Identification of multimodal mental health signatures in the young population using deep phenotyping

Supplementary Material

Contents

Supplementary text

Supplementary Methods and Materials

Metabolomics

Sample preparation

Serum samples (100 µl) were randomly distributed over 3 96-well plates (batches). A batch of serum was created before the sample preparation and stored at −80 °C, referred to as external control (EC) samples. Both EC samples and plate-specific pools all serum samples in each batch were analyzed for quality control purposes. Each plate included eight EC, four process blanks (PB, extraction in empty well), four pooled samples, and 80 analytical samples. All solvents were LCMS-grade and were purchased from Thermo Fisher Scientific (Waltham, MA, USA). Serum samples were kept at −20 °C until extraction. On the day the samples were removed from the freezer and kept at room temperature until thawed. 300 µL of 80 % methanol was added to the sample, with subsequent shaking for 15 min at 450 rpm at room temperature, and consecutively centrifuged at 4000 rpm for 30 min at 4 °C. 15 µL of extract was pipetted into a new 96-well polypropylene plate, which was then evaporated under nitrogen for 1 h at 60 L/min, at room temperature. The samples were reconstituted in 100 µL of reconstitution solution (comprised of 5% solvent B in 95% solvent A, see Metabolomics Profiling section), shaken at 600 rpm for 15 min, and then centrifuged at 3000 rpm for 10 min at 4 °C. Afterward the samples on the plate were pooled into a single well on a deep well plate, and pipetted into the four pool positions on the plate, which was then sealed with a silicone lid and centrifuged at 3000 rpm for 5 min at 4 °C. The plate was then run on the LC-MS/MS platform. All pipetting steps were performed on a Microlab STAR automated liquid handler (Hamilton Bonaduz AG, Bonaduz, Switzerland).

Metabolomics profiling

The LC-MS/MS platform consisted of timsTOF Pro mass spectrometer with an Apollo II ion-source for electrospray ionization, Bruker Daltonics (Billerica, MA, US) coupled to a UHPLC Elute LC system, Bruker Daltonics (Billerica, MA, US). The chromatographic separation system included a binary pump, an autosampler with cooling function, and a column oven with temperature control. For infusion of the reference solution, used for external and internal mass calibration, an additional isocratic pump, Azura Pump P4.1S (Knauer, Berlin, Germany) was used. The analytical separation was performed on an Acquity HSS T3 (100 Å, 2.1 mm x 100 mm, 1.8 µm) column (Waters, Milford, MA, US). The mobile phase consisted of solvent A (99.8% water and 0.2% formic acid) and B (49.9% methanol, 49.9% acetonitrile, and 0.2% formic acid). The analysis started with 99% mobile phase A for 1.5 min, thereafter a linear gradient to 95% mobile phase B during 8.5 min followed by an isocratic condition at 95% mobile phase B for 2.5 min before going back to 99% mobile phase A and equilibration for 2.4 min. Total run time for each injection was 15 min and the analysis time for a full 96-well

plate was approximately 25 h. Samples were maintained at +15°C in the autosampler, and 5 µL were loaded to the column with a flow rate of 0.4 mL/min and a column temperature of 40 °C.

Tandem mass spectrometric analysis on the timsTOF Pro was performed in the Q-TOF mode with TIMS off, and auto MS/MS on using the following settings: ionization mode set to positive ionization, mass range set to 20 – 1100 *m/z* and a Spectra Rate of 9 Hz (Sample time 0.11s). Source settings as Capillary: 4500 V, Nebulizer Gas: 2.2 Bar, Dry Gas flow: 10 l/min, Dry Gas temperature: 220 °C. Tune settings as follows: Funnel 1 RF and Funnel 2 RF: 200Vpp, isCID: 0 eV, Multipole RF: 60 Vpp, Deflection Delta: 60 V, Quadrupole Ion Energy: 5 eV with a low mass set to 60 *m/z*, Collision Cell Energy set to 7 eV with a pre-Pulse Storage of 5 µs. Stepping is used in Basic Mode with a Collison RF from 250 – 750 Vpp, Transfer Time 20 – 50 µs, and Timing set to 50% for both. For MS/MS only the collision energy ranges from 100% - 250% with timing set to 50% for both. Auto MS/MS was used with a predefined Cycle Time of 0.5 s, Active Exclusion was used with Exclusion after 3 Spectra, and a Release time set to 0.15 min. Dynamic MS/MS spectra acquisition was applied with a target intensity of 20 000 counts, max MS/MS spectra acquisition of 30 Hz (0.03 sec), and min MS/MS spectra acquisition of 16 Hz (0.06 sec). Sodium formate clusters were applied for instrument mass calibration and for internal recalibration of individual samples. A Precursor Exclusion list was used with Exclusion of mass range of 20-60.

Metabolomics preprocessing

Bruker .d files were exported to the .mzML format using ProteoWizard's MSConvert10 and subsequently preprocessed using the Ion Identity Network workflow in MZmine[1,2] (version 3.3.0). Mass lists were created by considering mass spectra with retention times of 0.4-14 minutes and retaining MS1 intensities above 3E2 and MS2 intensities above 0. The chromatogram was built through the ADAP chromatogram builder by using the following parameters, minimum group size of scans: 5, group intensity threshold: 3E2, minimum highest intensity: 9.0E2, and *m/z* tolerance: 0.002 *m/z* or 5 ppm. The chromatogram was smoothed with a filter width of 5. The local minimum search algorithm was used for deconvolution with parameters set to, chromatographic threshold: 85%, minimum RT range (min): 0.05, minimum relative height: 0%, minimum absolute height: 9.0E2, min ratio of peak top/edge: 2.2, peak duration range (min): 0.05-0.5. The peaks were deisotoped by using the isotopic peak grouper function, with parameters set to, *m/z* tolerance: 0.002 *m/z* or 5 ppm, retention time tolerance: 0.15 min, monotonic shape: on, maximum charge: 2, representative isotope: most intense. Peaks from all samples were aligned, by using the join aligner function with parameters set to, *m/z* tolerance: 0.002 *m/z* or 5 ppm, retention time tolerance: 0.15 min, weight for *m/z*: 75, weight for retention time: 25. Rows were then filtered using the duplicate peak filter with the new average filter mode and *m/z* tolerance set to 0.001 *m/z* or 5 ppm and RT tolerance 0.03 min. Gap-filling was performed using the same mz and RT range gap filler, with a *m/z* tolerance of 0.002 *m/z* or 5ppm and an RT tolerance of 0.03 minutes. The metaCorrelate function was used to find correlating peak shapes with parameters set to, RT tolerance: 0.1 min, min. height: 9.0E2, noise level: 5E2, min samples in all: 2 (abs), min samples in group: 0 (abs), min %-intensity overlap: 60%, exclude estimated

