinic de Barcelona approved the study protocol on the 24th of march of 2021 (HCB/2020/1051) in the context of the EU project: ERAPERMED2019-108 - TRAJECTOME. All the data were handled in compliance with the General Data Protection Regulation 2016/679 on data protection and privacy for all individuals within the European Union. The study was conducted in conformity with the Helsinki Declaration (Stronghold Version, Brazil, October 2013) and in accordance with the protocol and the relevant legal requirements (Biomedical Research Act 14/2007 of 3 July).

d. **SHIP Ethics statement:** The study followed the recommendations of the Declaration of Helsinki. The medical ethics committee of the University of Greifswald approved the study protocol, and oral and written informed consents were obtained from each of the study participants.

## 2. Supplementary Figures for results section *Dynamic Bayesian network analysis reveals seven MDD-related multimorbidity clusters*

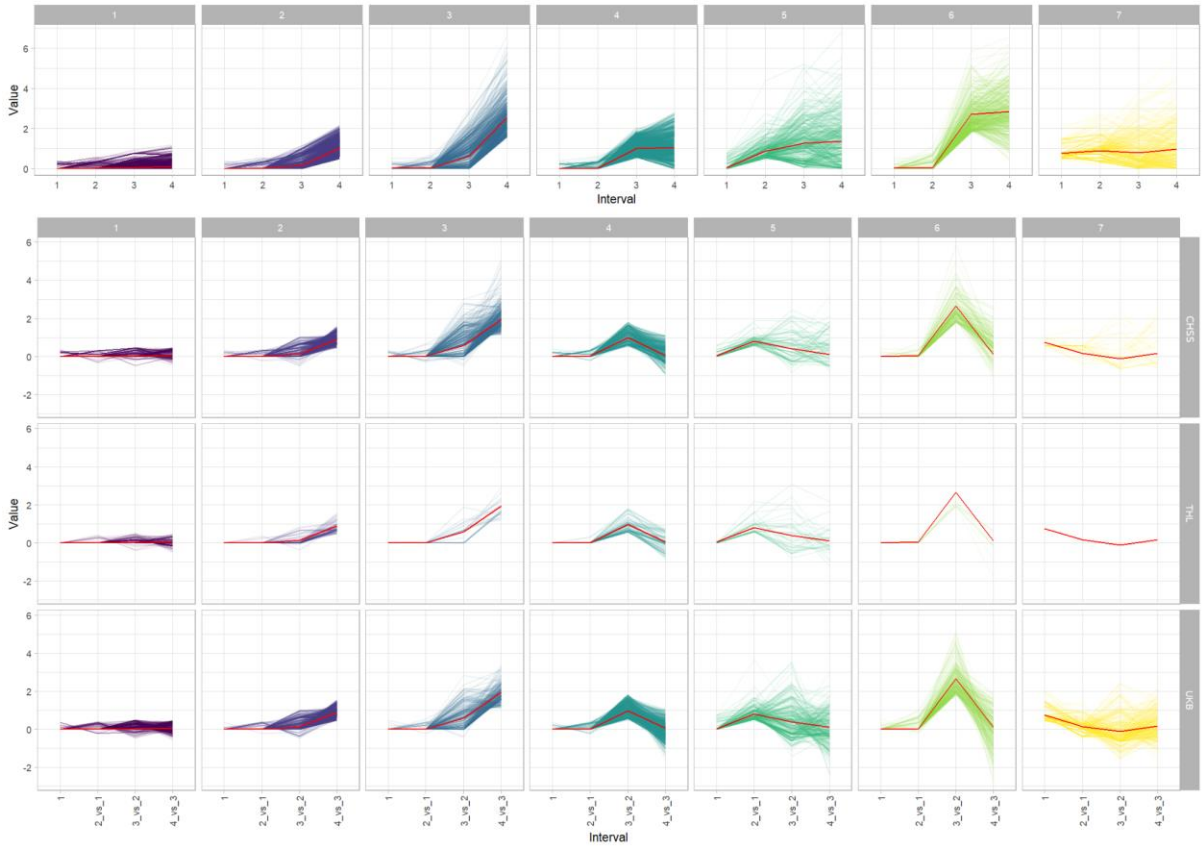



**Figure S1.1. The temporal trajectories of the MDD-associated multimorbidity burden in the discovery cohorts (UKB, THL, CHSS).** Boxes correspond to clusters. Colored lines denote the trajectories of randomly sampled individuals in each cluster, and the red lines denote the mean trajectories. Top: The trajectories according to all discovery cohorts as a function of the cumulative time intervals. Bottom: The trajectories according to each discovery cohort as a function of the first time interval and the difference between the subsequent consecutive time intervals. Cumulative time intervals 1: 0-20 years, 2: 0-40 years, 3: 0-60 years, 4: 0-70 years.

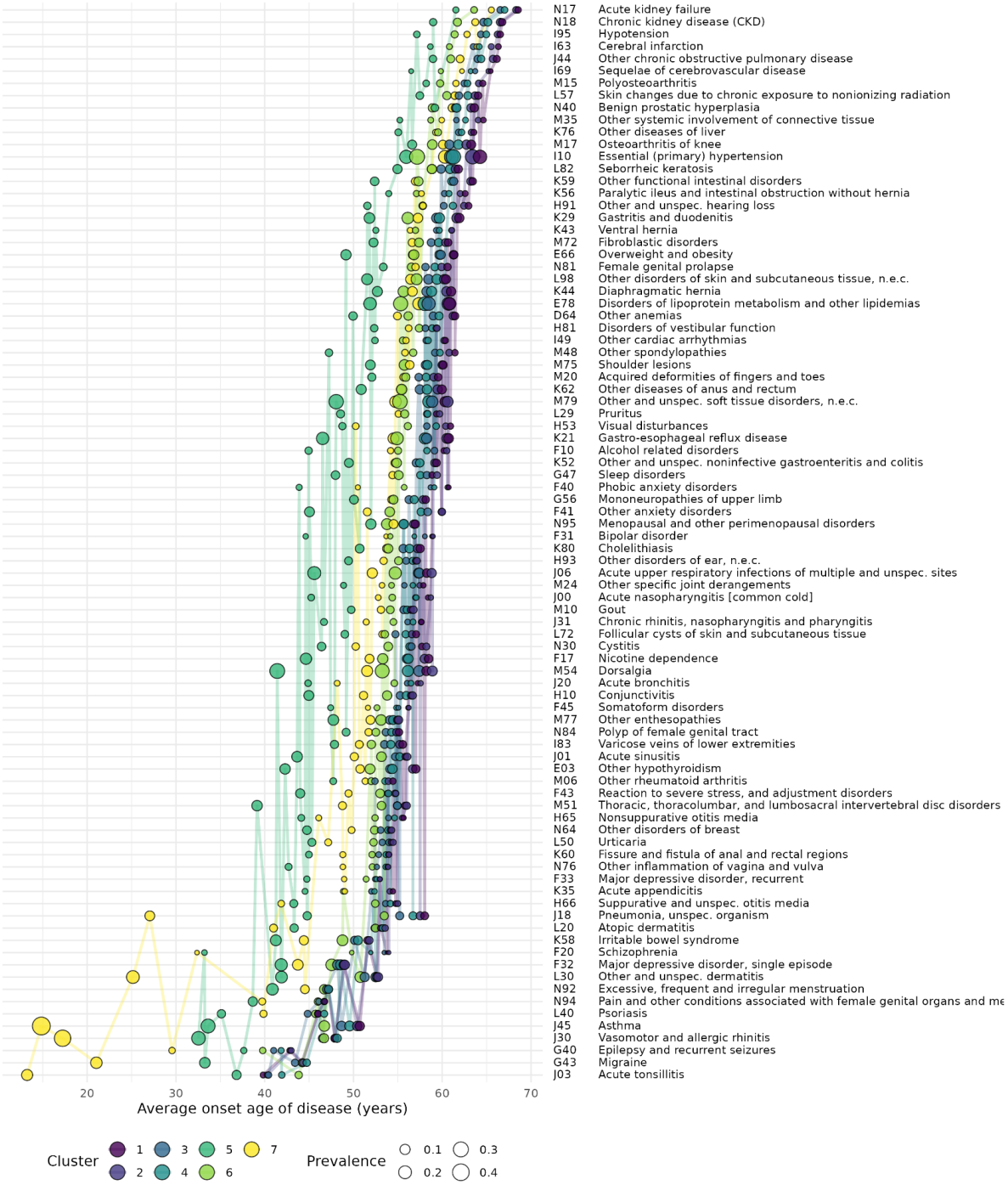



**Figure S1.2. The average onset ages of the consensual diseases according to the seven clusters in the UKB.** Average onsets are weighted by the cluster membership probabilities of each participant. The nodes' color indicates the cluster, and the node size is proportional to the observed prevalence of the disease in the cluster.

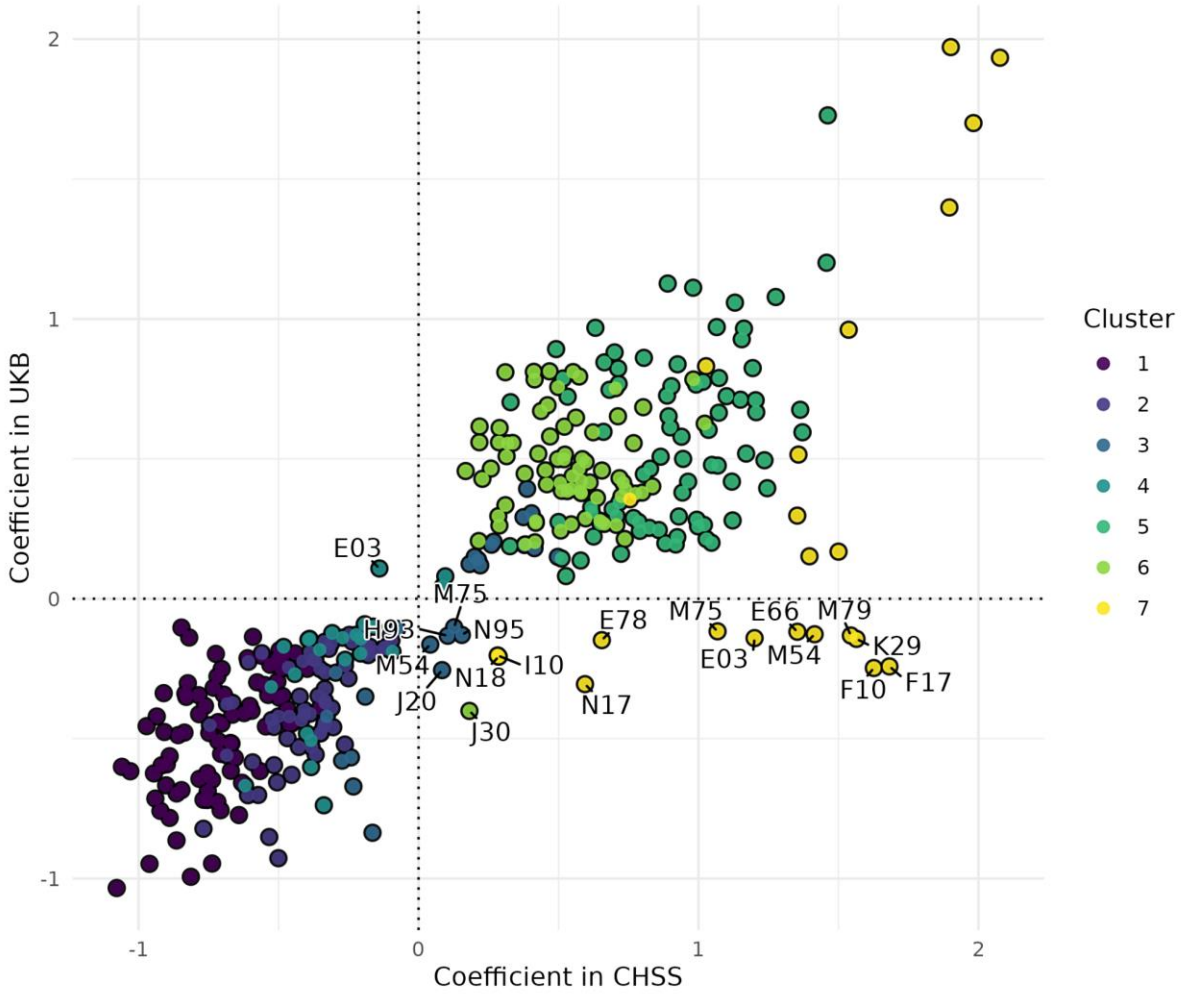



**Figure S2.1. Scatter plot of cluster membership variables' coefficients on the onset of all consensual diseases according to CHSS versus UKB**. Each point represents a disease according to a given cluster (indicated by the point's color). Only those diseases are shown which are statistically significant in both cohorts. Diseases for which the sign of the coefficients differ are indicated with labels.

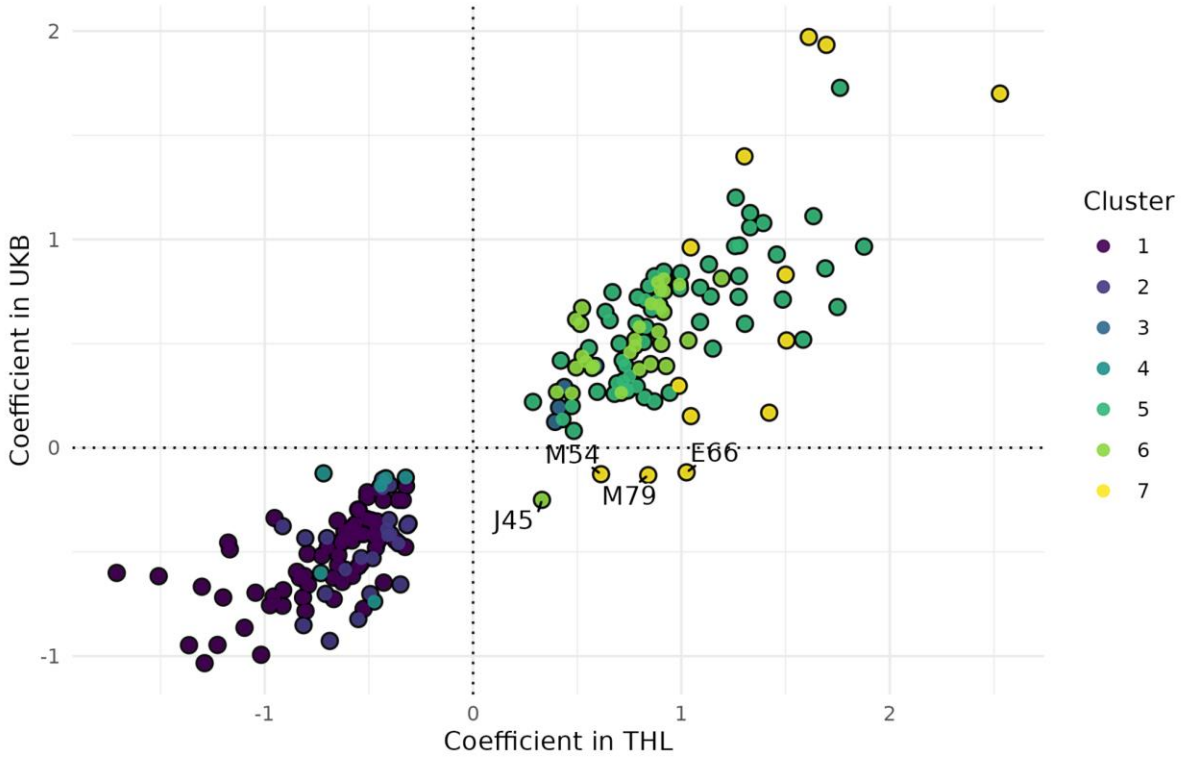



**Figure S2.2. Scatter plot of cluster membership variables' coefficients on the onset of all consensual diseases according to THL versus UKB.** Each point represents a disease according to a given cluster (indicated by the point's color). Only those diseases are shown which are statistically significant in both cohorts. Diseases for which the sign of the coefficients differ are indicated with labels.

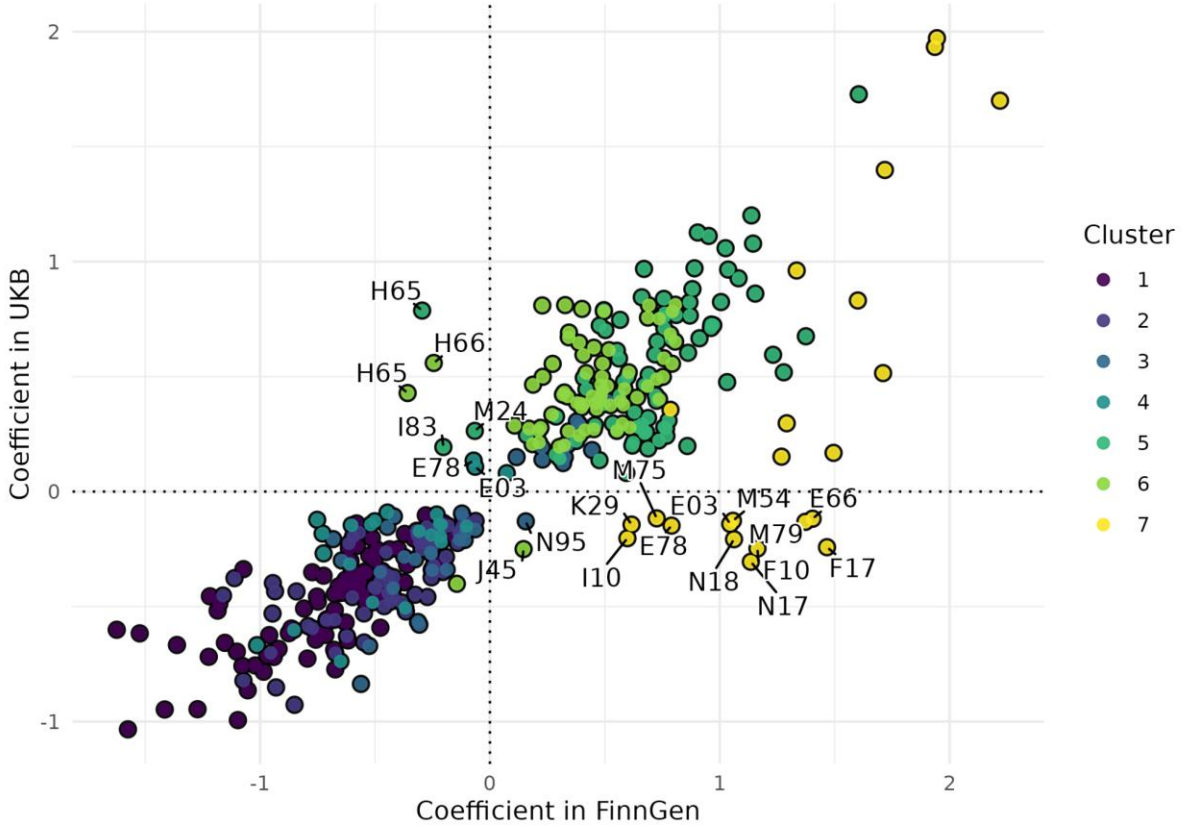



**Figure S2.3. Scatter plot of cluster membership variables' coefficients on the onset of all consensual diseases according to FinnGen versus UKB.** Each point represents a disease according to a given cluster (indicated by the point's color). Only those diseases are shown which are statistically significant in both cohorts. Diseases for which the sign of the coefficients differ are indicated with labels.

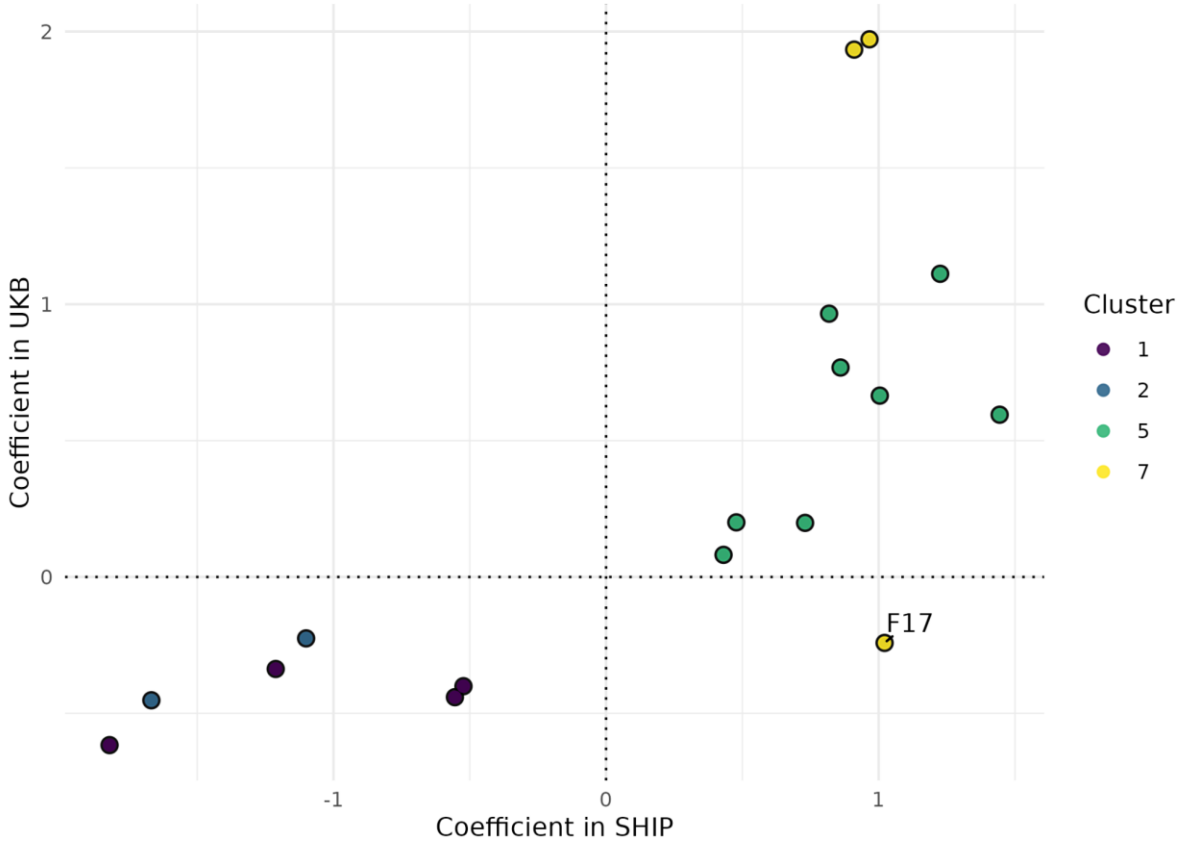



**Figure S2.4. Scatter plot of cluster membership variables' coefficients on the onset of all consensual diseases according to SHIP versus UKB**. Each point represents a disease according to a given cluster (indicated by the point's color). Only those diseases are shown which are statistically significant in both cohorts. The disease for which the sign of the coefficient differs is indicated with a label (F17).