features (gap-filled): on. Parameters for the correlation grouping were set as follows, min data points: 5, min data points on edge: 2, measure: Pearson, min feature shape correlation: 85%. Ion identity networking parameters were set to, *m/z* tolerance: 0.002 *m/z* or 5 ppm, check: one feature, min. height: 9.0E2E3 with ion identity library parameters set to, MS mode: positive, maximum charge: 2, maximum molecules/cluster: 2, adducts: M+H, M+Na, M+K, modifications: M-H2O, M-NH3. Further ion identity networks were added with *m/z* tolerance: 0.002 *m/z* or 5 ppm, min. height: 9.0E2, and ion identity library parameters set to, MS mode: positive, maximum charge: 2, maximum molecules/cluster: 6, adducts: M+H, M+Na, modifications: M-H2O, M-2H2O, M-3H2O, M-4H2O, M-5H2O, and *m/z* tolerance: 0.002 *m/z* or 5 ppm, min. height: 1E3, and annotation refinement on with parameters set to, delete smaller networks: link threshold: 4, delete networks without monomer: on, and ion identity library parameters set to MS mode: positive, maximum charge: 2, maximum molecules/cluster: 2, adducts: M+H, M+Na, M+K, modifications: M-H2O, M-NH3. Finally, two feature tables were exported in the .csv format. One feature table contains all extracted mass spectral features, and another feature table filtered for mass spectral features with associated fragmentation spectra (MS2). An aggregated list of MS2 fragmentation spectra was exported in the .mgf format and submitted to ion identity feature-based mass spectral molecular networking through the Global Natural Products Social Molecular Networking Platform (GNPS)[3,4].

Before statistical analysis, connected ion adducts were merged and mass spectral feature signals with a relative intensity less than 5 times the mean relative intensity in all paper blank samples were removed. Metabolite features present in less than 25% of the samples were removed and features present in fewer than 75% were treated as binary variables (present or absent). This resulted in a final dataset with a total of 1076 metabolite features measured, among which 433 features were continuous and 643 were binary variables. Missing values for metabolite features with continuous measurements were further subjected to imputation and batch correction procedures. Among the 433 metabolite features, 214 (49%) had less than 5% missing values. Missing values were imputed using missForest[5], with the maximum number of iterations set to 10 and the number of trees to 100. Batch correction was performed by centering and univariance scaling of each metabolite per batch.

Quality control procedures

Quality control procedures for metabolite profiling are divided into three main categories, system suitability test (SST), batch evaluation, and post-processing quality control.

In the SST, the mass spectral and chromatographic performance was evaluated prior to each batch by injecting two different standard samples. Standard sample A consisted of leucine enkephalin (1.8 μM in 50/50: H2O/ACN) and standard sample B consisted of a mix of amino acids and acylcarnitines in 50/50: H2O/ACN (Cambridge Isotope Laboratories, Tewksbury, MA, USA). System suitability was evaluated based on retention time deviation (<0.2 min), mass accuracy (<2 ppm), and relative standard deviation (<20%) for all compounds

in both standard samples A and B. Batch evaluation was performed by monitoring sixteen quality control metabolites in pooled sample extracts, EC samples and paper blanks. Potential carry-over is controlled by ensuring that quality control metabolites are not present in the paper-blank samples. Mass spectral and chromatographic performance is evaluated by monitoring retention time deviation (<0.2 min), mass accuracy (<2 ppm), and coefficient of variation (<20%) in EC samples and pooled sample extracts. Data for batch evaluation is presented in **Table S16**. Feature picking for the SST and batch evaluation was performed in Metaboscape (Bruker, Billerica, MA, United States).

Metabolite identification

To annotate mass spectral features to putative chemical structures, a mass spectral molecular network was created through the GNPS Platform (http://gnps.ucsd.edu) using the ion identity feature-based molecular networking workflow (https://ccms-ucsd.github.io/GNPSDocumentation/fbmn-iin/)[1,3,4]. The data was filtered by removing all MS/MS fragment ions within +/- 17 Da of the precursor *m/z*. MS/MS spectra were window-filtered by choosing only the top 6 fragment ions in the +/- 50 Da window throughout the spectrum. The precursor ion mass tolerance was set to 0.02 Da and an MS/MS fragment ion tolerance of 0.02 Da. A network was then created where edges were filtered to have a cosine score above 0.7 and more than 4 matched peaks. Further, edges between two nodes were kept in the network if and only if each of the nodes appeared in each other's respective top 10 most similar nodes. Finally, the maximum size of a molecular family was set to 100, and the lowest-scoring edges were removed from molecular families until the molecular family size was below this threshold. The spectra in the network were then searched against all GNPS' spectral libraries. The library spectra were filtered in the same manner as the input data. All matches kept between network spectra and library spectra were required to have a score above 0.7 and at least 4 matched peaks. Furthermore, *in silico* structural annotation prediction was performed using Sirius+CSI:FingerID[6]. Chemical class annotations were performed using deep neural networks in CANOPUS[7] and followed the ClassyFire chemical ontology[8].

Supplementary Tables

See **Extended Data**

Supplementary Figures

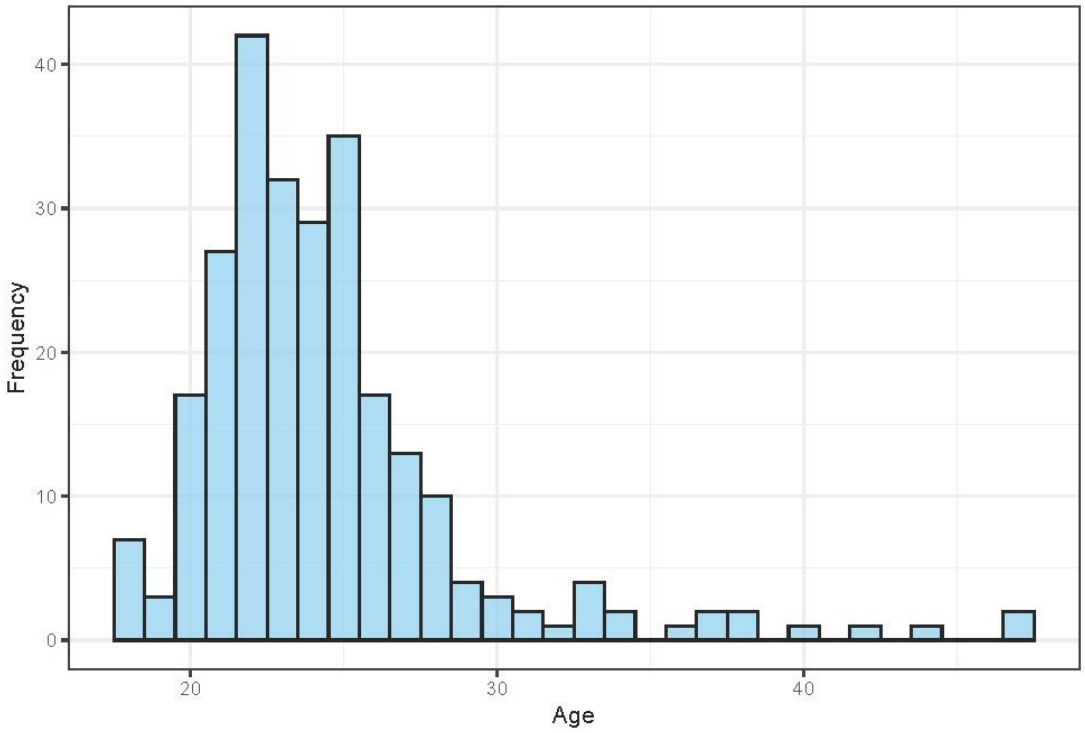
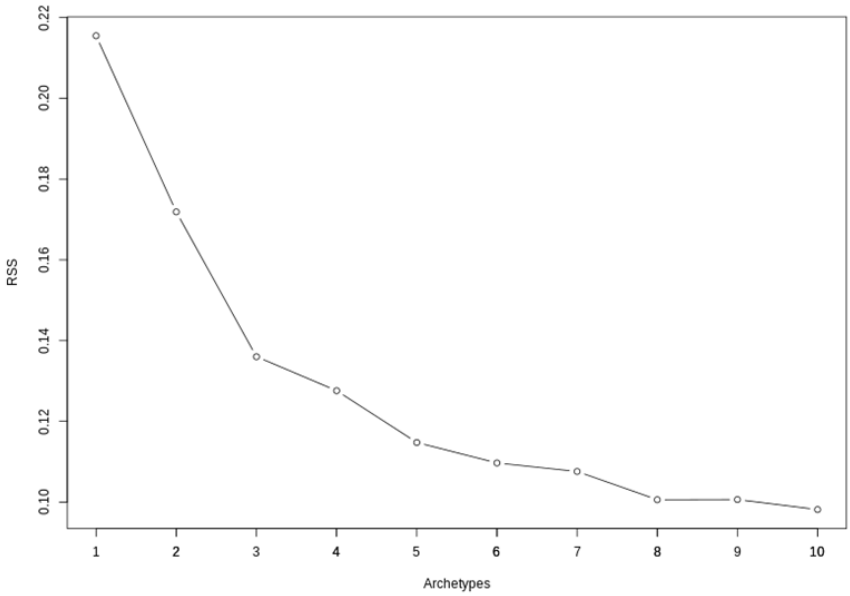


Figure S1 | Age distribution of the sample. This histogram shows the age distribution of the study participants. The x-axis represents age in years, and the y-axis shows the frequency of individuals within each age group. The distribution is centered around the early to mid-twenties, which aligns with the average age of onset for the psychiatric disorders measured in the sample.

A



B

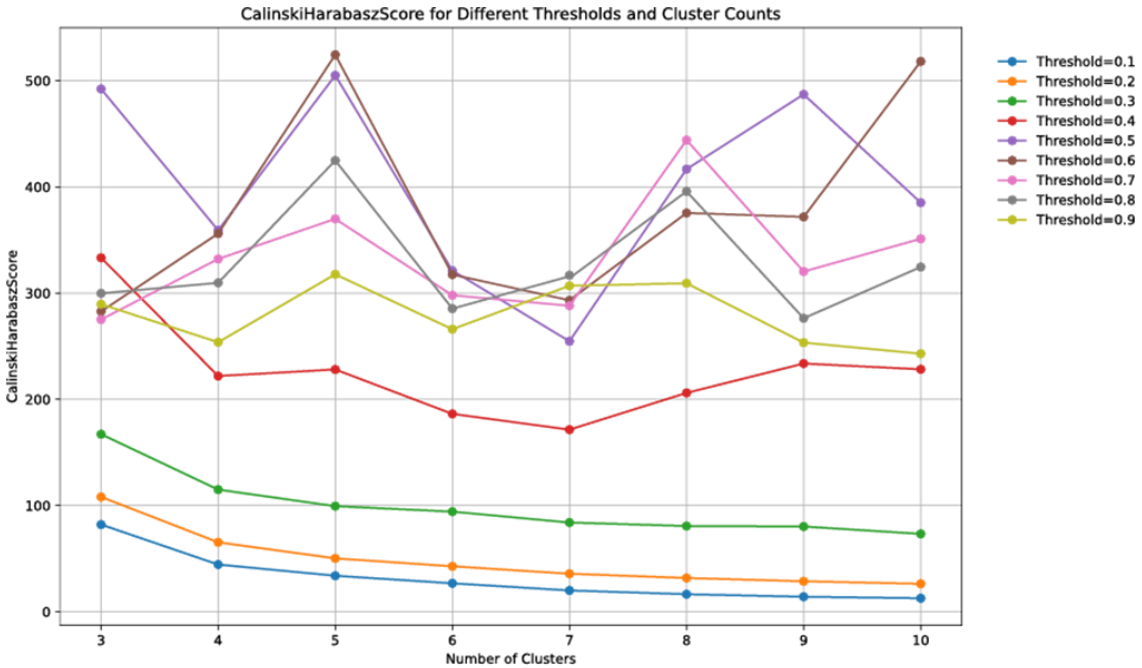


Figure S2 | Model Selection and Cluster Validation for Archetype Analysis. **(A)** Scree plot of the minimized residual sum of squares (RSS) for k archetypes (k = 1 to 10) across 100 restarts of the archetype algorithm. The plot shows a plateau in intra-cluster variance reduction at k = 5. **(B)** The Calinski-Harabasz index supports this, indicating similar model fitness at thresholds of 0.5 and 0.6.

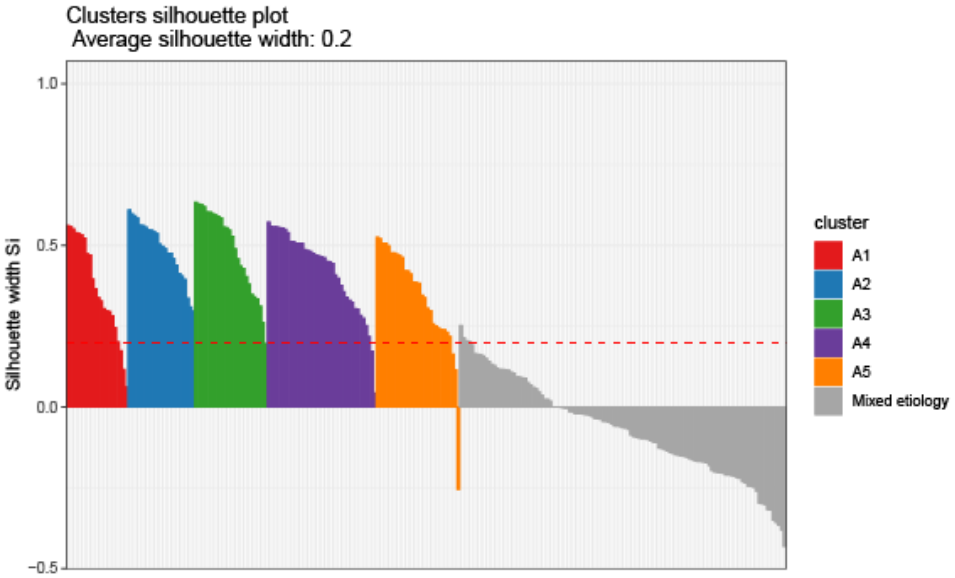


Figure S3 | Silhouette Analysis of Clustering for Archetype Memberships. Silhouette analysis of clustering for individuals with extreme archetype values and the mixed etiology group. The analysis reveals that individuals with extreme archetype memberships (greater than 0.5) for each of the five archetypes are well-clustered, whereas the mixed etiology group does not form a cohesive cluster.