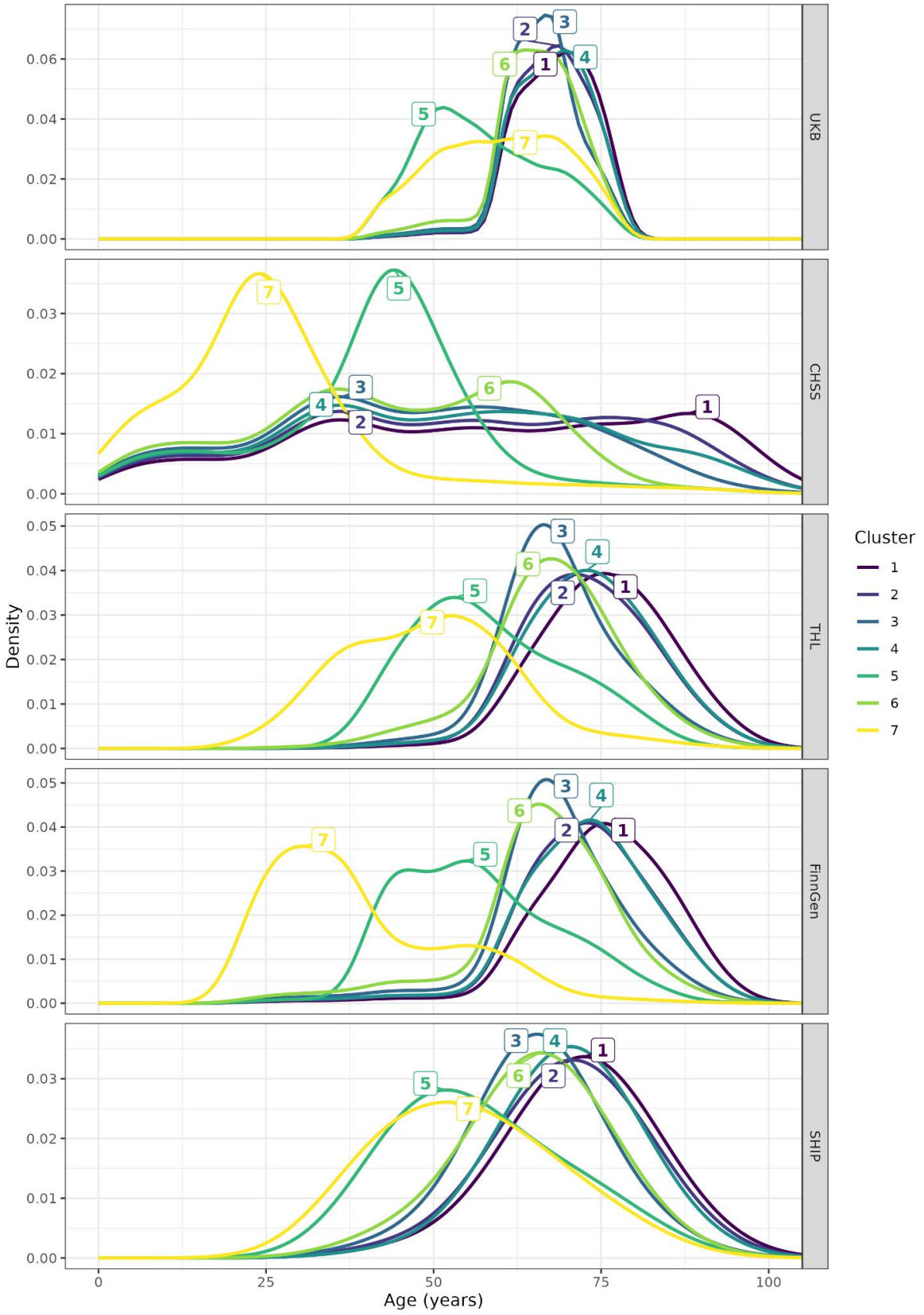



**Figure S3. Density plot of the age pattern of each cluster in each cohort.** Participants are weighted by the posterior probability of belonging to a cluster. Participants were excluded if both being under 60 years and having a maximum posterior probability < 0.25 to any of the clusters. UKB N = 364,008; CHSS N = 645,913; THL N = 23,786; FinnGen N = 277,252; SHIP N = 1126.

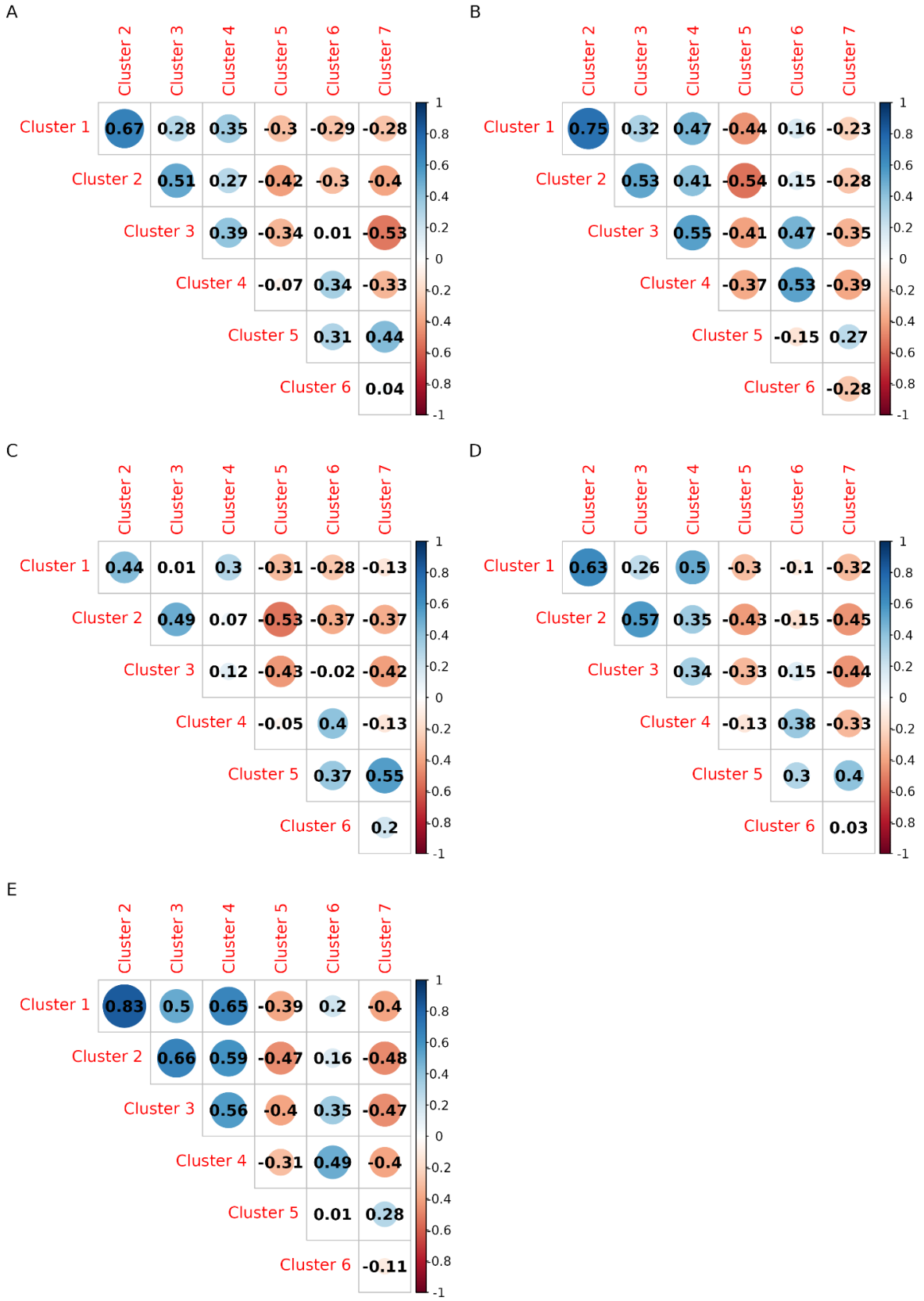



**Figure S4. Kendall's correlation matrix for cluster membership probabilities for each cohort.** Kendall correlation coefficients were computed on filtered samples; participants were excluded if both being under 60 years and having a maximum posterior probability < 0.25 to any of the clusters. A. UKB (N = 364,008); B. CHSS (N = 645,913); C. THL (N = 23,786); D. FinnGen (N = 277,252); E. SHIP (N = 1126).

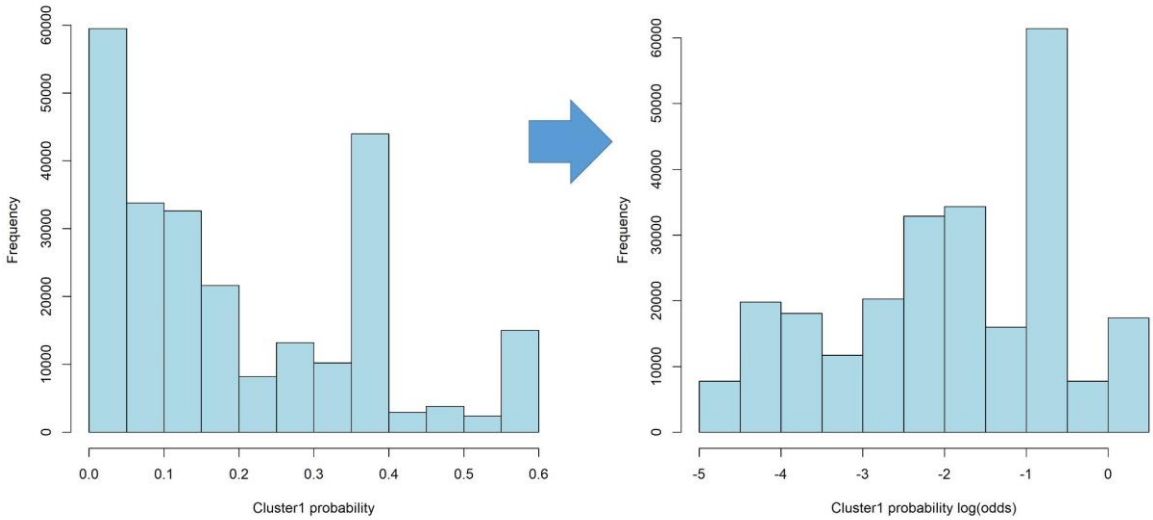



**Figure S5.1. Distribution of MDD-related cluster membership for Cluster 1 in UKB data (N = 247,325).** Raw probability and posterior log-odds.

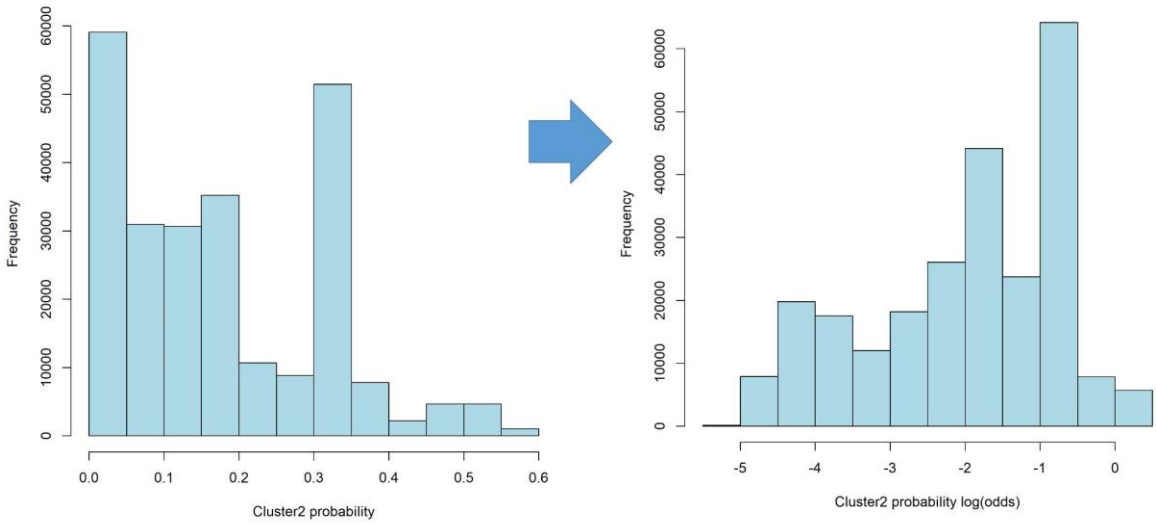



**Figure S5.2. Distribution of MDD-related cluster membership for Cluster 2 in UKB data (N = 247,325).** Raw probability and posterior log-odds.

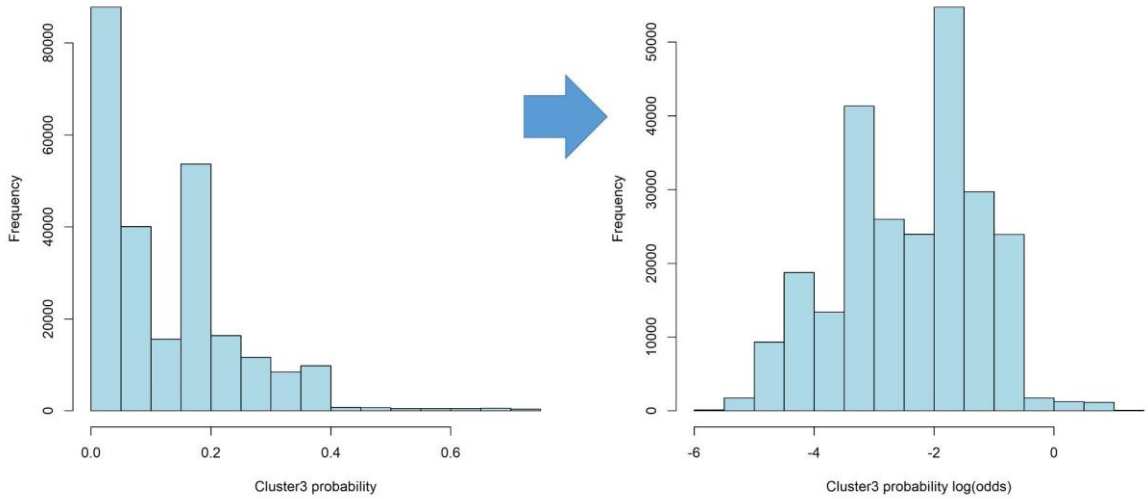



**Figure S5.3. Distribution of MDD-related cluster membership for Cluster 3 in UKB data (N = 247,325).** Raw probability and posterior log-odds.

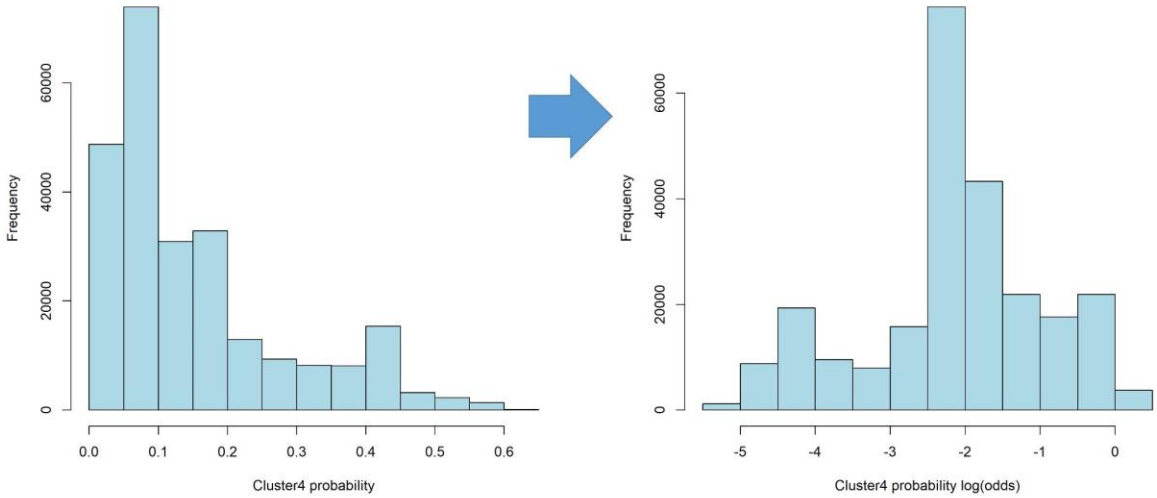



**Figure S5.4. Distribution of MDD-related cluster membership for Cluster 4 in UKB data (N = 247,325).** Raw probability and posterior log-odds.

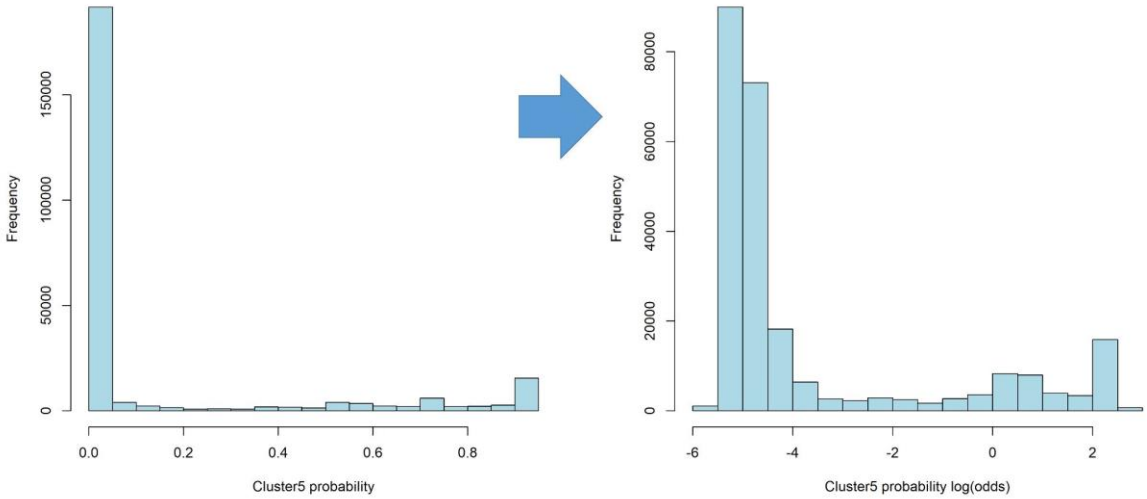



**Figure S5.5. Distribution of MDD-related cluster membership for Cluster 5 in UKB data (N = 247,325).** Raw probability and posterior log-odds.

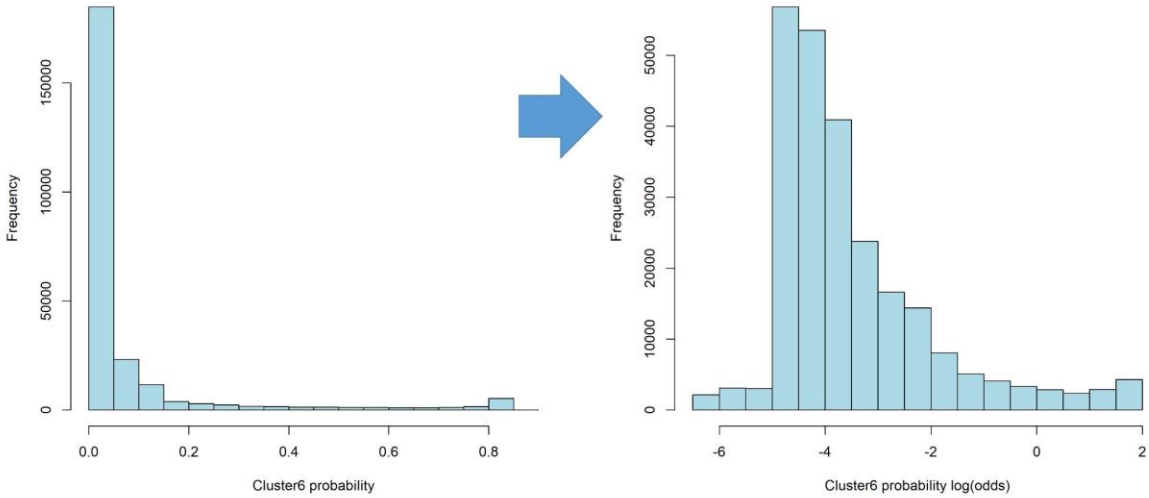



**Figure S5.6. Distribution of MDD-related cluster membership for Cluster 6 in UKB data (N = 247,325).** Raw probability and posterior log-odds.

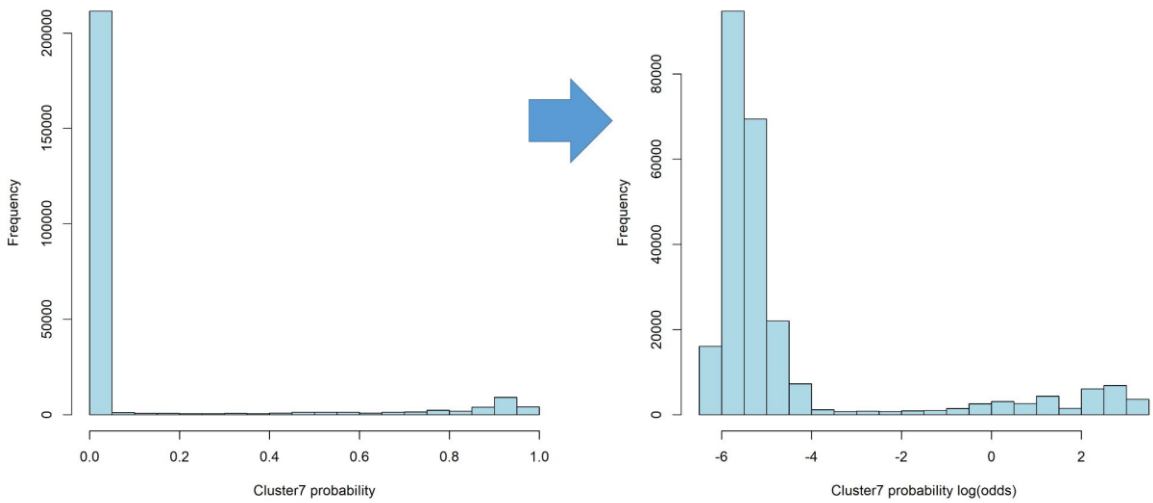



**Figure S5.7. Distribution of MDD-related cluster membership for Cluster 7 in UKB data (N = 247,325).** Raw probability and posterior log-odds.