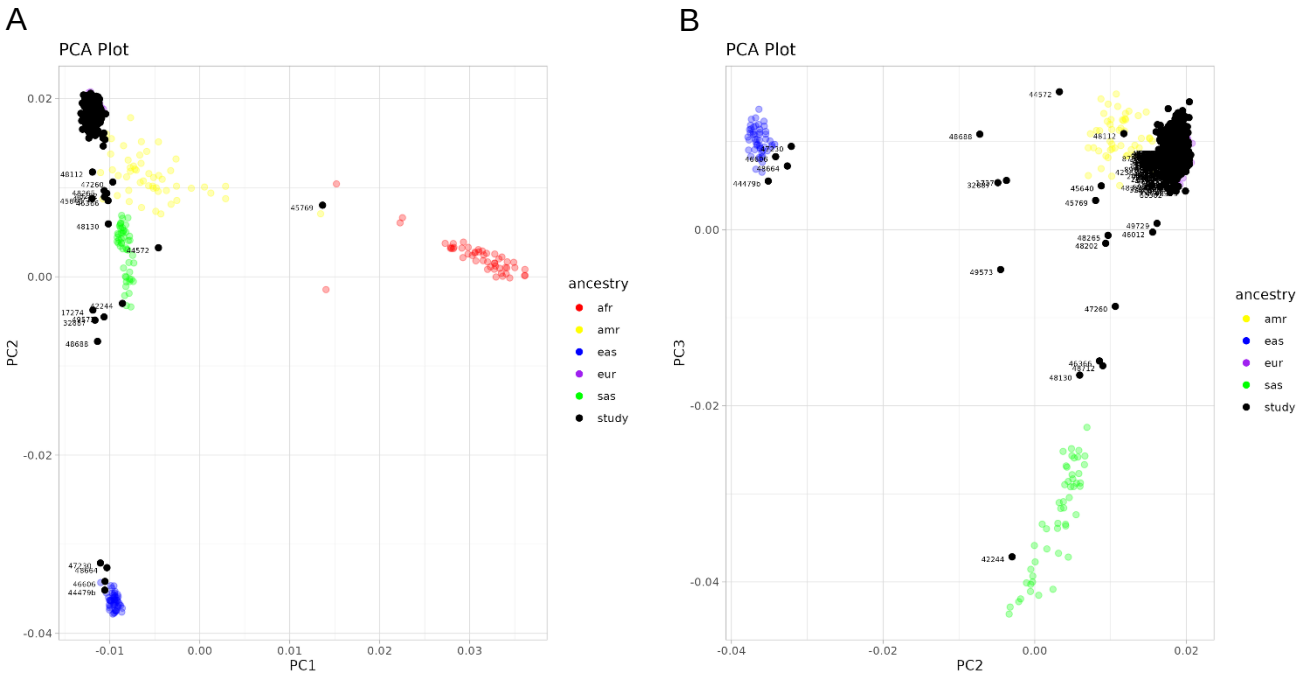
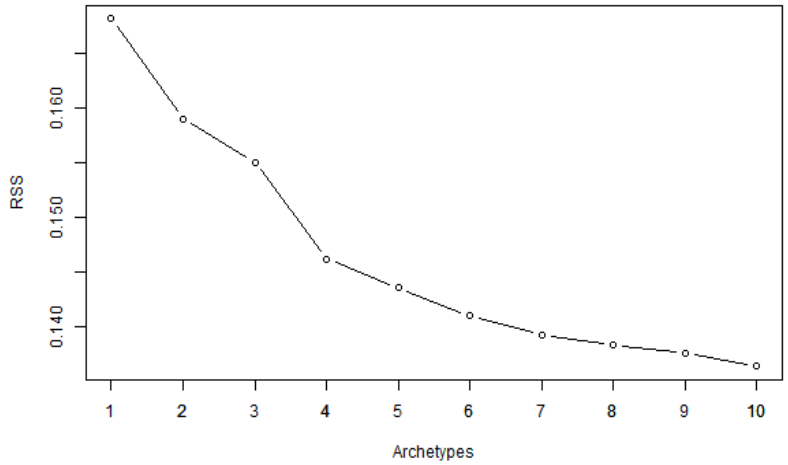


Figure S4 | Principal Component Analysis of Genetic Populations. Plots of the first three principal components from LDAK (A: PC1 and PC2, B: PC2 and PC3). Solid black circles represent 265 individuals from the Danish subset of CA18106 that donated DNA for the study. Samples from the 1000 genome project belonging to various genetic populations are presented with different colored circles: red = African (AFR), yellow = Ad Mixed American (AMR), blue = East Asian (EAS), purple = European (EUR), and green = South Asian (SAS).

A



B

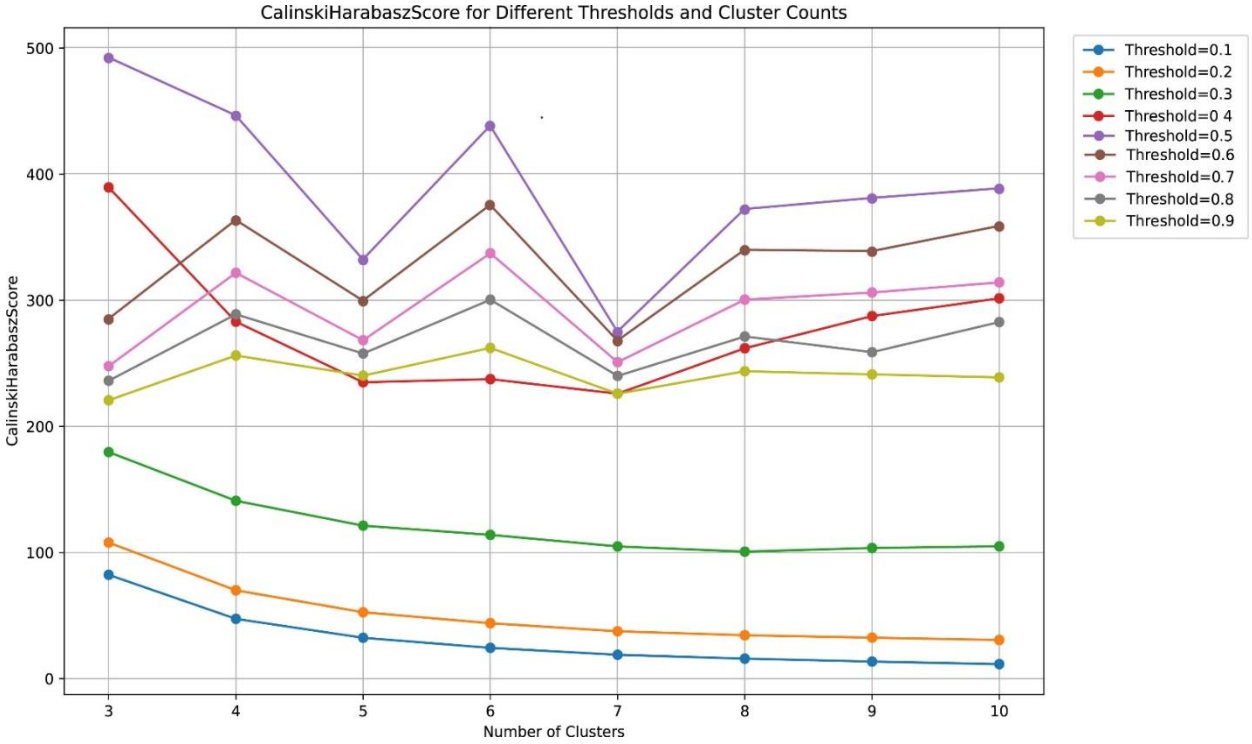


Figure S5 | Model Selection and Cluster Validation for PGS Archetype Analysis. **(A)** Scree plot of the minimized residual sum of squares (RSS) for k PGS archetypes (k = 1 to 10) across 100 restarts of the archetype algorithm. The plot shows a plateau in intra-cluster variance reduction at around k = 4. **(B)** The Calinski-Harabasz index supports this, indicating optimal model fitness at k = 4 and a threshold of 0.

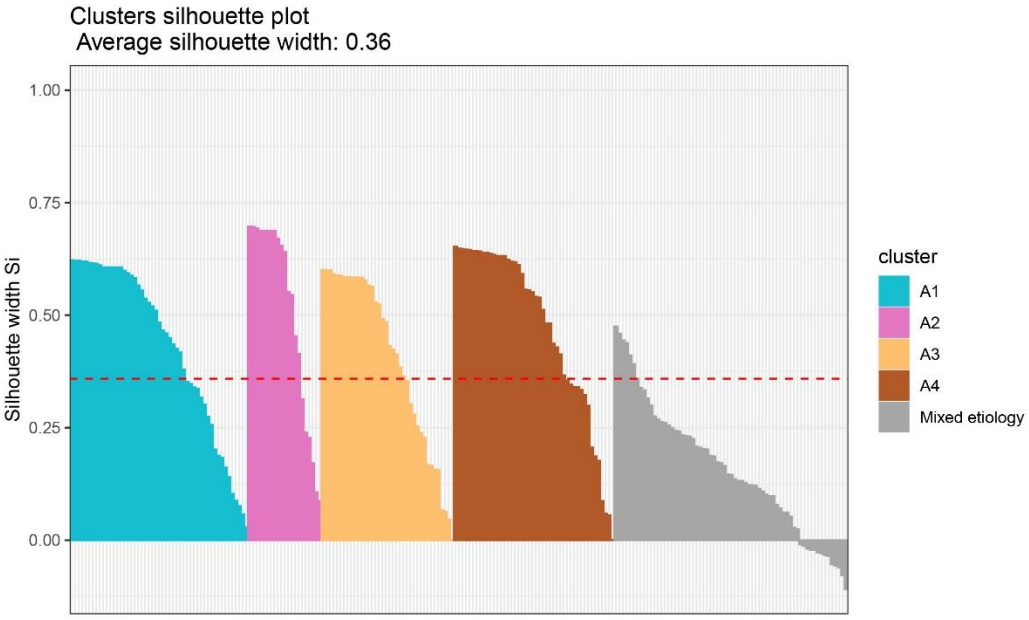


Figure S6 | Silhouette Analysis of Clustering for PGS Archetypes. Silhouette analysis of clustering for individuals with extreme archetype values and the mixed etiology group. The analysis reveals that individuals with extreme archetype memberships (greater than 0.5) for each of the six archetypes are well-clustered, whereas the mixed etiology group does not form a cohesive cluster.

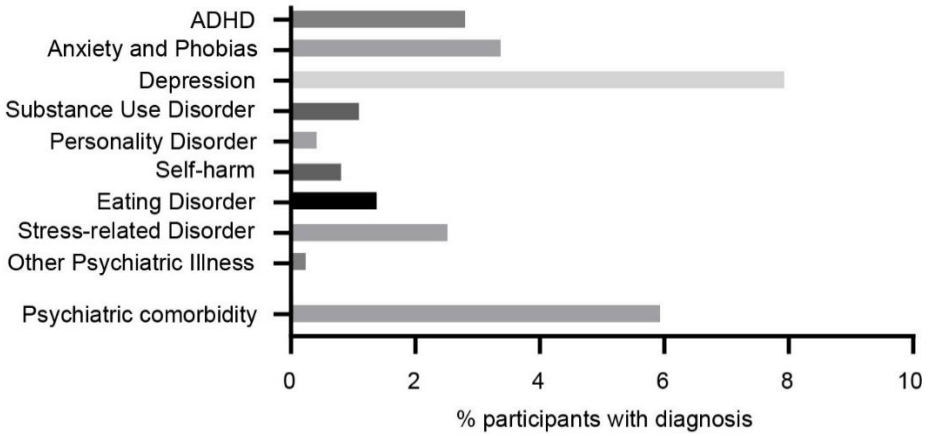


Figure S7 | Prevalence of self-reported MDx diagnoses within the sample. The y-axis represents different MDx diagnoses, while the x-axis shows the proportion or number of individuals reporting each diagnosis. This distribution highlights the frequency of various MDx conditions within the sample population.