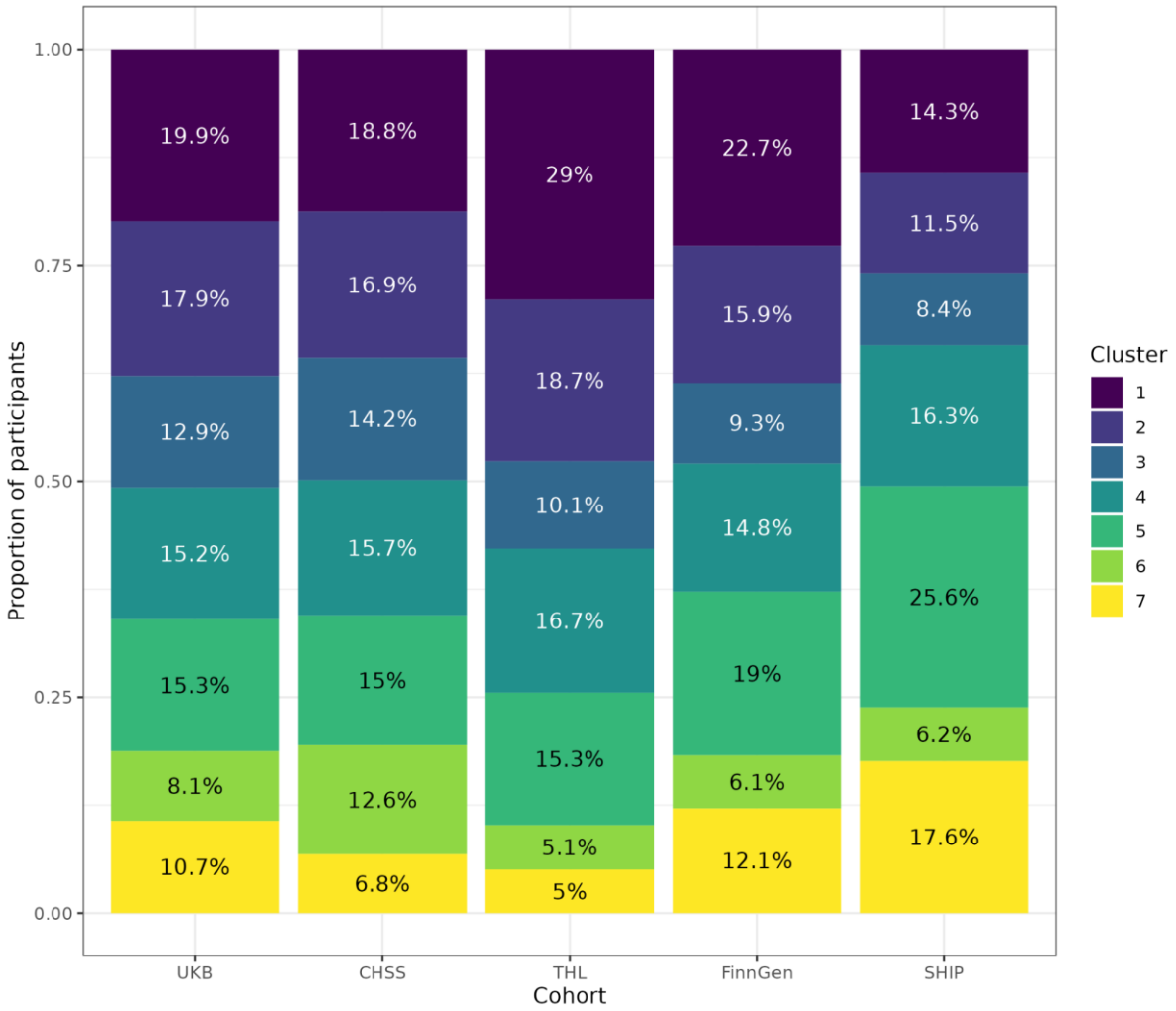



**Figure S6. Count distribution of MDD-related cluster membership for Clusters 1-7 in the individual cohorts.** Colored bars show the expected proportion of participants belonging to each cluster. Each individual is weighted by their cluster membership probability; i.e., to compute the expected number of participants in a cluster, we summed each participant's corresponding cluster membership probability.

## 3. Supplementary Figures for results section *GWAS analysis of MDD-related multimorbidity clusters in the UKB cohort identifies immune system-related genetic profiles*

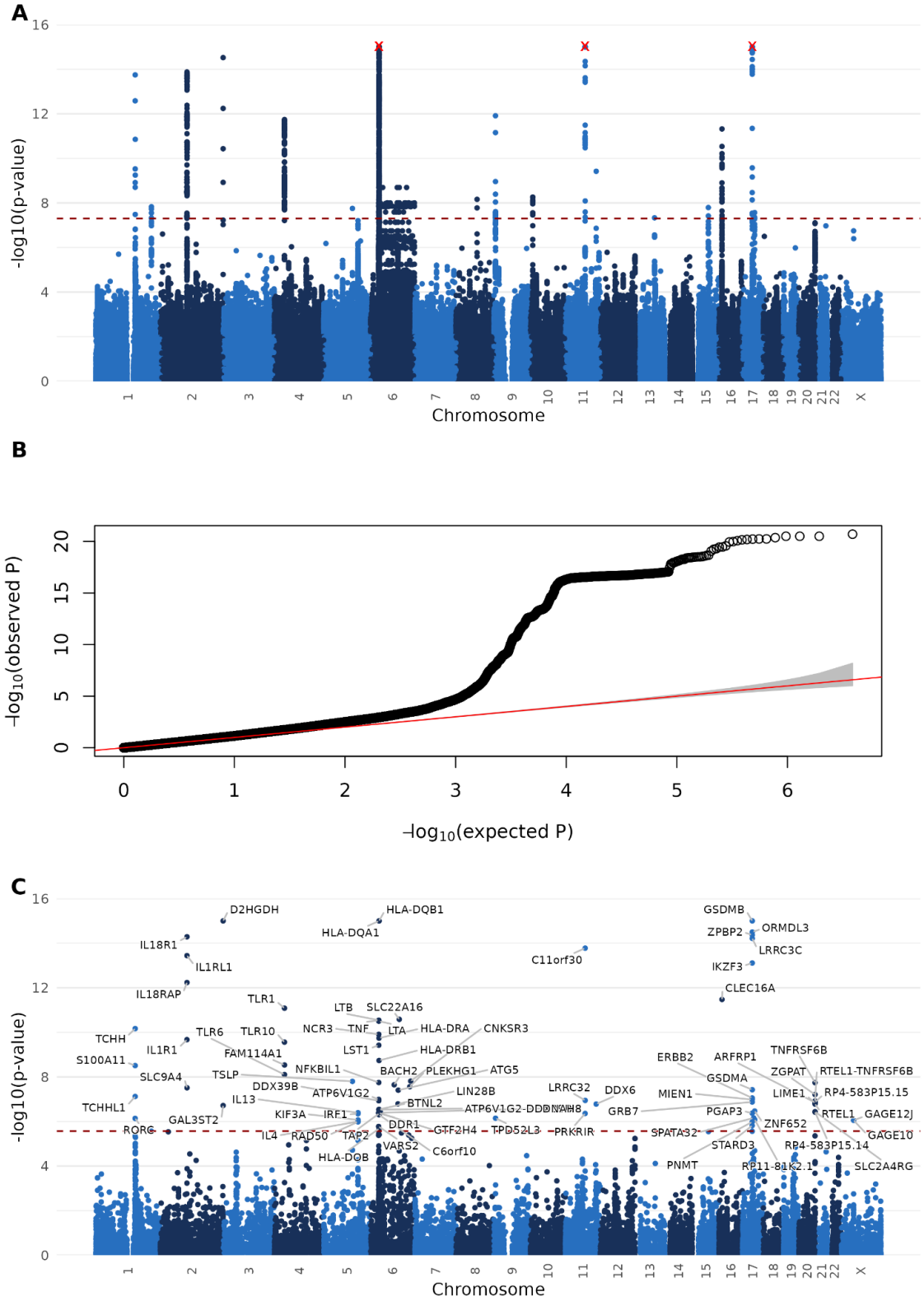



**Figure S7.1. GWAS results for MDD-related Cluster 1 membership in UKB data (N = 247,325).** A. SNP-based genome-wide Manhattan plot. Results capped at $10^{-15}$ are indicated with red crosses. B. QQ-plot. C. Gene-based genome-wide Manhattan plot. The statistically significant genes are indicated with labels.

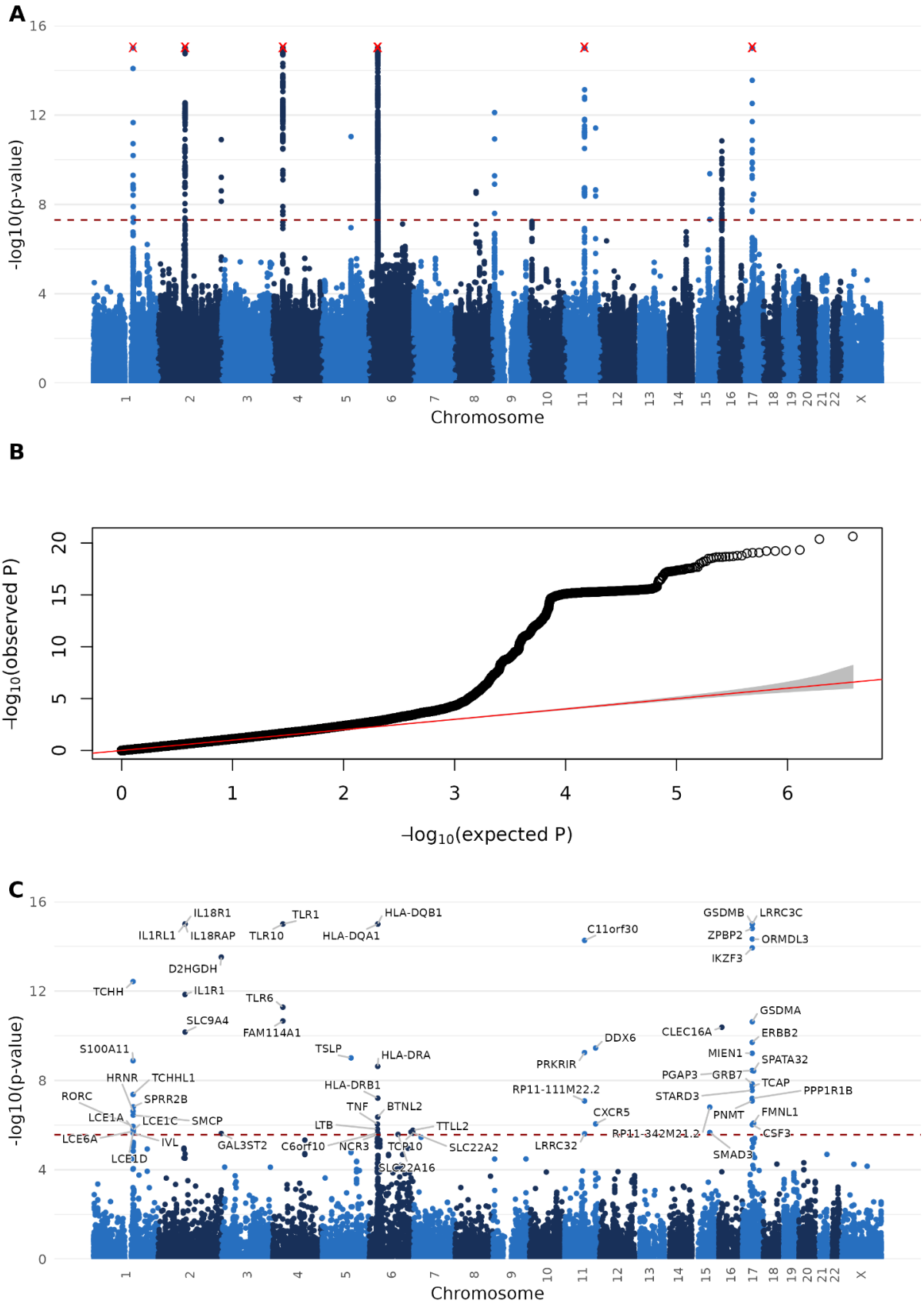



**Figure S7.2. GWAS results for MDD-related Cluster 2 membership in UKB data (N = 247,325).** A. SNP-based genome-wide Manhattan plot. Results capped at $10^{-15}$ are indicated with red crosses. B. QQ-plot. C. Gene-based genome-wide Manhattan plot. The statistically significant genes are indicated with labels.

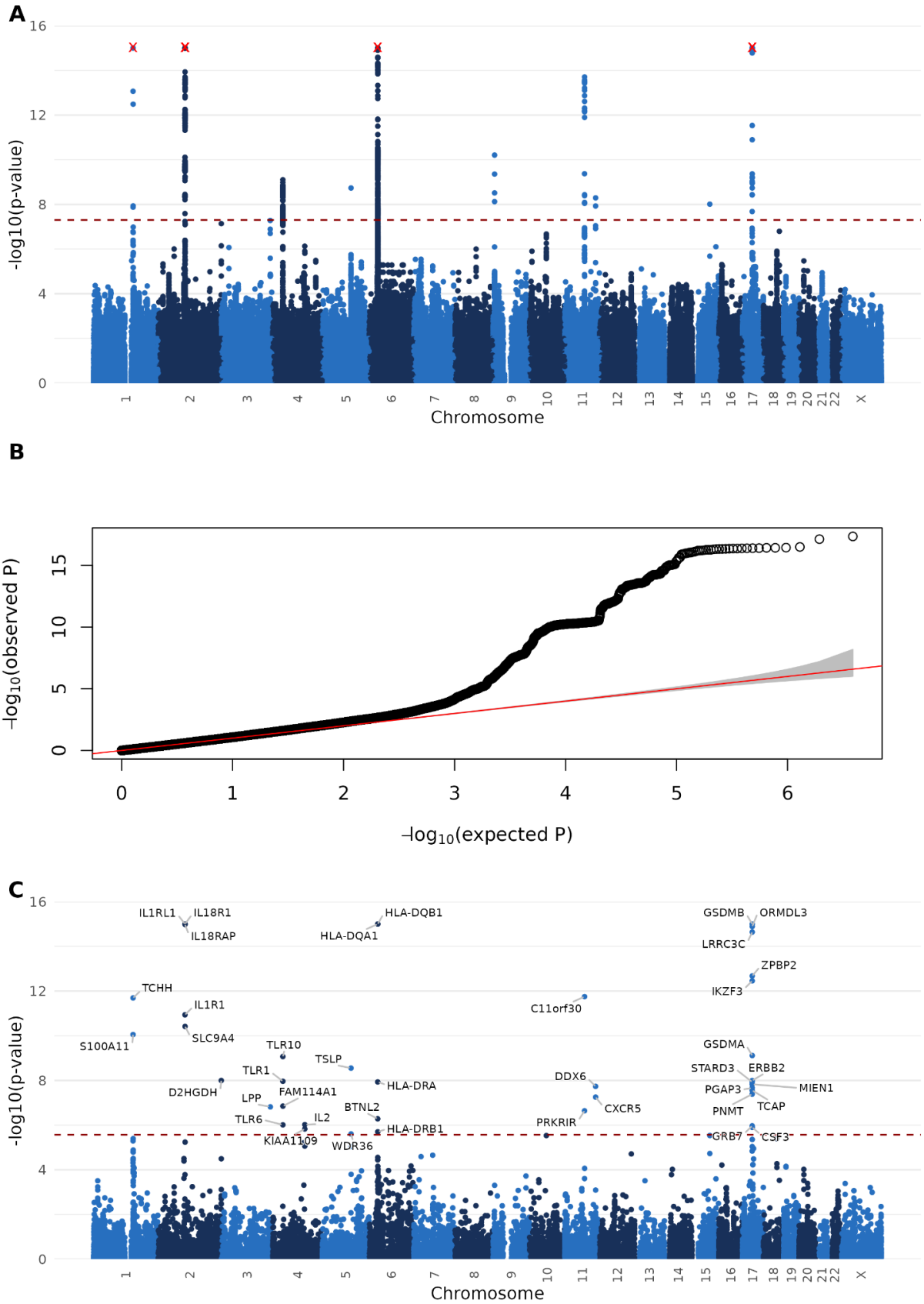



**Figure S7.3. GWAS results for MDD-related Cluster 3 membership in UKB data (N = 247,325).** A. SNP-based genome-wide Manhattan plot. Results capped at $10^{-15}$ are indicated with red crosses. B. QQ-plot. C. Gene-based genome-wide Manhattan plot. The statistically significant genes are indicated with labels.

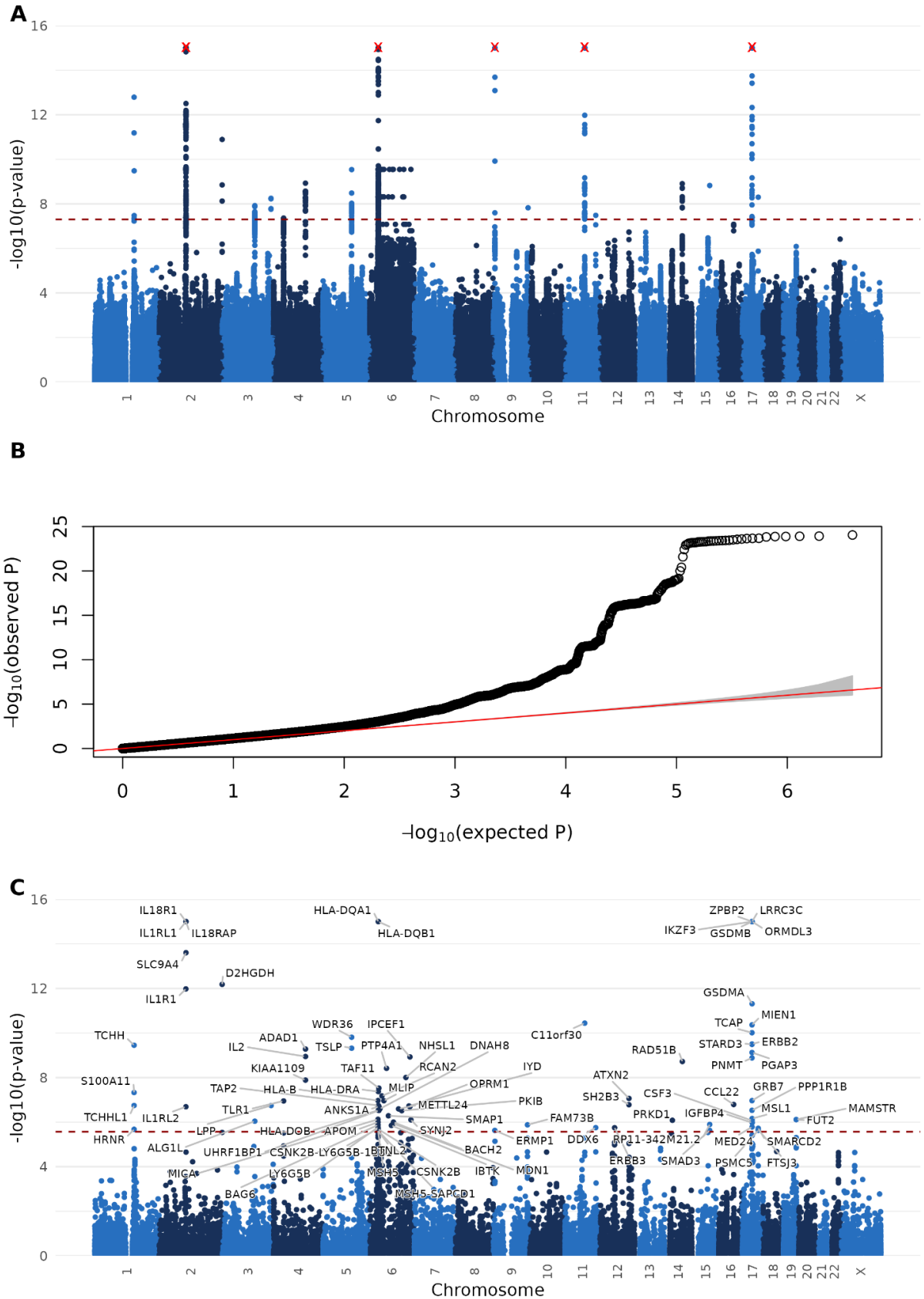



**Figure S7.4. GWAS results for MDD-related Cluster 4 membership in UKB data (N = 247,325)**. A. SNP-based genome-wide Manhattan plot. Results capped at $10^{-15}$ are indicated with red crosses. B. QQ-plot. C. Gene-based genome-wide Manhattan plot. The statistically significant genes are indicated with labels.