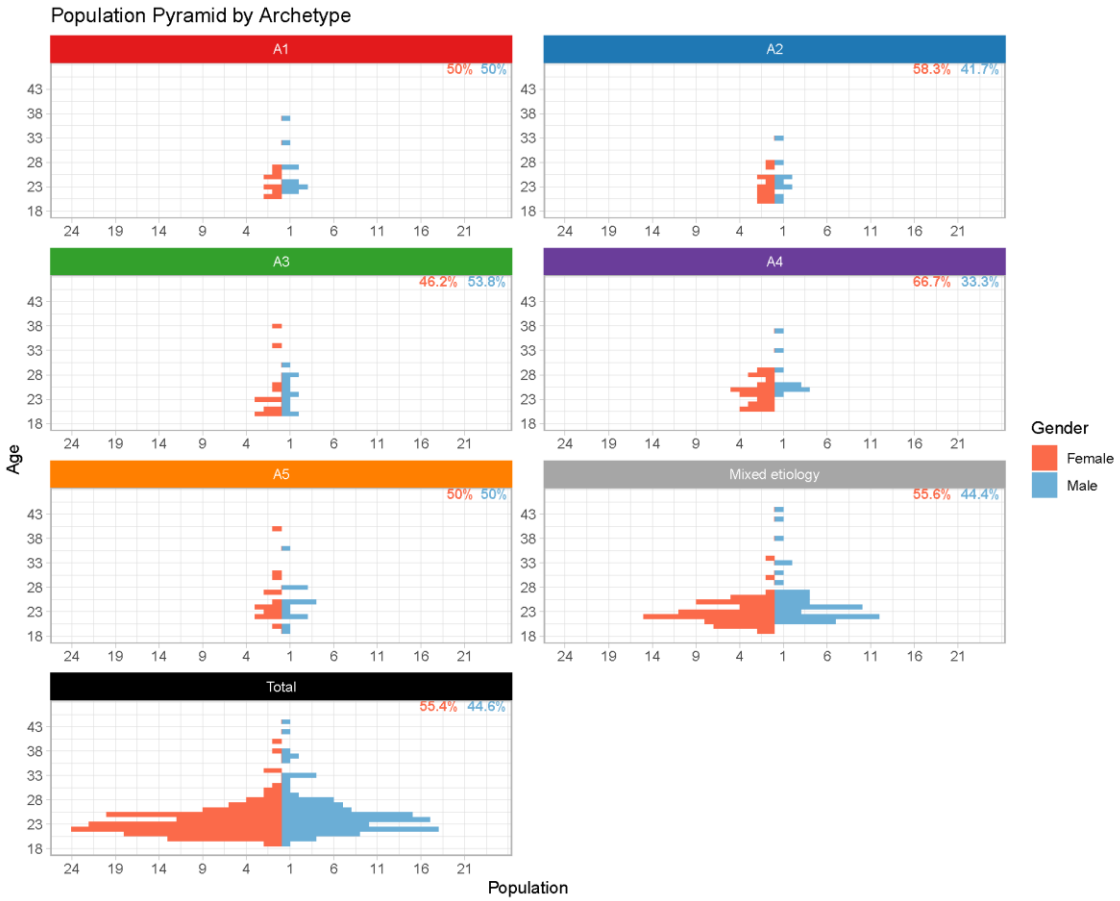


Figure S8 | Population Pyramid of Age Distribution by Archetype and Gender. This population pyramid displays the age distribution of males and females across each archetype, as well as for the total cohort. The y-axis represents age groups, and the x-axis shows the number of individuals, separated by gender. The plot provides a comparative view of age distribution patterns within each archetype and across the overall sample.

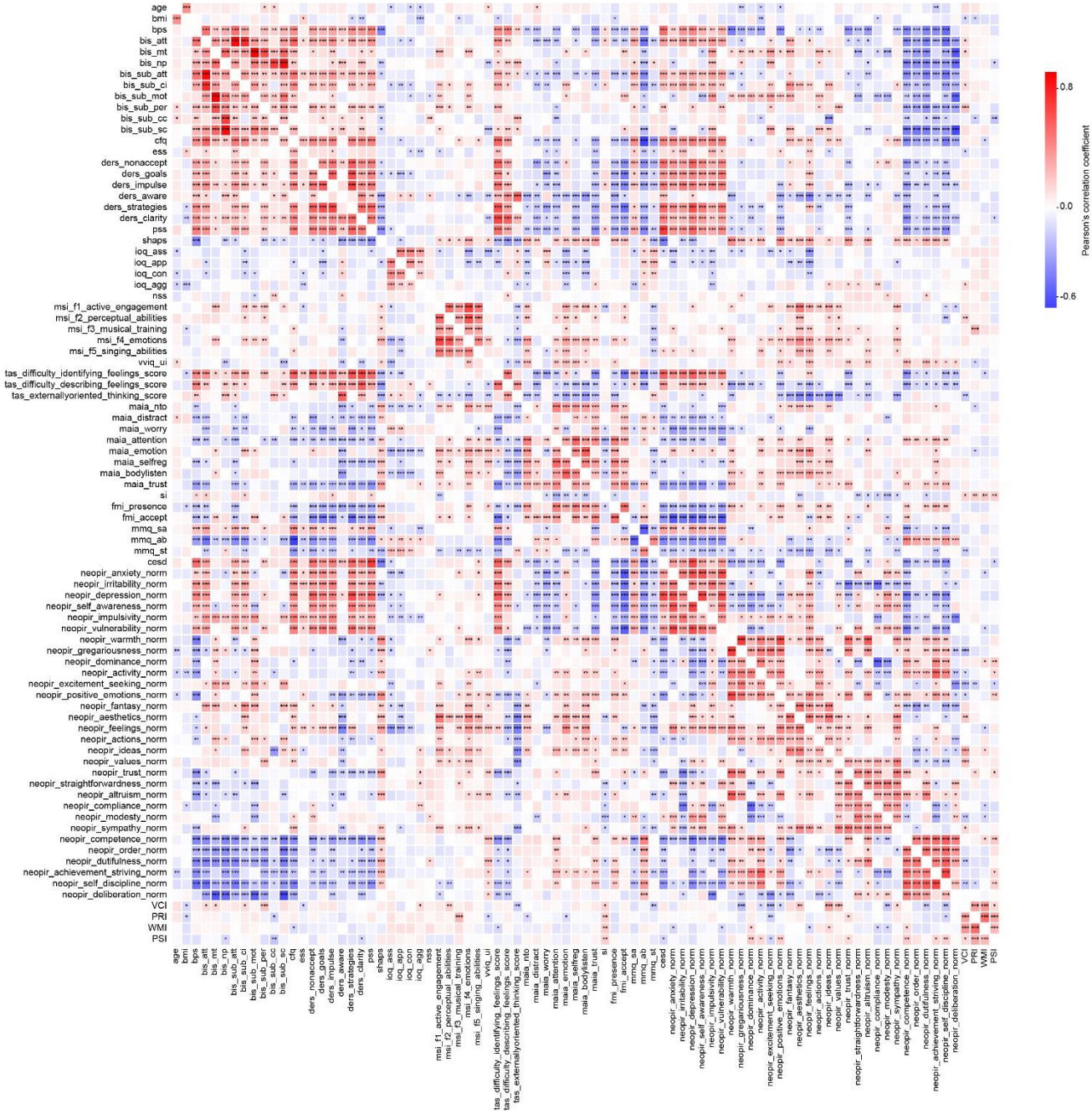


Figure S9 | Correlation Heatmap of Psychometric Variables. This heatmap displays Pearson's correlation coefficients between psychometric variables, with shades of red and blue indicating the strength and direction of the correlations. Significant correlations are marked with asterisks: * for p<0.05, ** for p<0.01, and *** for p<0.001. Darker shades represent stronger correlations, with red indicating positive correlations and blue indicating negative correlations.