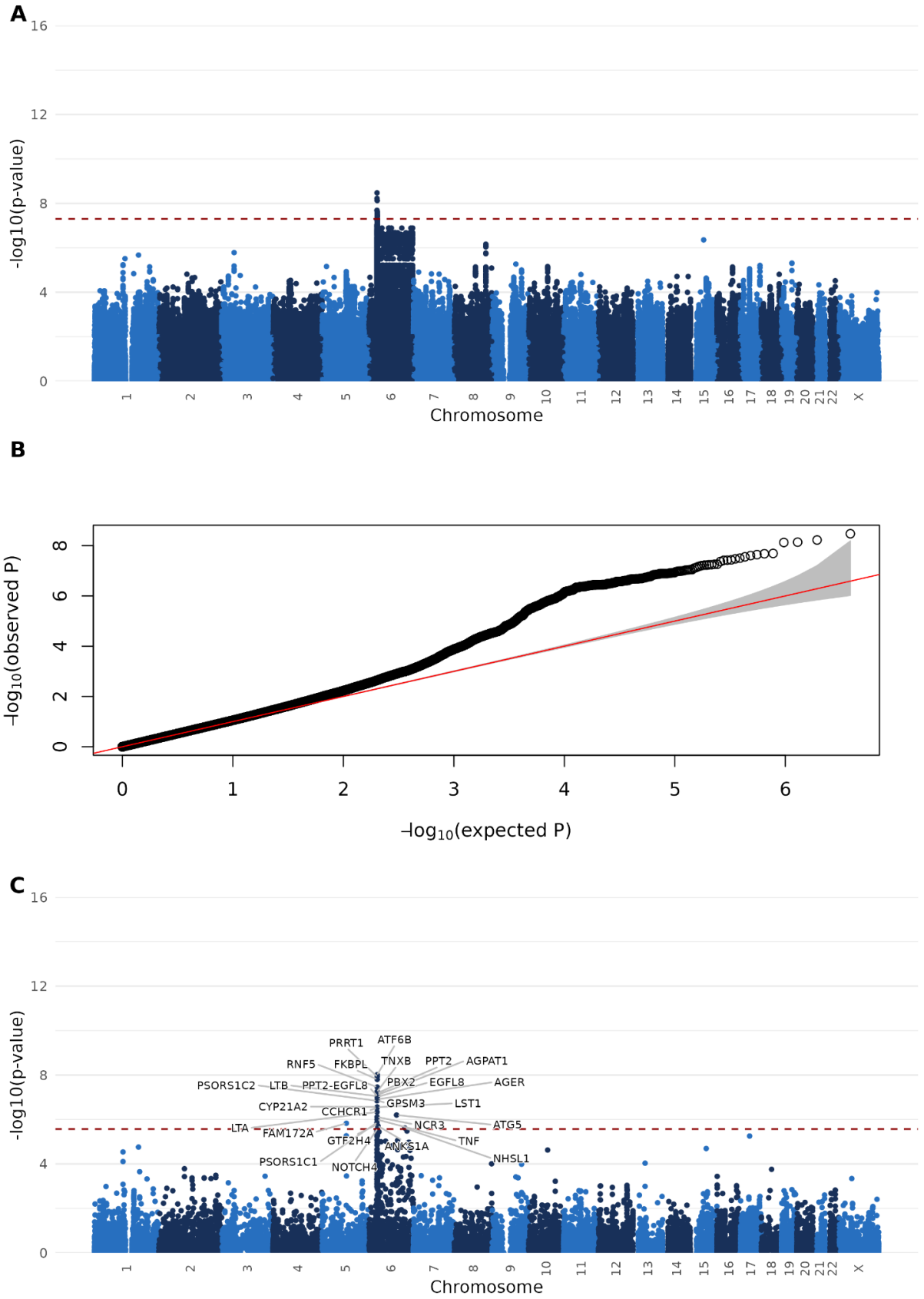



**Figure S7.5. GWAS results for MDD-related Cluster 5 membership in UKB data (N = 247,325).** A. SNP-based genome-wide Manhattan plot. B. QQ-plot. C. Gene-based genome-wide Manhattan plot. The statistically significant genes are indicated with labels.

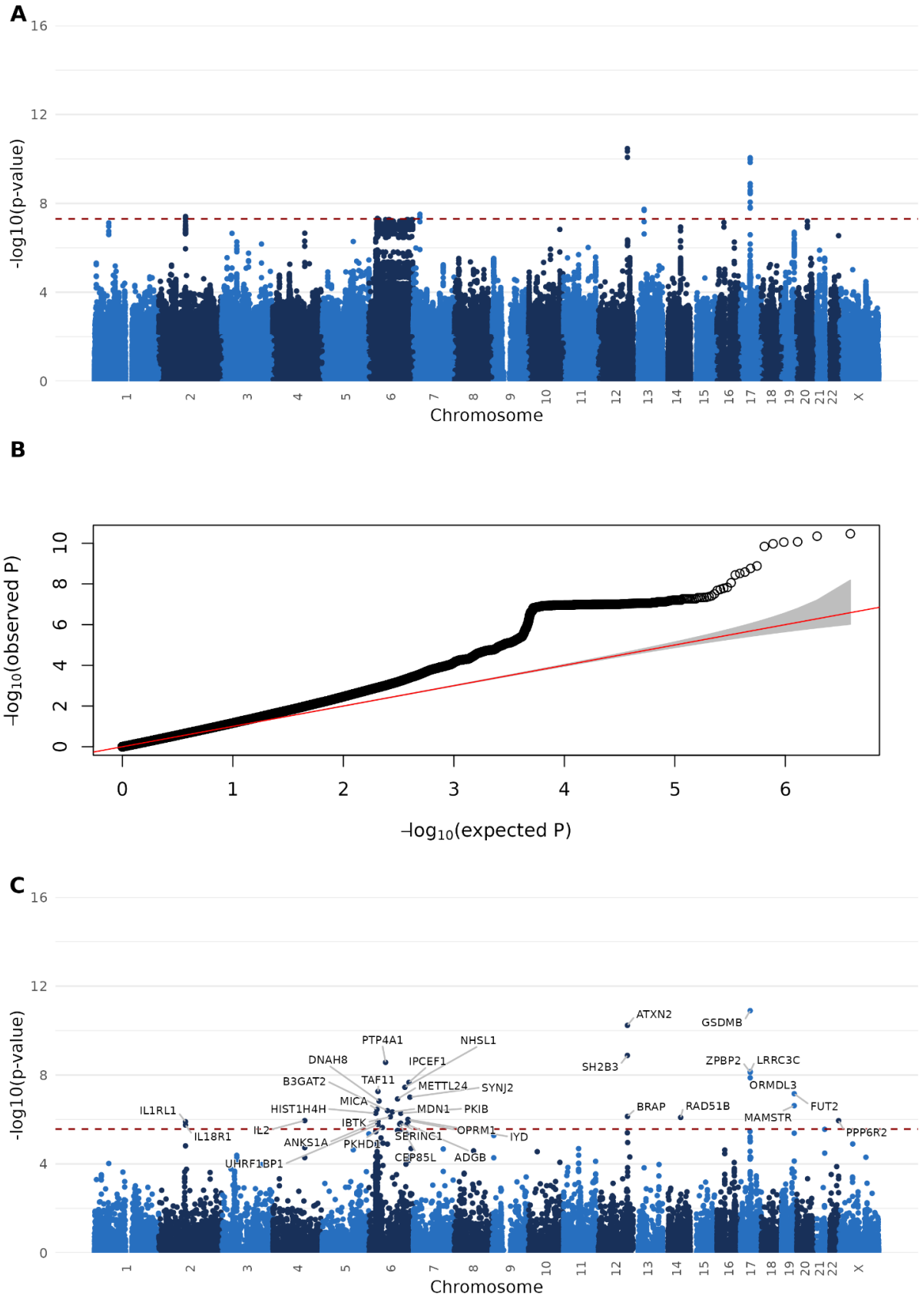



**Figure S7.6. GWAS results for MDD-related Cluster 6 membership in UKB data (N = 247,325).** A. SNP-based genome-wide Manhattan plot. B. QQ-plot. C. Gene-based genome-wide Manhattan plot. The statistically significant genes are indicated with labels.

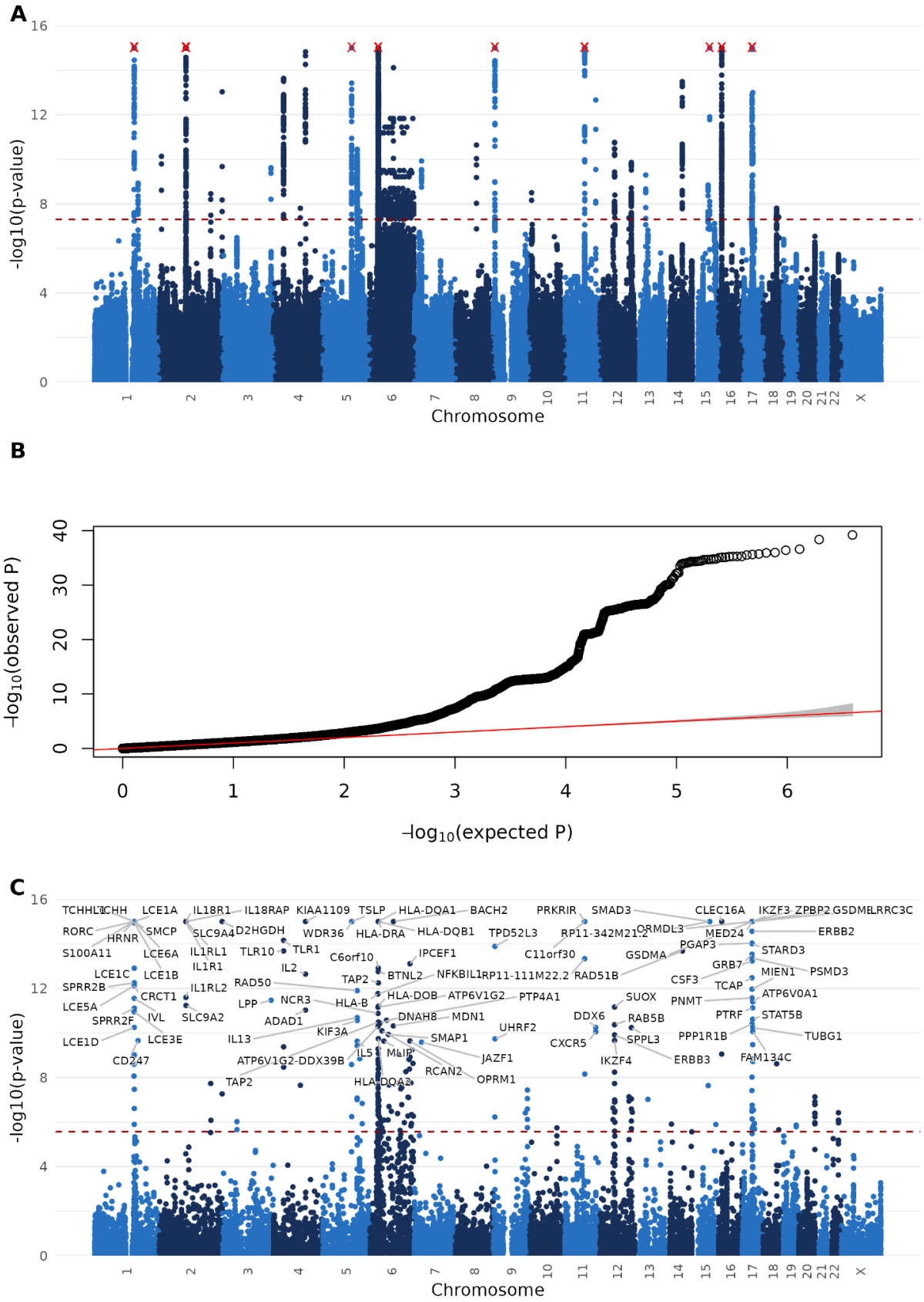
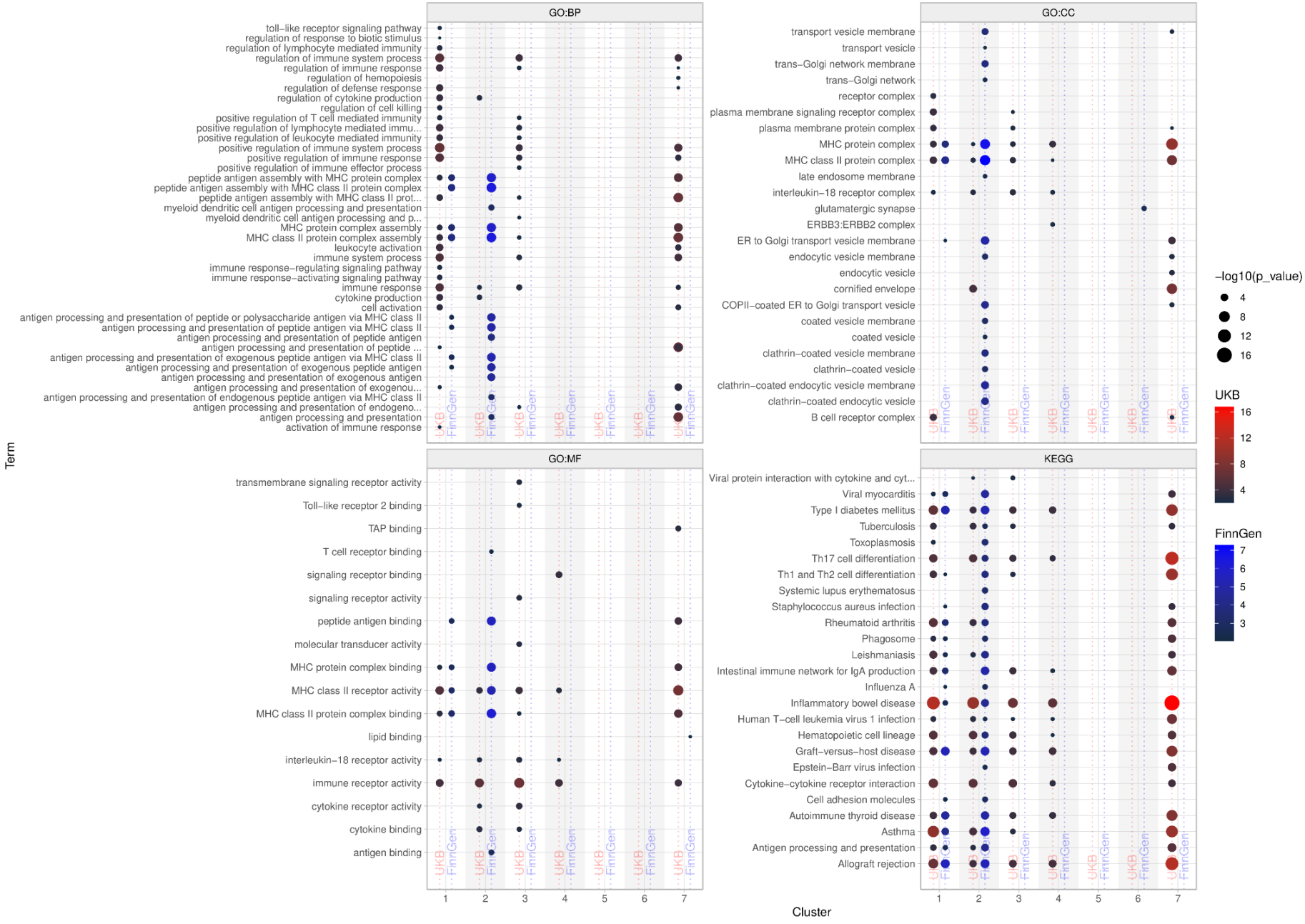



**Figure S7.7. GWAS results for MDD-related Cluster 7 membership in UKB data (N = 247,325)**. A. SNP-based genome-wide Manhattan plot. Results capped at $10^{-15}$ are indicated with red crosses. B. QQ-plot. C. Gene-based genome-wide Manhattan plot. The top 100 statistically significant genes are indicated with labels.


GO:BP
toll−like receptor signaling pathway
regulation of response to biotic stimulus
regulation of lymphocyte mediated immunity
regulation of immune system process
regulation of immune response
regulation of hemopoiesis
regulation of defense response
regulation of cytokine production
regulation of cell killing
positive regulation of T cell mediated immunity
positive regulation of lymphocyte mediated immu...
positive regulation of leukocyte mediated immunity
positive regulation of immune system process
positive regulation of immune response
positive regulation of immune effector process
peptide antigen assembly with MHC protein complex
peptide antigen assembly with MHC class II protein complex
peptide antigen assembly with MHC class II prot...
myeloid dendritic cell antigen processing and presentation
myeloid dendritic cell antigen processing and p...
MHC protein complex assembly
MHC class II protein complex assembly
leukocyte activation
immune system process
immune response−regulating signaling pathway
immune response−activating signaling pathway
immune response
cytokine production
cell activation
antigen processing and presentation of peptide or polysaccharide antigen via MHC class II
antigen processing and presentation of peptide antigen via MHC class II
antigen processing and presentation of peptide antigen
antigen processing and presentation of peptide ...
antigen processing and presentation of exogenous peptide antigen via MHC class II
antigen processing and presentation of exogenous peptide antigen
antigen processing and presentation of exogenous antigen
antigen processing and presentation of exogenou...
antigen processing and presentation of endogenous peptide antigen via MHC class II
antigen processing and presentation of endogeno...
antigen processing and presentation
activation of immune response
GO:CC
transport vesicle membrane
transport vesicle
trans−Golgi network membrane
trans−Golgi network
receptor complex
plasma membrane signaling receptor complex
plasma membrane protein complex
MHC protein complex
MHC class II protein complex
late endosome membrane
interleukin−18 receptor complex
glutamatergic synapse
ERBB3:ERBB2 complex
ER to Golgi transport vesicle membrane
endocytic vesicle membrane
endocytic vesicle
cornified envelope
COPII−coated ER to Golgi transport vesicle
coated vesicle membrane
coated vesicle
clathrin−coated vesicle membrane
clathrin−coated vesicle
clathrin−coated endocytic vesicle membrane
clathrin−coated endocytic vesicle
B cell receptor complex
GO:MF
transmembrane signaling receptor activity
Toll−like receptor 2 binding
TAP binding
T cell receptor binding
signaling receptor binding
signaling receptor activity
peptide antigen binding
molecular transducer activity
MHC protein complex binding
MHC class II receptor activity
MHC class II protein complex binding
lipid binding
interleukin−18 receptor activity
immune receptor activity
cytokine receptor activity
cytokine binding
antigen binding
KEGG
Viral protein interaction with cytokine and cyt...
Viral myocarditis
Type I diabetes mellitus
Tuberculosis
Toxoplasmosis
Th17 cell differentiation
Th1 and Th2 cell differentiation
Systemic lupus erythematosus
Staphylococcus aureus infection
Rheumatoid arthritis
Phagosome
Leishmaniasis
Intestinal immune network for IgA production
Influenza A
Inflammatory bowel disease
Human T−cell leukemia virus 1 infection
Hematopoietic cell lineage
Graft−versus−host disease
Epstein−Barr virus infection
Cytokine−cytokine receptor interaction
Cell adhesion molecules
Autoimmune thyroid disease
Asthma
Antigen processing and presentation
Allograft rejection
UKB
FinnGen
1
2
3
4
5
6
7
Term
Cluster
−log10(p_value)
4
8
12
16
UKB
16
12
8
4
FinnGen
7
6
5
4
3

335

**Figure S8. Functional overrepresentation analysis of UKB and FinnGen GWAS.** Results of overrepresentation analysis based on the g:Profiler method. Each dot represents a statistically significantly overrepresented Gene Ontology term or KEGG pathway (adjusted *p*-value < 0.01, based on g:SCS method). The color scales correspond to the cohorts, red dots indicate UKB results (at the left part of each column), and the blue dots indicate the FinnGen results (shown at the right part of each column). The size of the dots indicate the strength of the significance, measured by the negative log-transformed *p*-values.

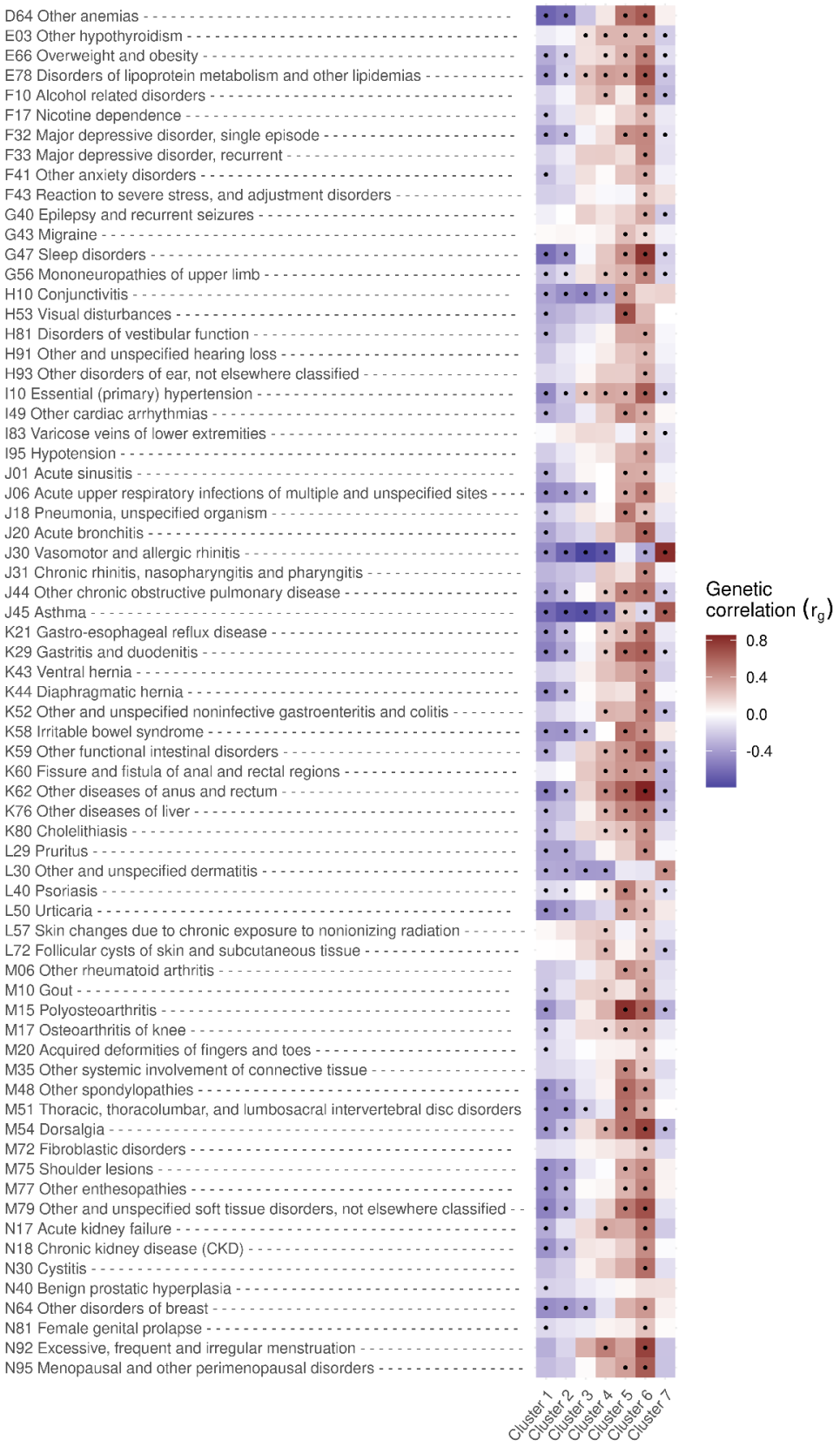
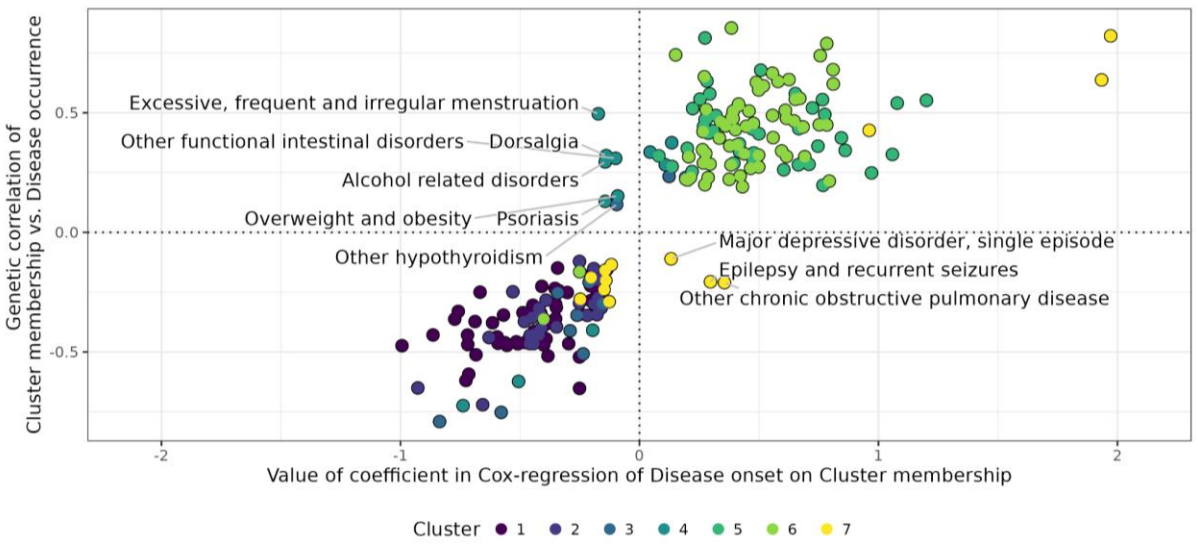



**Figure S9. Genetic correlation between clusters and the consensual diseases on which the definitions of the clusters are based.** Only those diseases are shown for which at least one cluster showed a statistically significant genetic correlation. *p*-values were adjusted using Benjamini-Hochberg method. Significant genetic correlations are indicated with dots. The GWAS of the cluster memberships and all diseases were performed on the UKB cohort using the same covariates and filtering procedure (see online Methods).




346

347 **Figure S10. Comparison of the value of coefficient in Cox regression of cluster membership on disease onset**
348 **versus the genetic correlation between the cluster membership and the disease occurrence in the UKB data.**
349 Only those diseases are shown that are statistically significant according to both methods. Each point represents
350 a disease according to a given cluster (indicated by the color of the point). Diseases for which the sign of the
351 effects differ are indicated with labels.

352

353

354

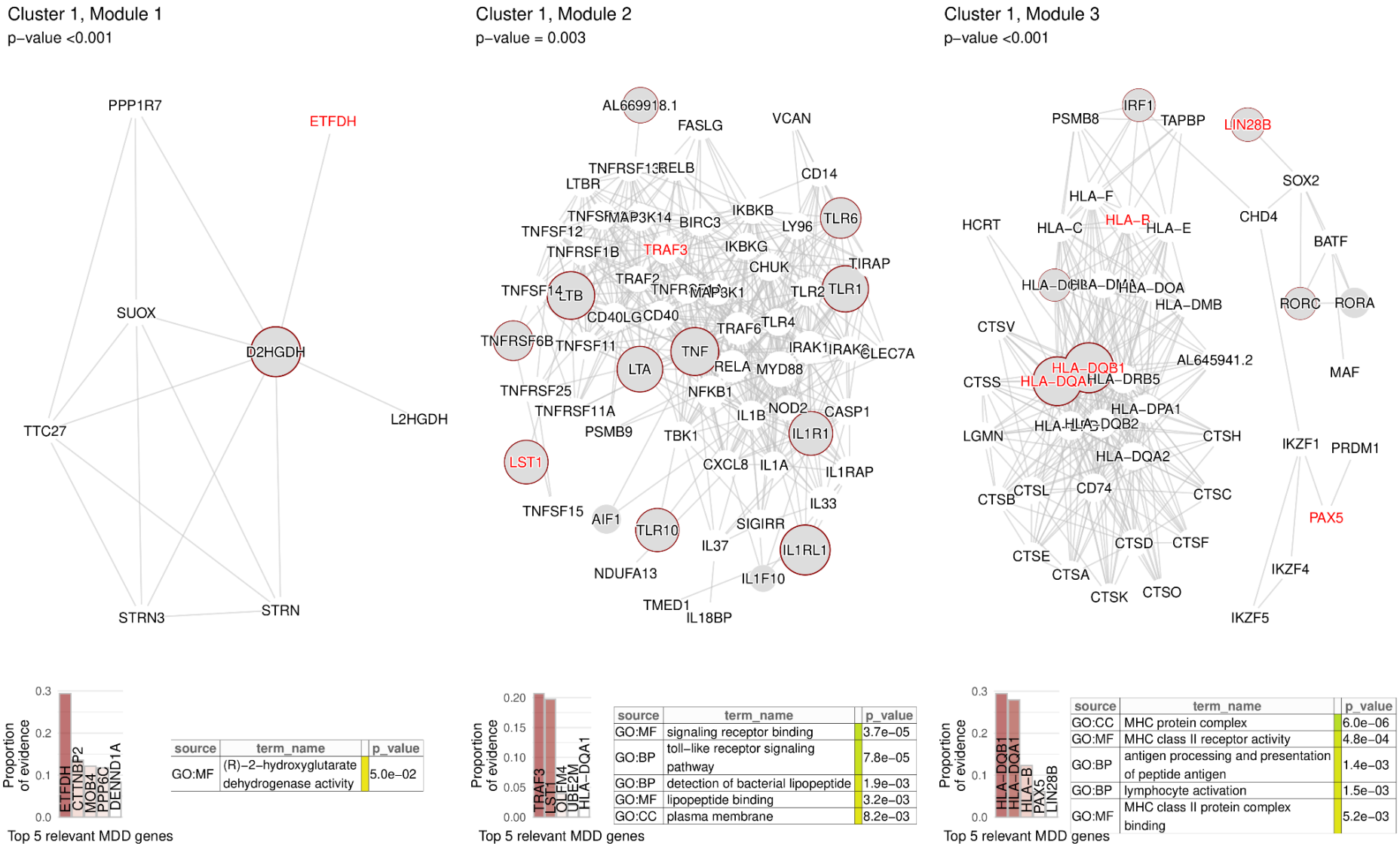
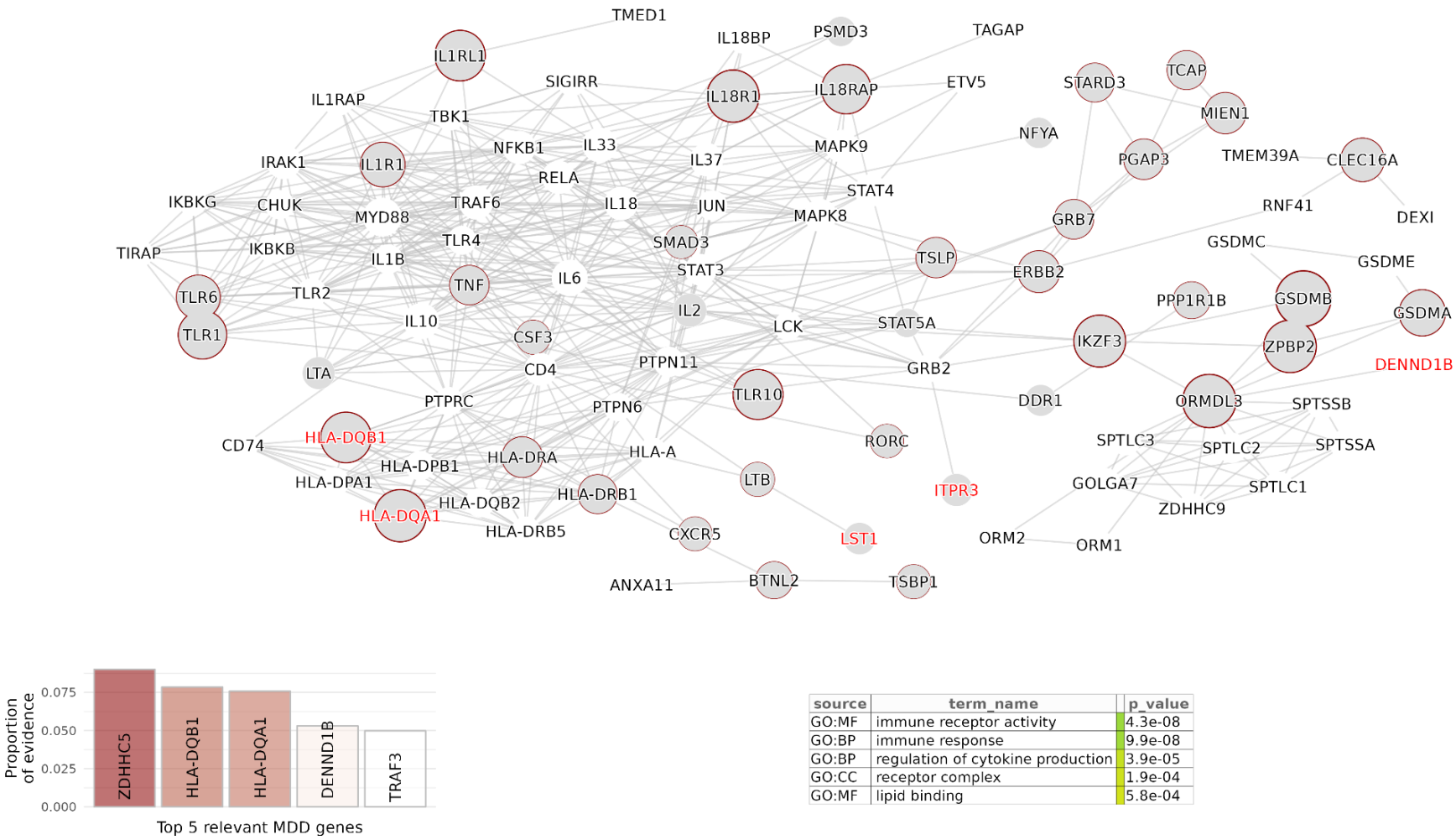


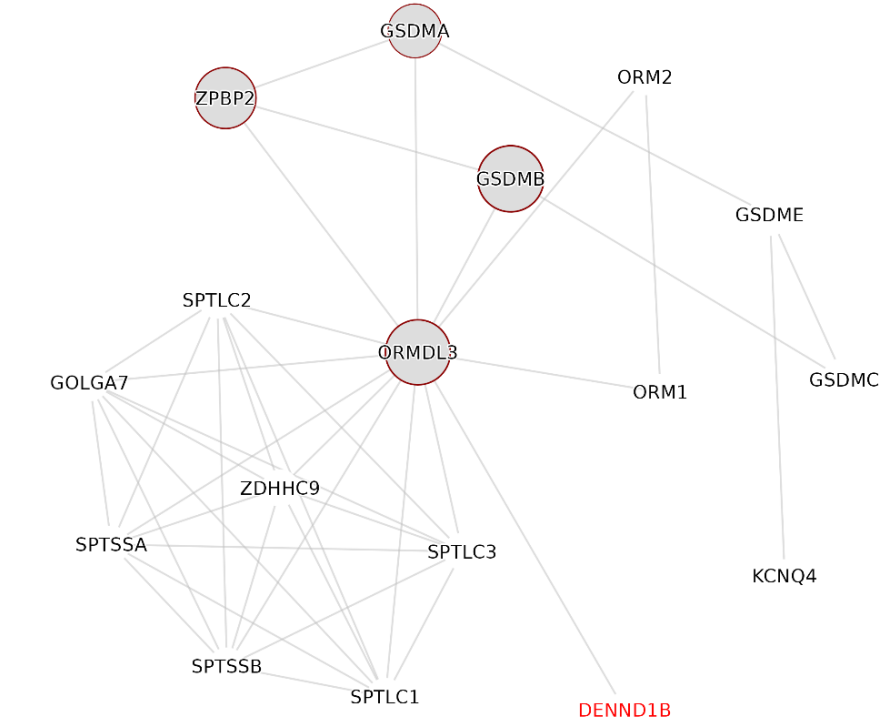

**Figure S11.1. Functional modules of the human interactome in Cluster 1 that are pleiotropic with MDD.** Top: Pleiotropic functional modules. Gray nodes were used as evidences in a network propagation algorithm used to score all genes in the STRING PPI network. The size of the nodes indicate the propagated score. Nodes with red border are statistically significant according to the GWAS of Cluster 1. Modules of the top scoring genes were identified, and those modules are shown in which the propagated effect of MDD-associated genes (shown with red labels, according to a GWAS meta-analysis of Howard et al.) are significantly enriched. Bottom left: The

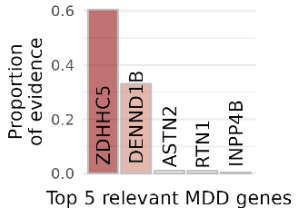
relative contribution of the top 5 MDD-associated genes to the module. Bottom right: The at most top 5 most relevant statistically significant Gene Ontology terms overrepresented in the module.



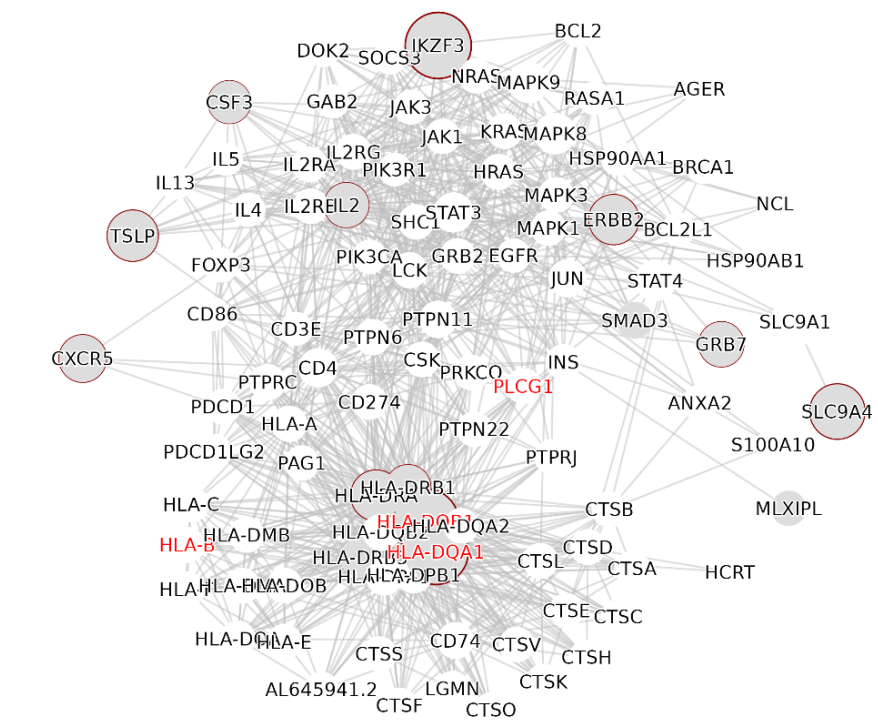

| source | term_name | p_value |
|---|---|---|
| GO:MF | immune receptor activity | 4.3e-08 |
| GO:BP | immune response | 9.9e-08 |
| GO:BP | regulation of cytokine production | 3.9e-05 |
| GO:CC | receptor complex | 1.9e-04 |
| GO:MF | lipid binding | 5.8e-04 |

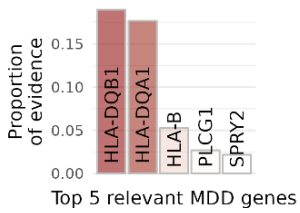
**Figure S11.2. Functional module of the human interactome in Cluster 2 that is pleiotropic with MDD.** See caption of Figure S11.1. for details.




| source | term_name | p_value |
|---|---|---|
| GO:MF | phosphatidylinositol phosphate binding | 7.0e-04 |
| GO:MF | phosphatidylserine binding | 1.4e-02 |
| GO:BP | pyroptosis | 1.6e-02 |




| source | term_name | p_value |
|---|---|---|
| GO:CC | MHC protein complex | 1.3e-07 |
| GO:MF | peptide antigen binding | 1.7e-06 |
| GO:MF | MHC class II receptor activity | 4.1e-06 |
| GO:BP | regulation of immune system process | 4.8e-05 |
| GO:BP | regulation of hemopoiesis | 7.4e-05 |

**Figure S11.3. Functional modules of the human interactome in Cluster 3 that are pleiotropic with MDD.** See caption of Figure S11.1. for details.

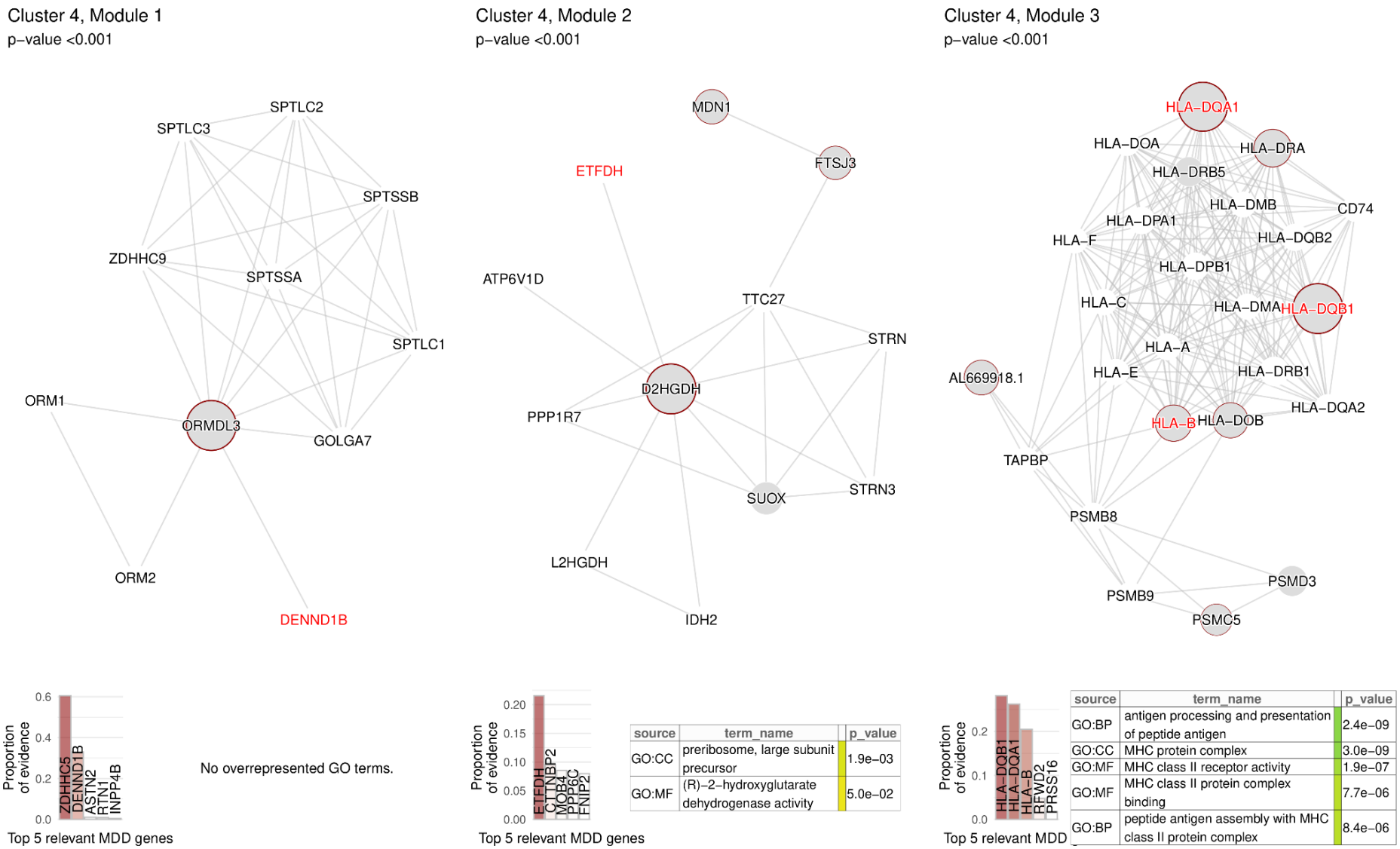



**Figure S11.4. Functional modules of the human interactome in Cluster 4 that are pleiotropic with MDD.** See caption of Figure S11.1. for details.