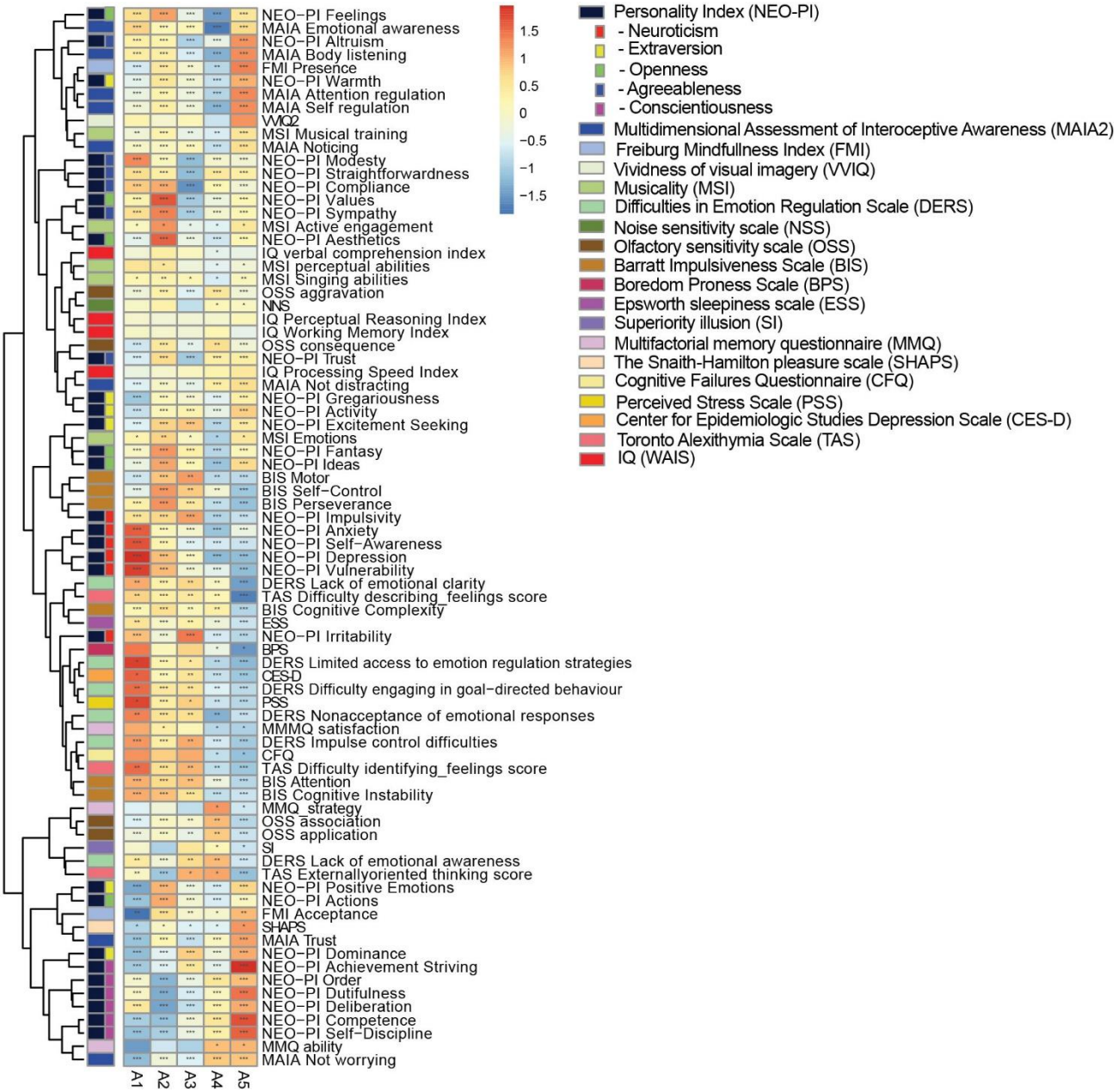


Figure S10 | Heatmap of Associations Between Archetypes and Clustering Variables. This heatmap shows the associations between archetypes and clustering variables, with shades of red and blue indicating the strength and direction of the associations. Darker shades of red represent stronger positive associations, while darker shades of blue indicate stronger negative associations. The clustering variables are indicated in the legend and are color-coded to match the heatmap, providing a clear reference for interpreting the associations. Significant correlations are marked with asterisks: * for $p<0.05$, ** for $p<0.01$, and *** for $p<0.001$.

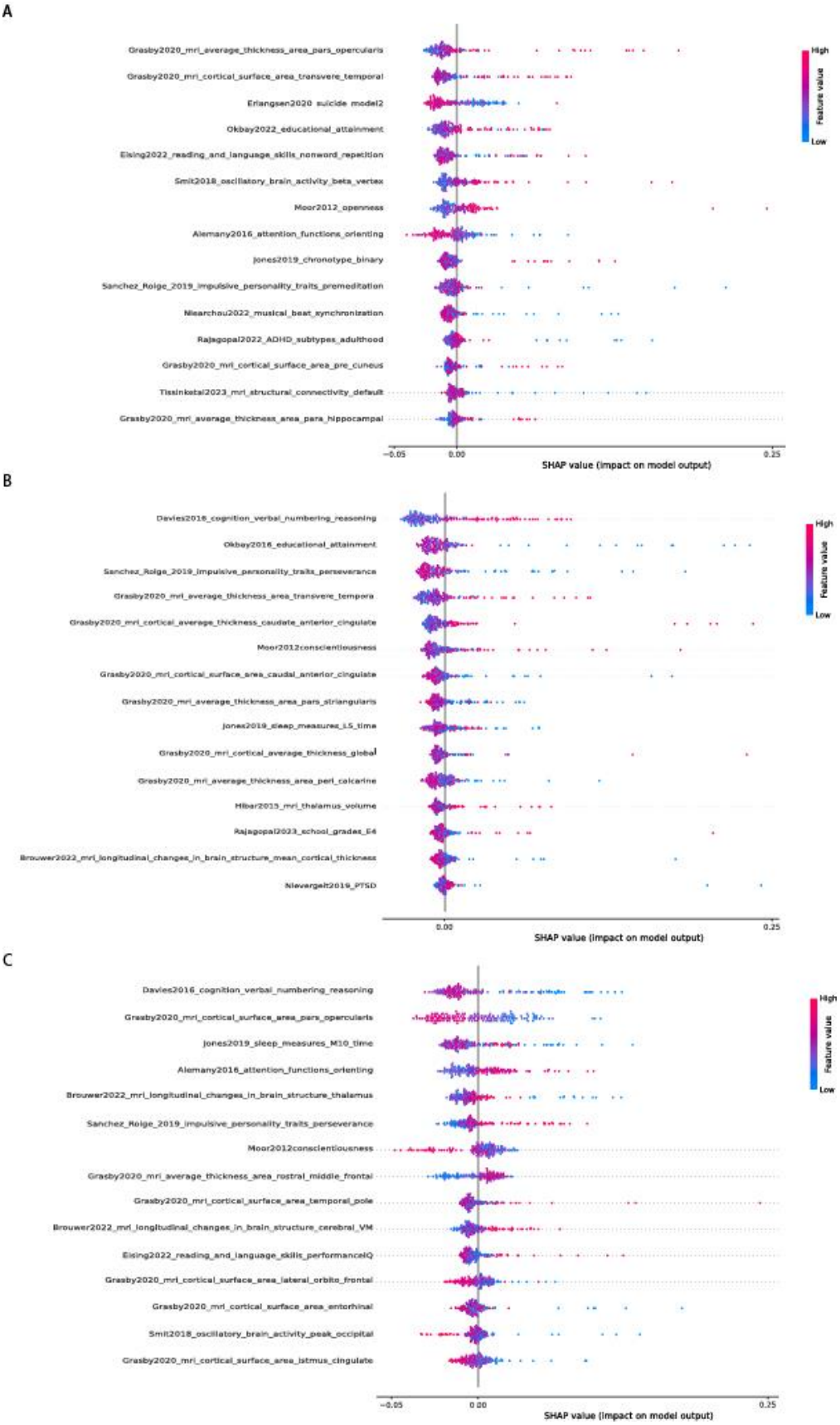


Figure S11 | SHAP Analysis for PGS-Based Archetypes. SHAP (SHapley Additive exPlanations) analyses for the following PGS-based archetypes are shown: (**A**) A2 Archetype; (**B**) A3 Archetype; (**C**) A4 Archetype; Each panel displays the contribution of individual features to the model predictions for the respective archetype, providing insight into the importance of different variables in defining each archetype.