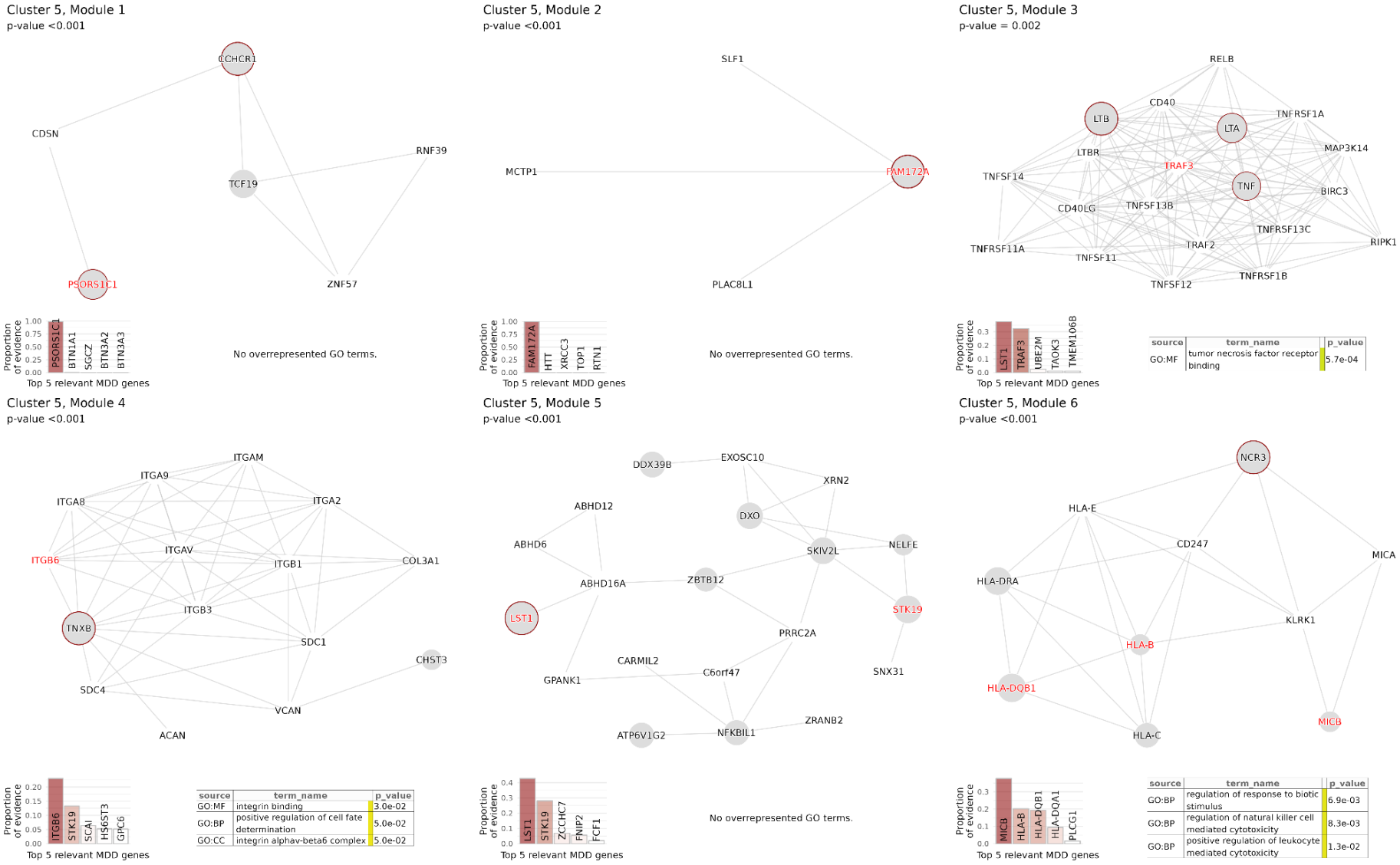



**Figure S11.5. Functional modules of the human interactome in Cluster 5 that are pleiotropic with MDD.** See caption of Figure S11.1. for details.

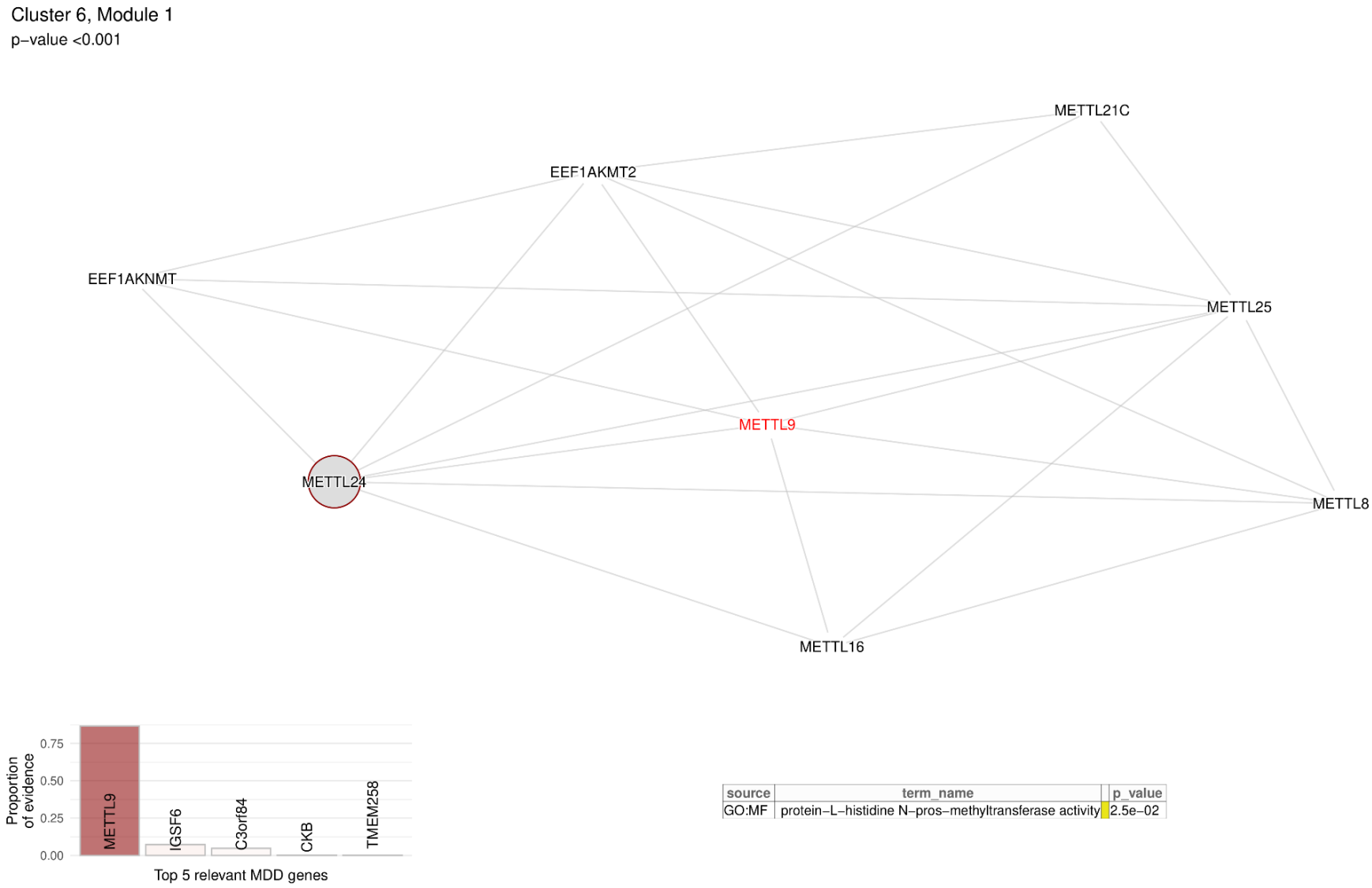



**Figure S11.6. Functional module of the human interactome in Cluster 6 that is pleiotropic with MDD.** See caption of Figure S11.1. for details.

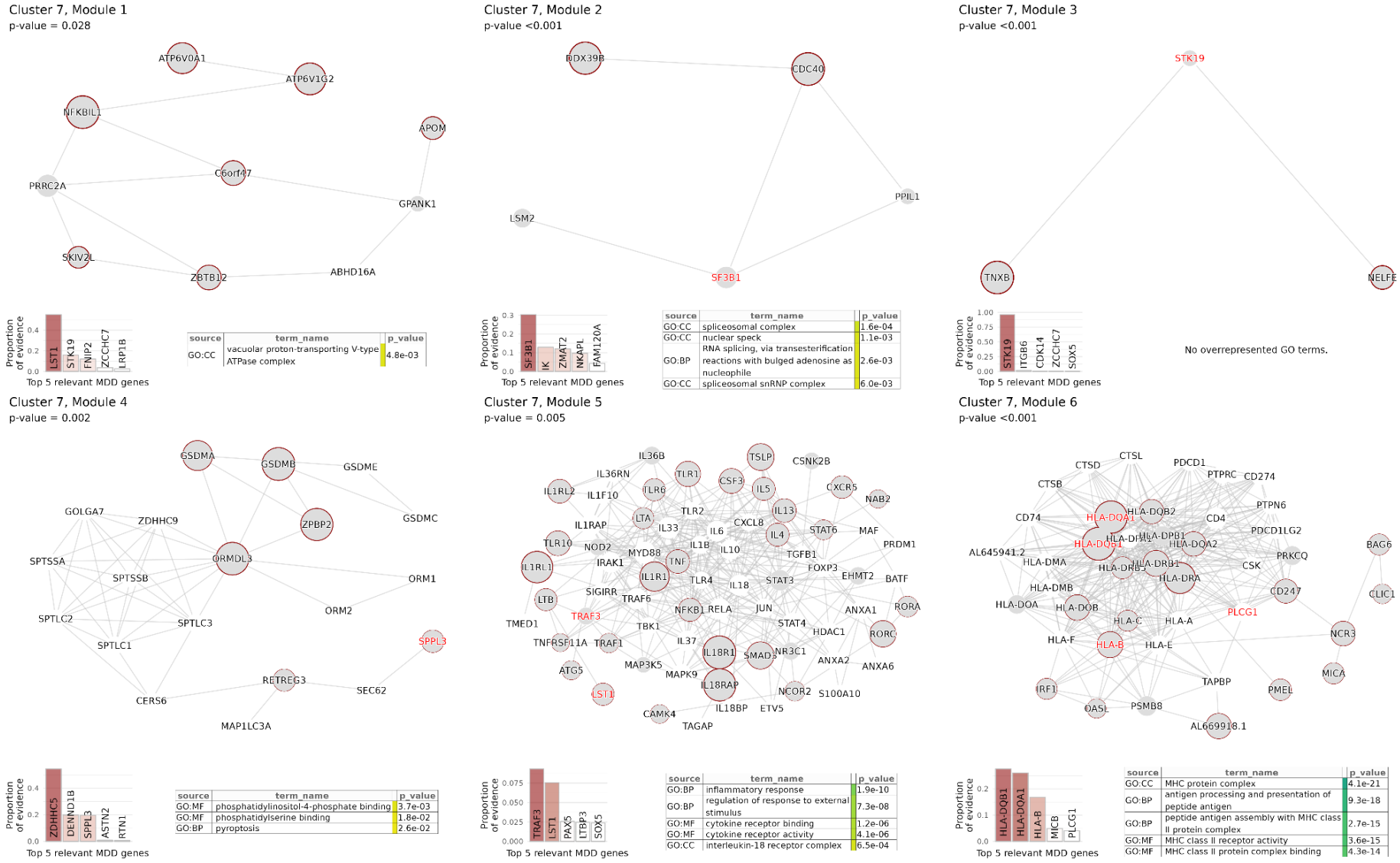



**Figure S11.7. Functional modules of the human interactome in Cluster 7 that are pleiotropic with MDD.** See caption of Figure S11.1. for details.

## 4. Supplementary Figures for results section *Modifiable risk-factor profiles of MDD-related multimorbidity clusters in the UKB cohort*

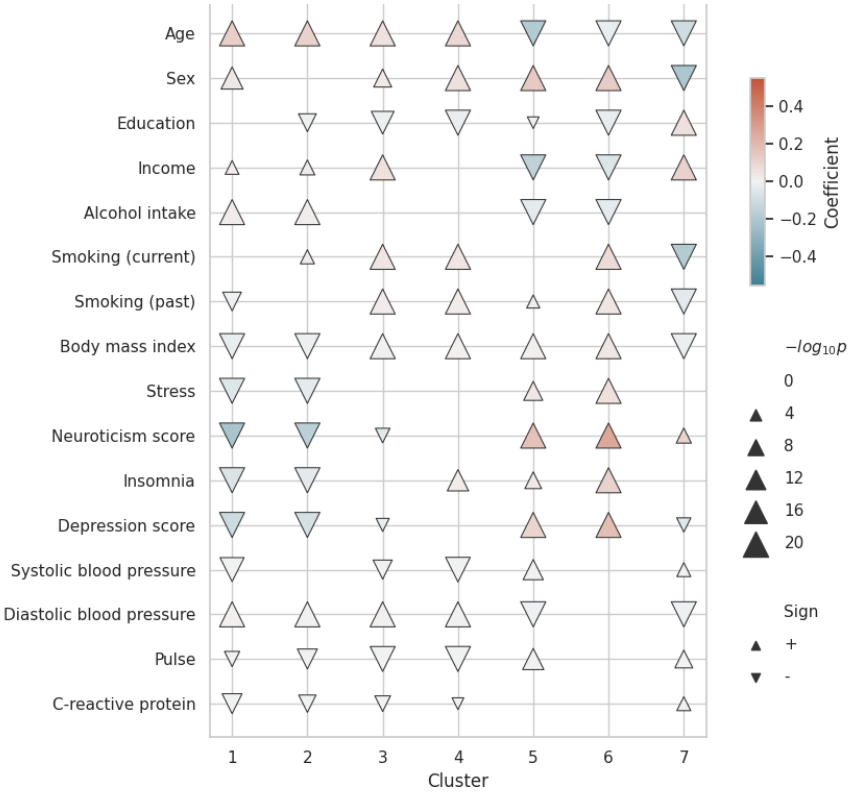



**Figure S12. Complex linear regression models involving available modifiable factors included in the UK Biobank dataset for each cluster (N = 247,325).** The "complex regression model" uses all available risk factors at once in a single model. The posterior log odds of being in a given cluster serves as the target variable. The direction of the triangles corresponds to the sign of the coefficient (positive - upwards; negative - downwards) and the color details the magnitude. The size of the triangles is proportional to the -log10 *p*-value, and only nominally significant values are shown (-log10*p* > 4). CRP: C-reactive protein; bp: blood pressure; sex: 1-males, 2-females.

## 5. Supplementary Figures for results section *Validation of MDD-related multimorbidity profiles at the genetic and modifiable risk-factor levels*

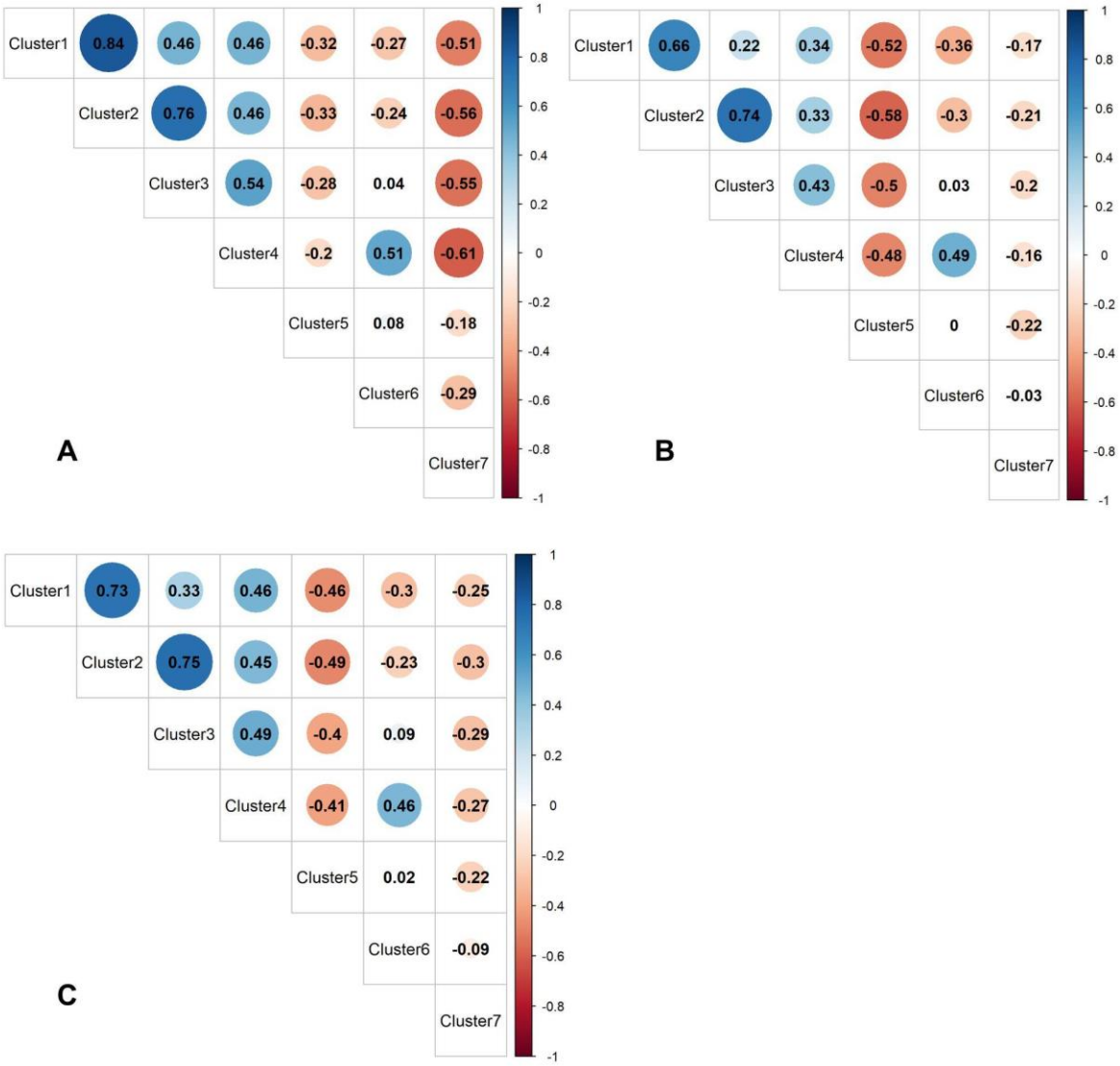



**Figure S13. Pearson correlation pattern of GWAS derived beta values between MDD-related clusters. (**A) UK Biobank (N = 247,325); (B) THL (N = 30,961); (C) FinnGen (N = 277,252).

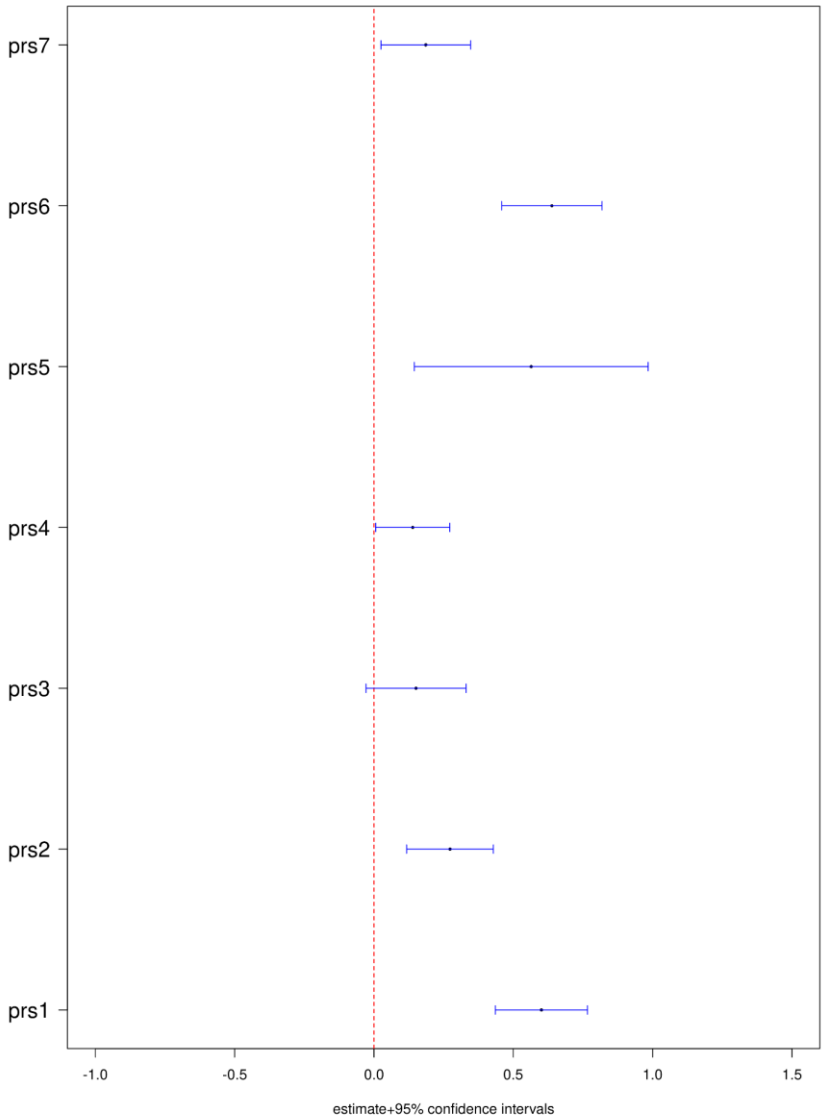



**Figure S14. Results for polygenic risk score (PRS) analysis in THL (N = 30,961).** Results of linear regression analyses between posterior log-odds cluster membership variable and PRS for cluster membership based on UKB GWAS results for each of the seven clusters. Adjusted for age, sex, cohort, batch, east/west-background. Estimate and 95% confidence interval displayed.

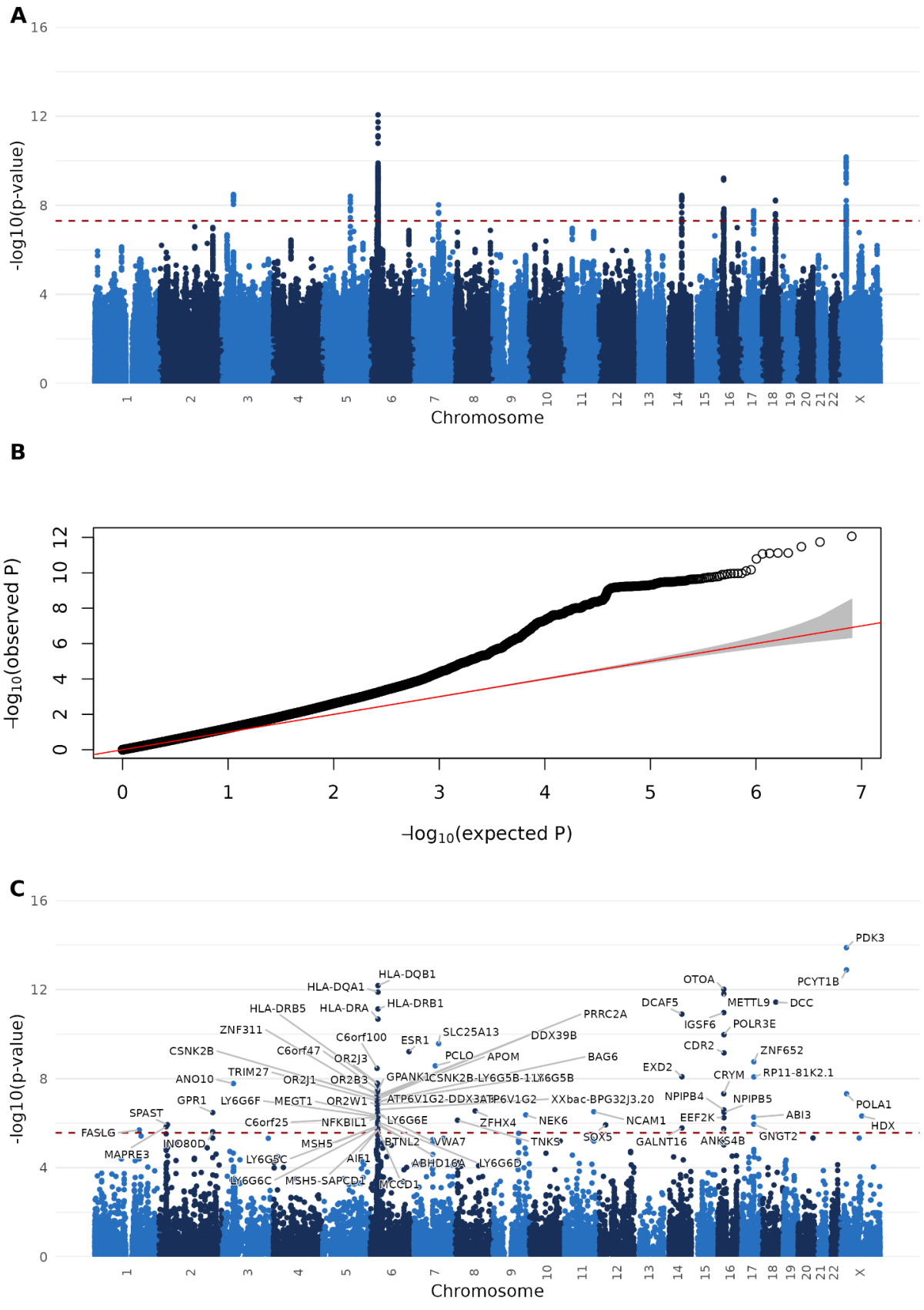



**Figure S15.1. GWAS results for MDD-related Cluster 1 membership in FinnGen data (N = 277,252).** A. SNP-based genome-wide Manhattan plot. B. QQ-plot. C. Gene-based genome-wide Manhattan plot. The statistically significant genes are indicated with labels.

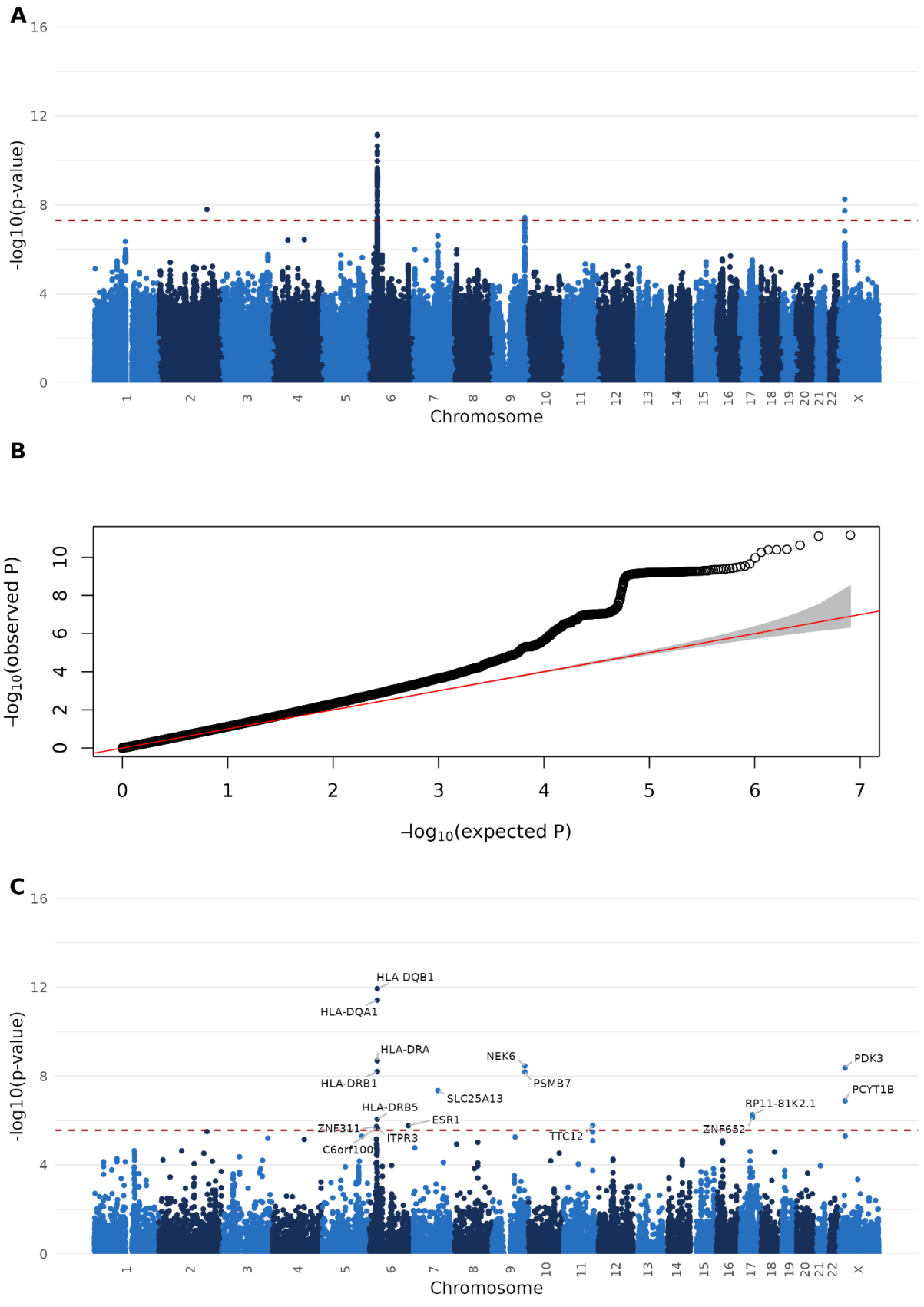



**Figure S15.2. GWAS results for MDD-related Cluster 2 membership in FinnGen data (N = 277,252).** A. SNP-based genome-wide Manhattan plot. B. QQ-plot. C. Gene-based genome-wide Manhattan plot. The statistically significant genes are indicated with labels.

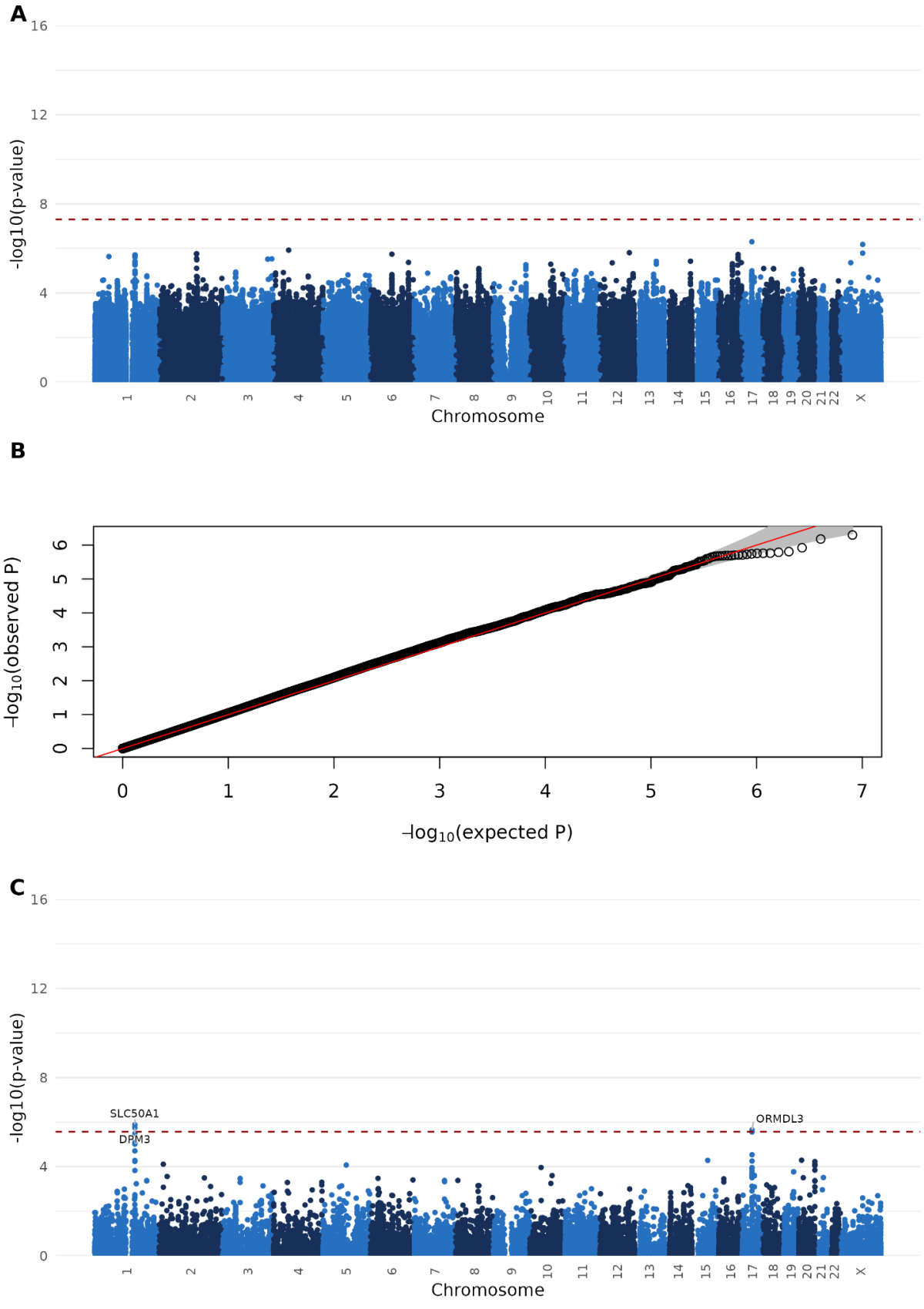



**Figure S15.3. GWAS results for MDD-related Cluster 3 membership in FinnGen data (N = 277,252).** A. SNP-based genome-wide Manhattan plot. B. QQ-plot. C. Gene-based genome-wide Manhattan plot. The statistically significant genes are indicated with labels.

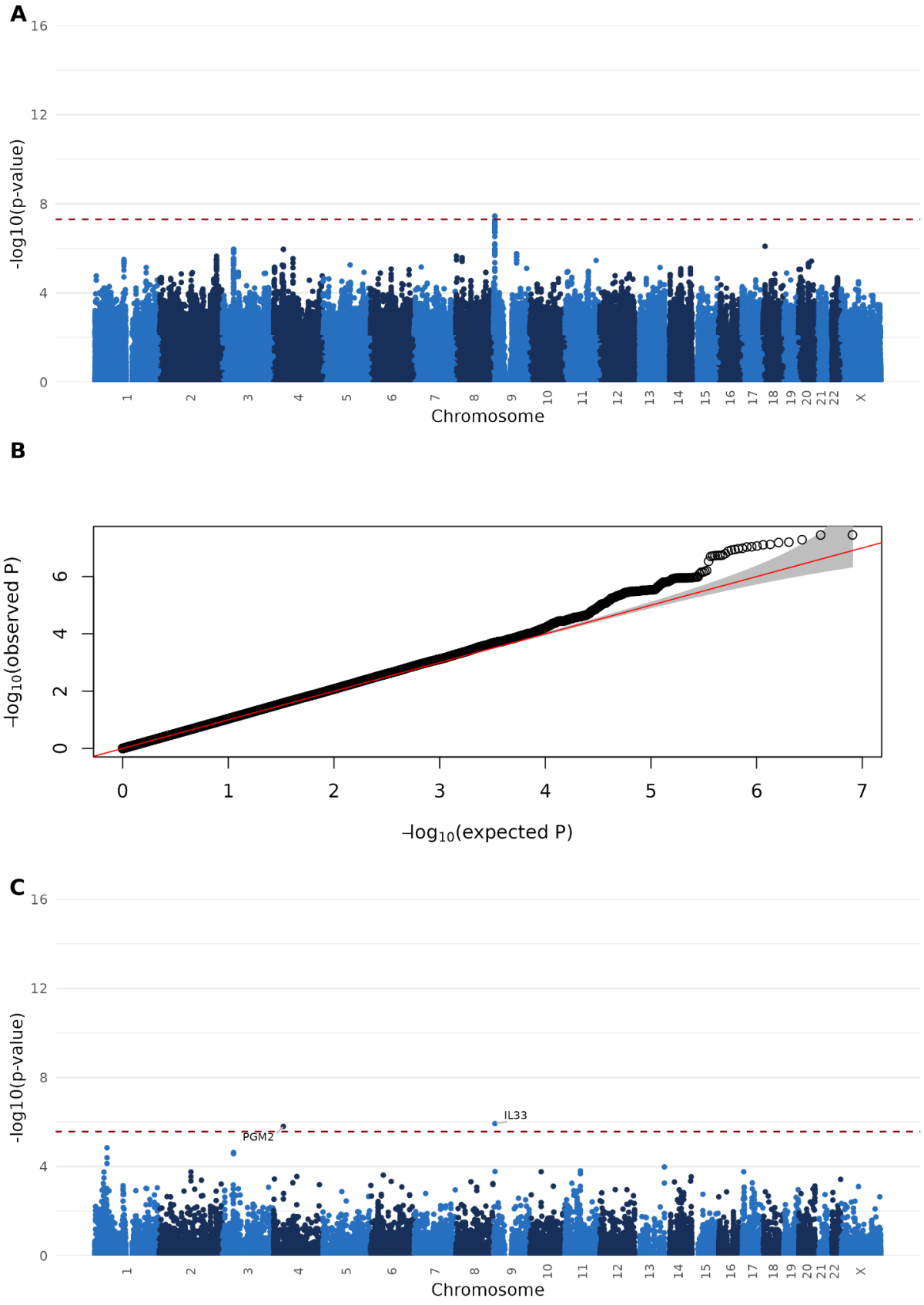



**Figure S15.4. GWAS results for MDD-related Cluster 4 membership in FinnGen data (N = 277,252).** A. SNP-based genome-wide Manhattan plot. B. QQ-plot. C. Gene-based genome-wide Manhattan plot. The statistically significant genes are indicated with labels.

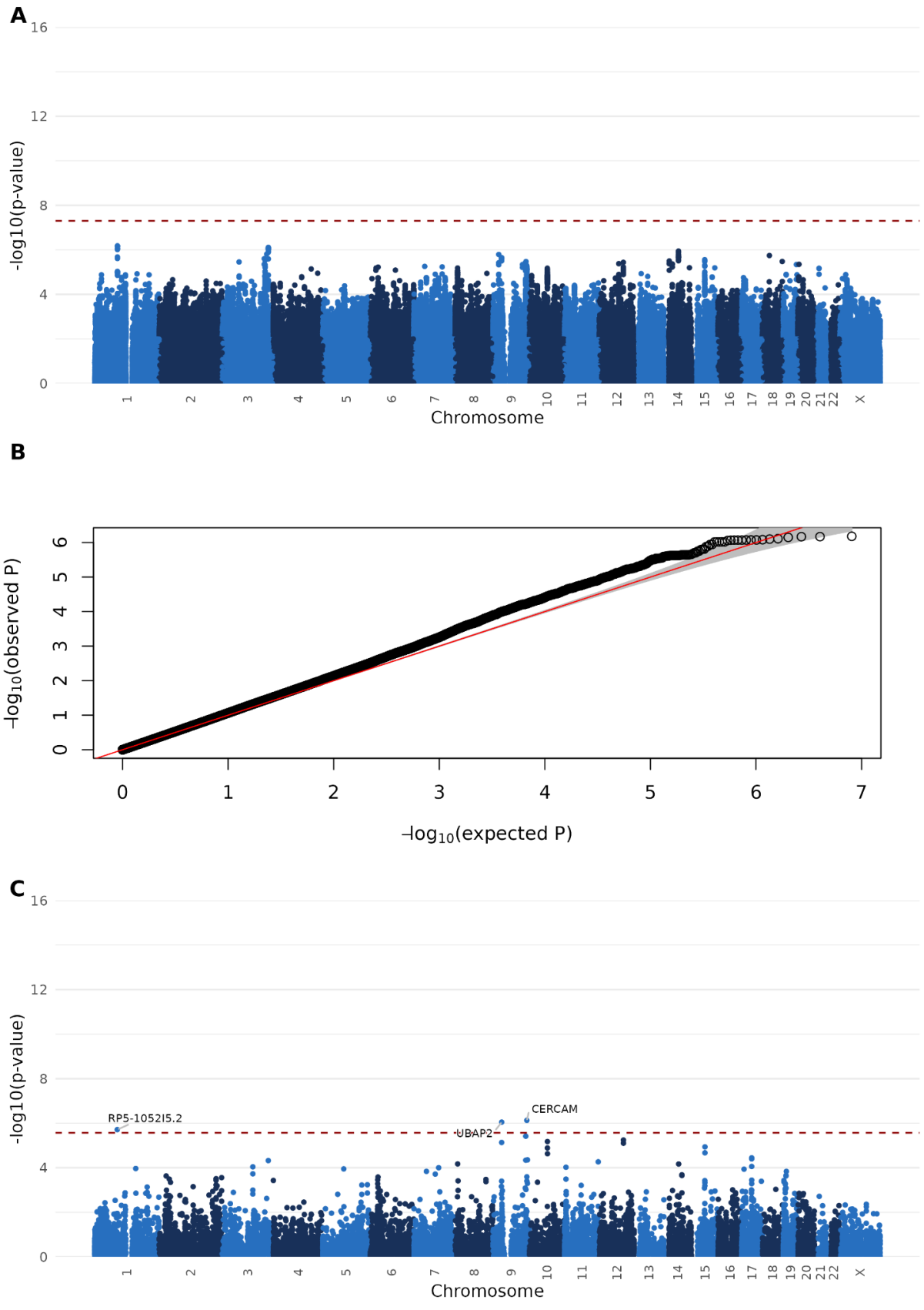



**Figure S15.5. GWAS results for MDD-related Cluster 5 membership in FinnGen data (N = 277,252).** A. SNP-based genome-wide Manhattan plot. B. QQ-plot. C. Gene-based genome-wide Manhattan plot. The statistically significant genes are indicated with labels.

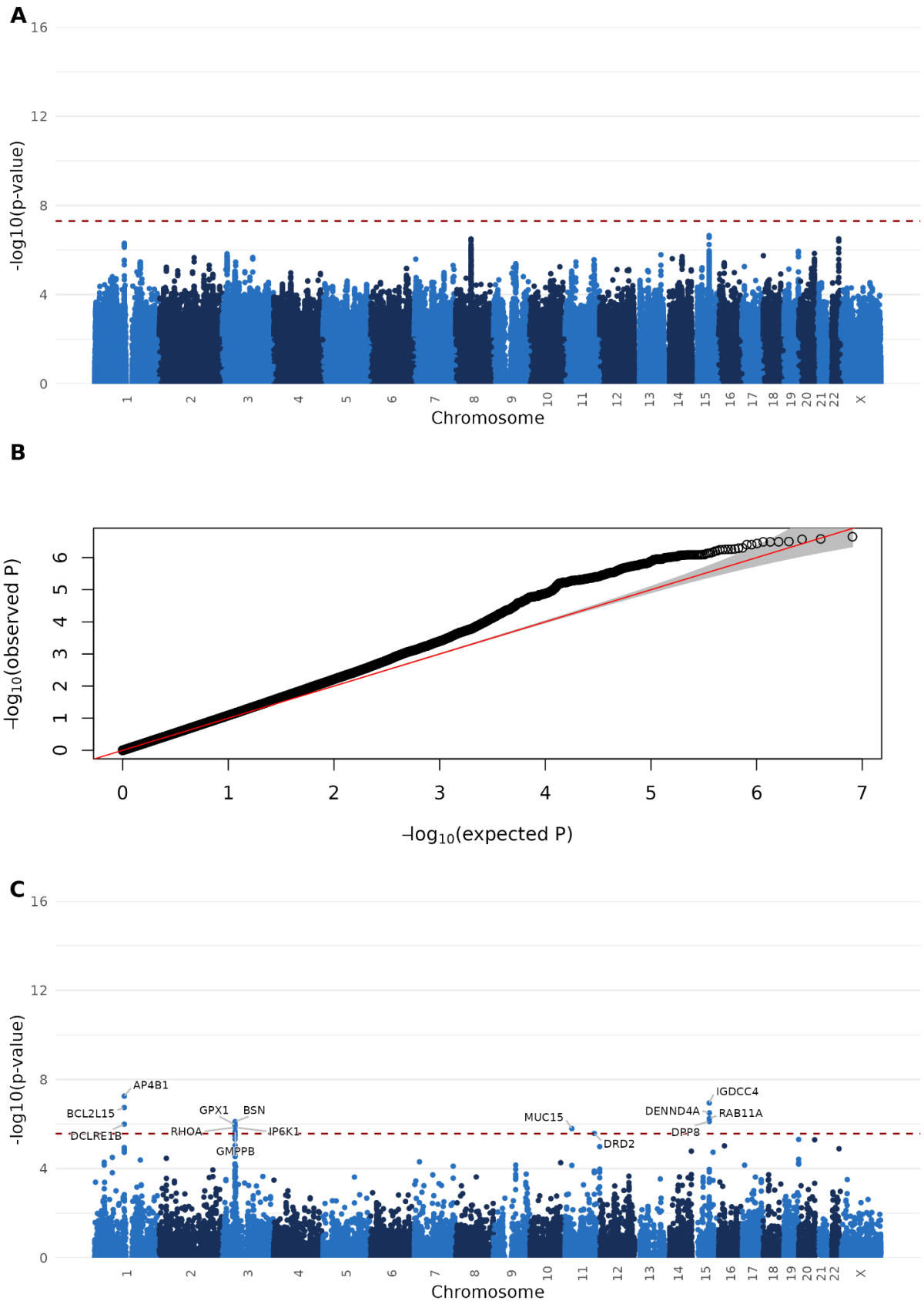



**Figure S15.6. GWAS results for MDD-related Cluster 6 membership in FinnGen data (N = 277,252).** A. SNP-based genome-wide Manhattan plot. B. QQ-plot. C. Gene-based genome-wide Manhattan plot. The statistically significant genes are indicated with labels.