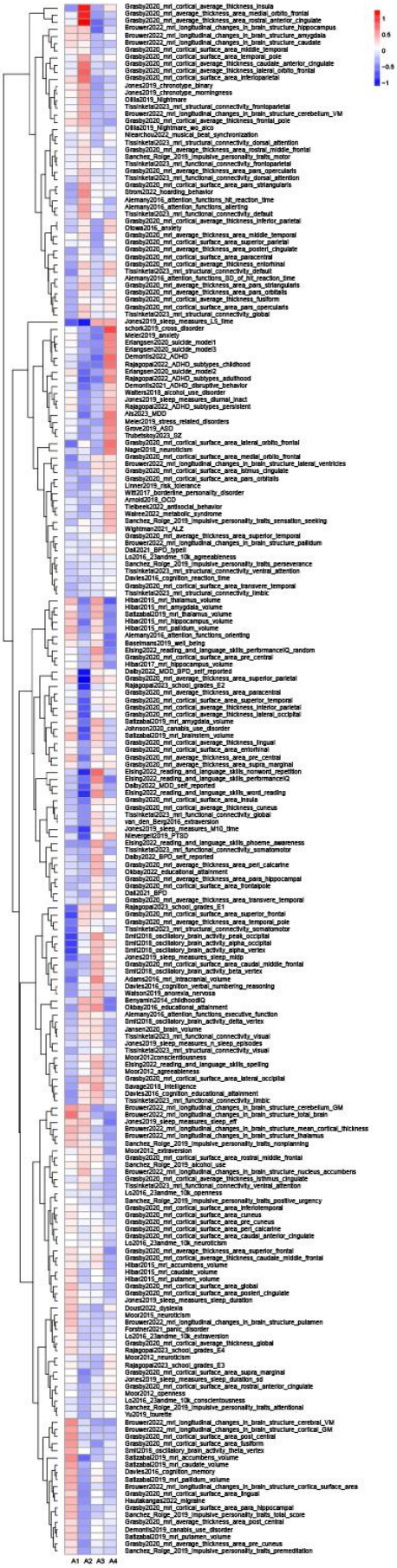


Figure S12 | Heatmap of Associations Between PGSs and the Four Identified PGS-Based Archetypes. This heatmap displays the associations between PGSs and the four identified PGS-based archetypes. All variables were rank-normally transformed. Shades of red and blue indicate the strength and direction of the associations, with red representing stronger positive associations and blue representing stronger negative associations. None of the associations were statistically significant ($p<0.05p<0.05$).

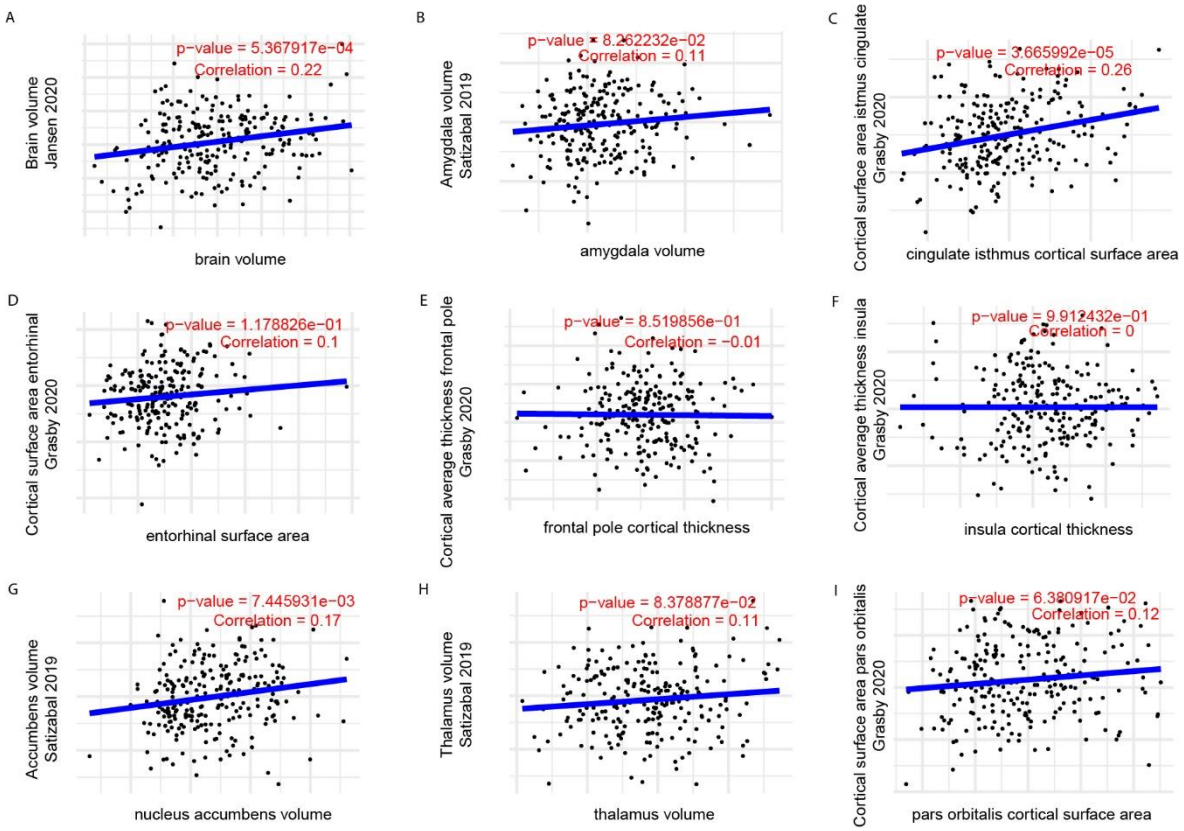


Figure S13 | Correlation Between MRI-Based Neuroimaging Traits and PGSs. This figure displays the correlation between MRI-based measures of neuroimaging traits and PGSs, calculated using data from GWAS on the same traits measured in large cohorts. Panels are as follows:(**A**) Whole Brain Volume; (**B**) Amygdala Volume; (**C**) Cortical Surface Area (Isthmus Cingulate); (**D**) Cortical Surface Area (Entorhinal); (**E**) Cortical Average Thickness (Frontal Pole); (**F**) Cortical Average Thickness (Insula); (**G**) Accumbens Volume; (**H**) Thalamus Volume; (**I**) Cortical Surface Area (Pars Orbitalis). Each panel shows the strength and direction of the correlations between the neuroimaging traits and PGSs.

Supplemental References

1. Schmid, R. *et al.* Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. *Nat. Commun.* **12**, 3832 (2021).
2. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**, 395 (2010).
3. Nothias, L.-F. *et al.* Feature-based molecular networking in the GNPS analysis environment. *Nat. Methods* **17**, 905–908 (2020).
4. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
5. Stekhoven, D. J. & Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2012).
6. Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).
7. Dührkop, K. *et al.* Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat. Biotechnol.* **39**, 462–471 (2021).
8. Djoumbou Feunang, Y. *et al.* ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.* **8**, 61 (2016).