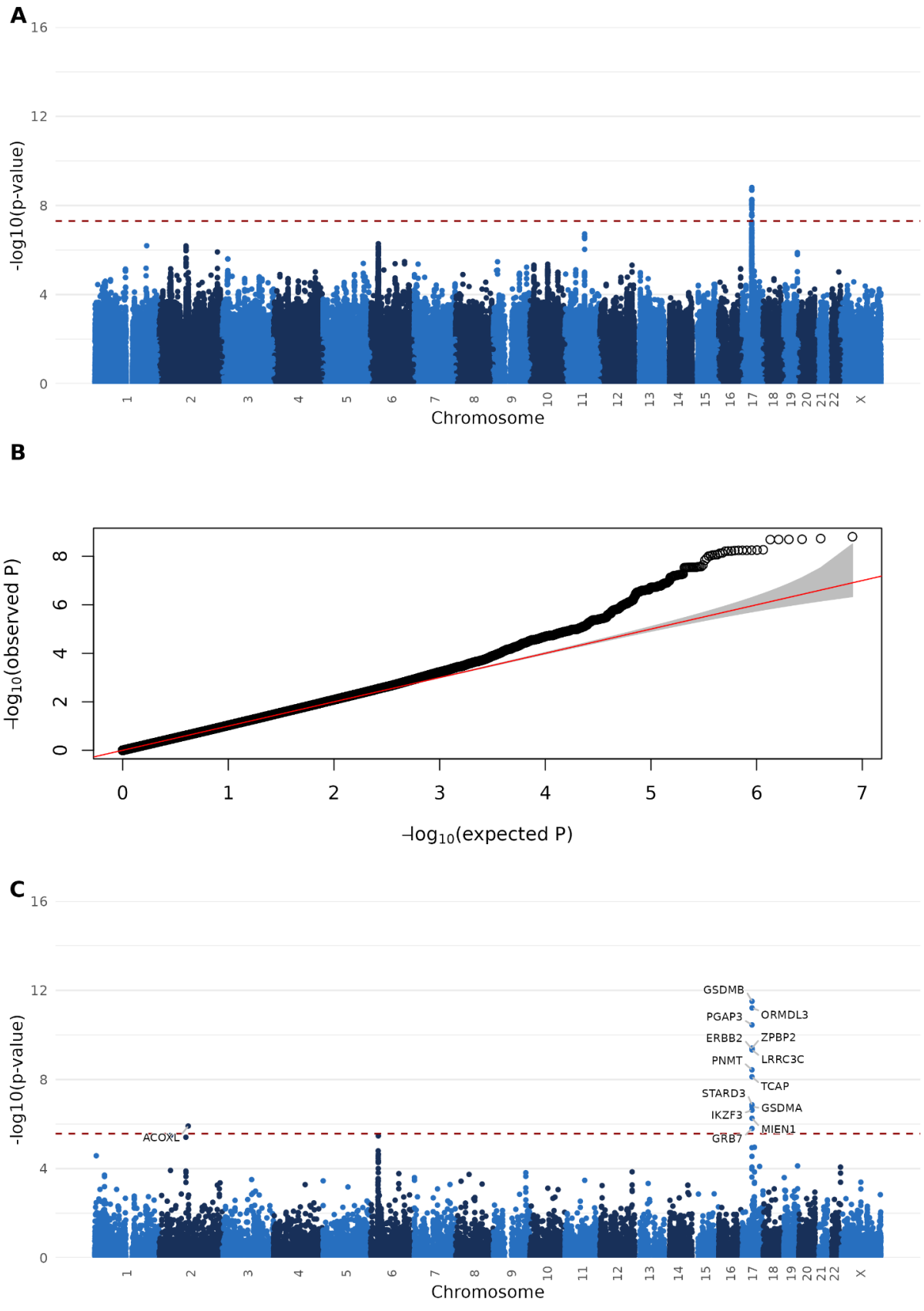



**Figure S15.7. GWAS results for MDD-related Cluster 7 membership in FinnGen data (N = 277,252).** A. SNP-based genome-wide Manhattan plot. B. QQ-plot. C. Gene-based genome-wide Manhattan plot. The statistically significant genes are indicated with labels.

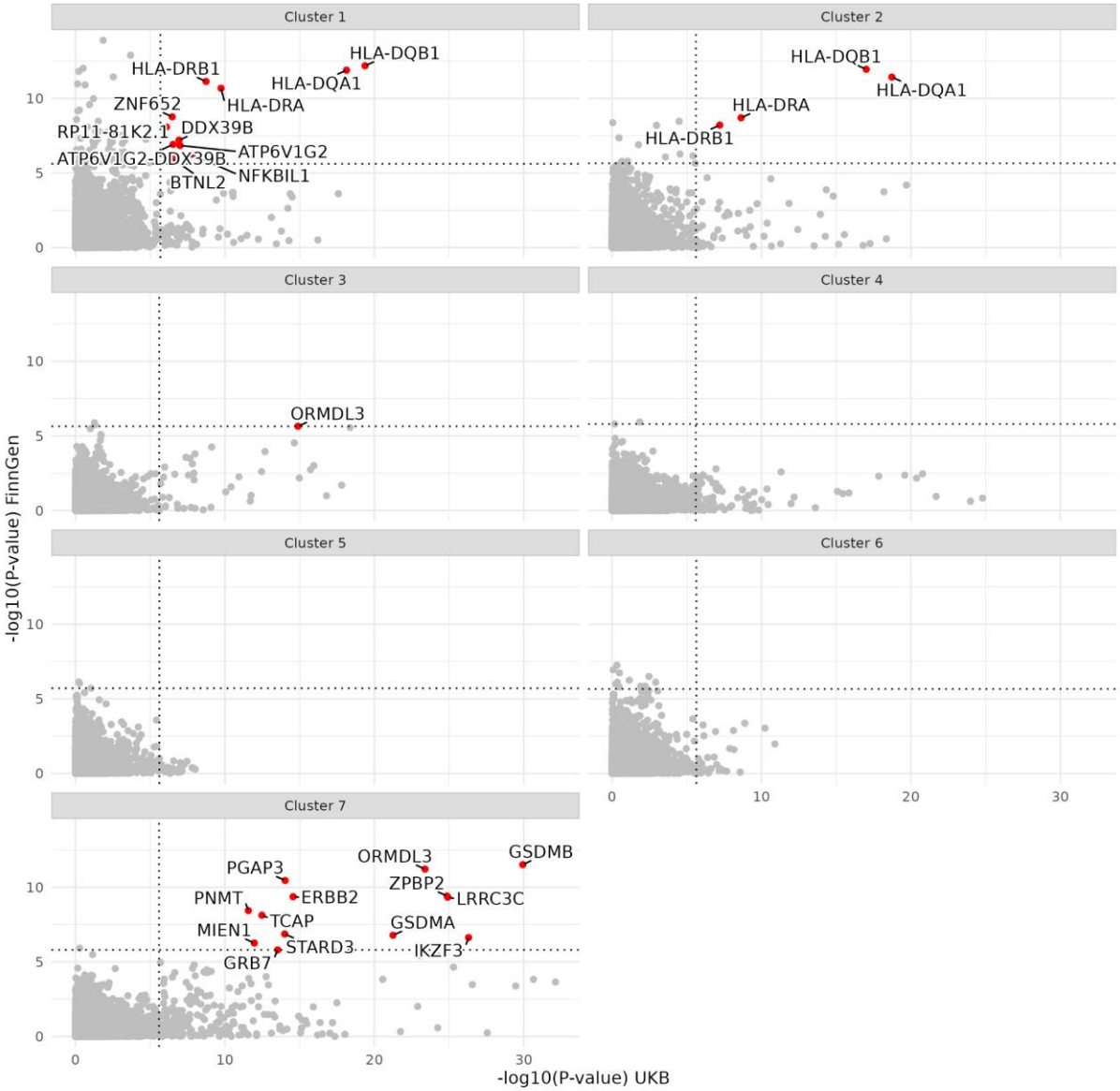



**Figure S16. Scatter plot of gene-level aggregated (negative log-transformed) *p*-values in UKB versus FinnGen.** Gene-level *p*-values were aggregated with MAGMA. Genes that are significant according to Holm's method are indicated with red dots and shown with labels.

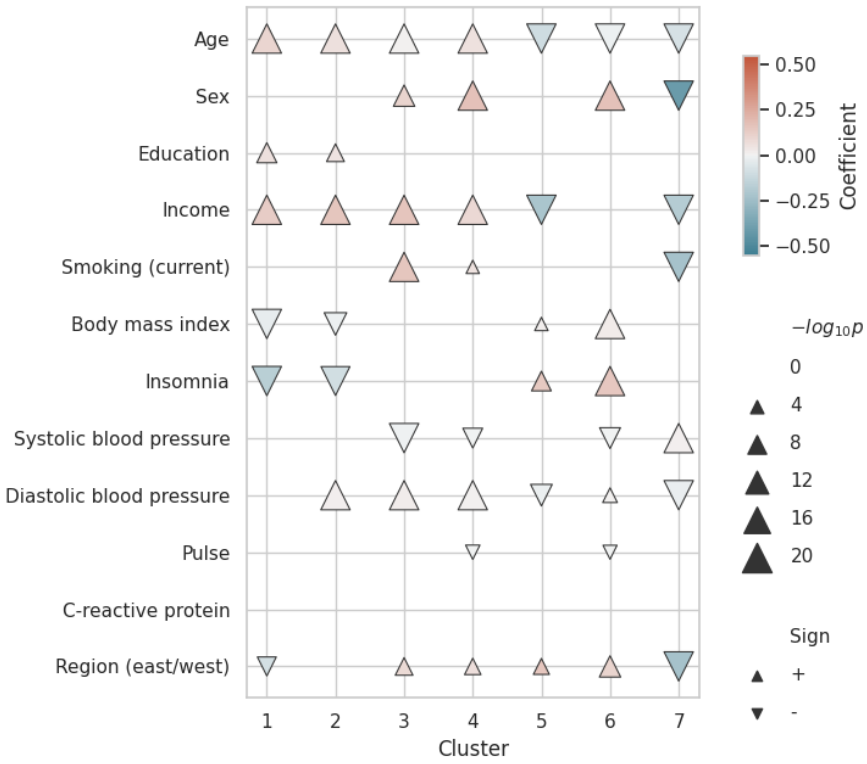



**Figure S17. Complex linear regression models involving available modifiable factors included in the THL data for each cluster (N = 23,786).** The "complex regression model" uses all available risk factors at once in a single model. The posterior log odds of being in a given cluster serves as the target variable. The direction of the triangles corresponds to the sign of the coefficient (positive - upwards; negative - downwards) and the color details the magnitude. The size of the triangles is proportional to the -log10 *p*-value, and only nominally significant values are shown (-log10*p* > 4). CRP: C-reactive protein; bp: blood pressure; sex: 1-males, 2-females.

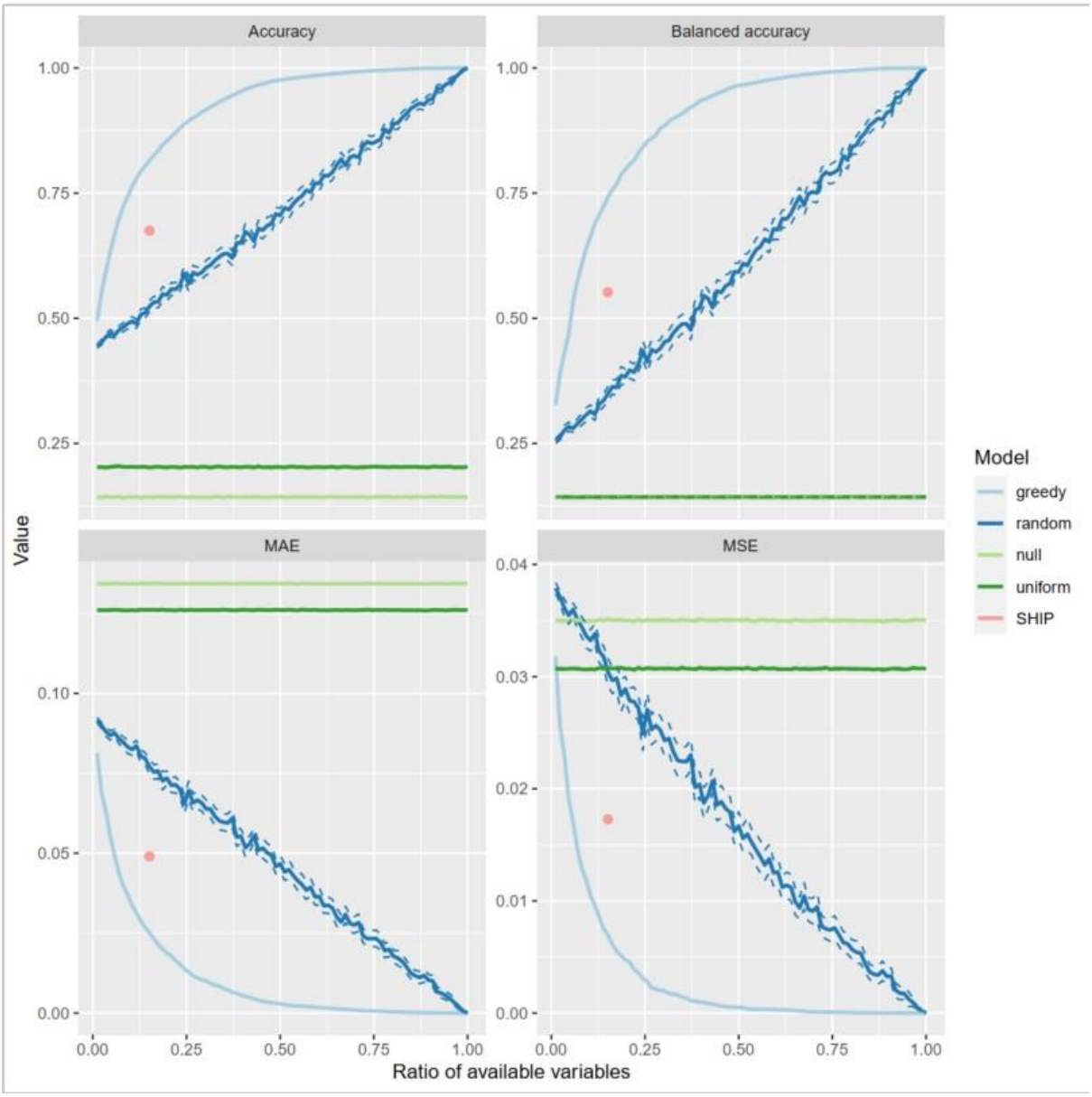
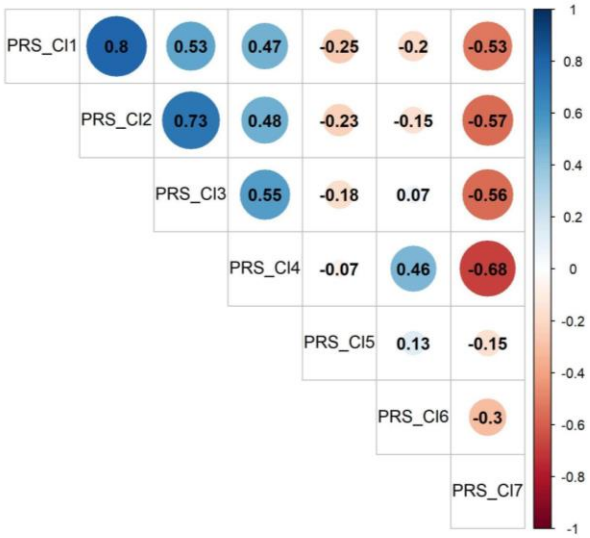



**Figure S18. Four measures that show the performance of the clustering method using only a fraction of all consensual disease variables from UKB.** The green lines show various baseline null models (a totally random (null) and a uniform cluster assignment). The dark blue line shows how a randomly chosen variable set performs, and the light blue curve shows how a greedily selected (~optimal) variable set performs. The red dot indicates the performance of the SHIP variable set. See methods for details. MAE: Mean absolute error, MSE: Mean squared error.




485

486 **Figure S19. Pearson correlation matrix for polygenic risk scores (PRS) in SHIP (N = 1126)**. PRS were calculated
487 as described in the online methods and based on MDD-related cluster GWAS from UKB. Correlation between
488 the PRS revealed the same pattern on the genetic level as has been observed in the other cohorts.

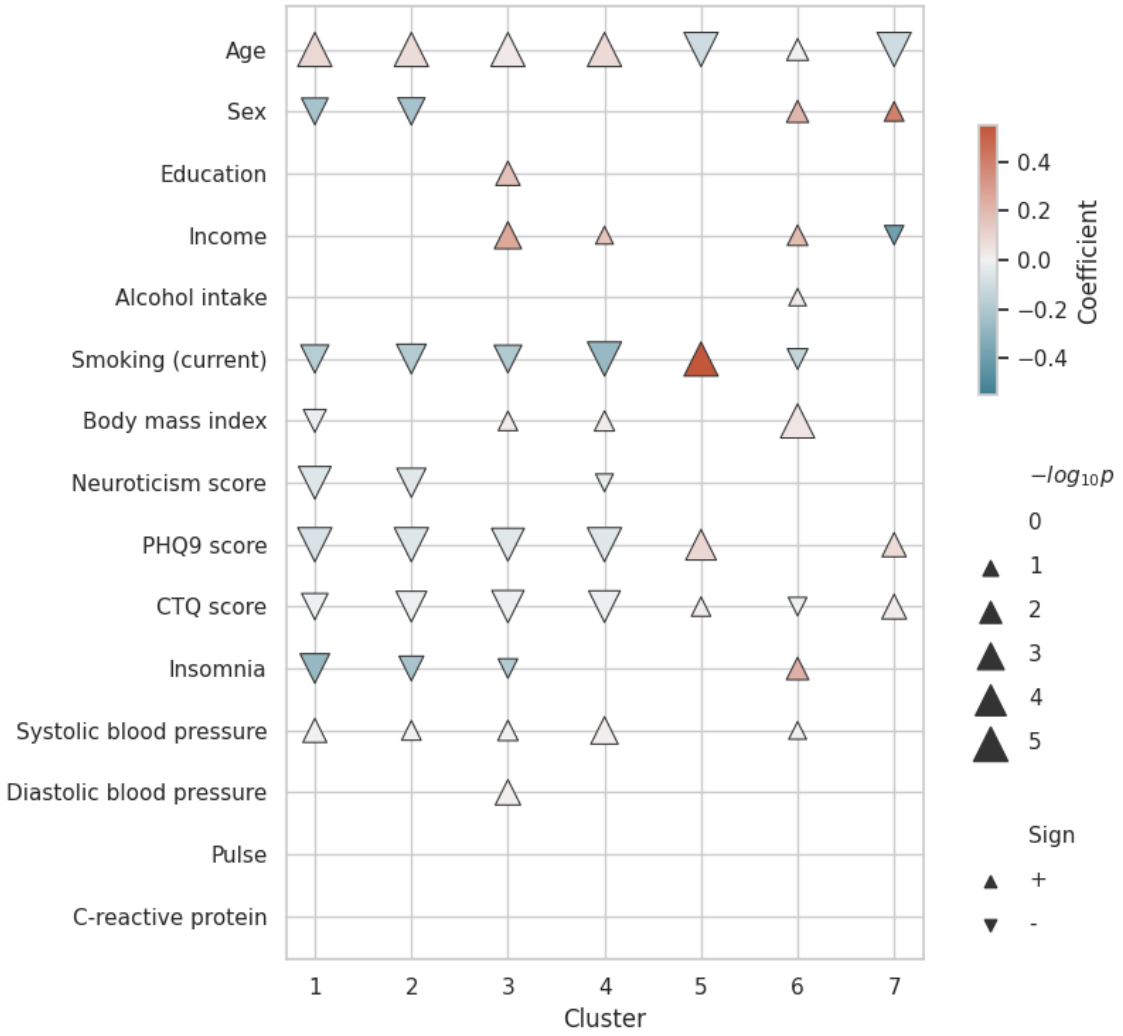



**Figure S20. Simple linear regression models involving modifiable risk factors included in the SHIP dataset for each cluster (N = 1126).** The simple linear regression model involves one factor at a time with age and sex as covariates for each cluster. The posterior log odds of being in a given cluster serves as the target variable. The direction of the triangles corresponds to the sign of the coefficient (positive - upwards; negative - downwards) and the color details the magnitude. The size of the triangles is proportional to the -log10 *p*-value, and only values $p > 0.05$ are shown ($-\log10p > 1.3$). CTQ: childhood trauma questionnaire; sex: 1-males, 2-females.