# Supporting Information

## Chen et al. 10.1073/pnas.1308091110

### SI Methods

**Participants.** The Vietnam Era Twin Study of Aging (VETSA) project has been described previously (1). The VETSA sample was drawn from the Vietnam Era Twin (VET) Registry (2), a sample of male–male twin pairs born between 1939 and 1957 who had both served in the United States military between 1965 and 1975. The study sample is not a VA or patient group; the majority of individuals were not exposed to combat. For this analysis, 474 individual VETSA participants were included. Of those, 406 were paired (i.e., 203 twin pairs): 110 monozygotic (MZ) and 93 dizygotic (DZ) pairs. Zygosity for 92% of the sample was determined by analysis of 25 satellite markers that were obtained from blood samples. For the remainder of the sample, zygosity was determined through a combination of questionnaire and blood-group methods (3).

Mean age of the MRI participants was 55.8 (2.6) years (range, 51–59), mean years of education was 13.9 (SD = 2.1). There were 88.3% non-Hispanic white, 5.3% African-American, 3.4% Hispanic, and 3.0% "other" participants. There were no significant demographic differences between MZ and DZ twins (1,4). The VETSA sample is representative of US men in their age range based on sociodemographic and health characteristics determined by US census and Center for Disease Control data (1,4).

All participants gave informed consent to participate in the research, and the study was approved by the Institutional Review Boards of the University of California, San Diego, Boston University, and the Massachusetts General Hospital.

**Image Acquisition.** Sagittal T1-weighted MPRAGE images were acquired on Siemens 1.5 Tesla scanners (241 at University of California, San Diego; 233 at Massachusetts General Hospital). Scan parameters were: Inversion time (TI) = 1,000 ms, echo time (TE) = 3.31 ms, repetition time (TR) = 2,730 ms, flip angle = 7 degrees, slice thickness = 1.33 mm, voxel size 1.3 × 1.0 × 1.3 mm. Data were reviewed for quality, registered, and averaged to improve signal-to-noise. Of the 493 scans available at the time of these analyses, quality control measures excluded 0.6% (3 cases) due to scanner artifact and 3% (16 cases) due to inadequate image-processing results (e.g., poor contrast caused removal of nonbrain to fail). The resultant 474 available cases included 203 twin pairs (406 individuals) that were used in the present study.

**Image Processing.** The cortical surface was reconstructed to measure surface area and cortical thickness at each surface location (a total of more than 160,000 locations for each hemisphere) using a semiautomated approach provided by the FreeSurfer software (5–7). Variations in image intensity due to radio frequency (RF) coil sensitivity inhomogeneities were corrected, a normalized intensity image was created, and the skull (nonbrain) was removed from this image. A preliminary segmentation was then partitioned using a connected component algorithm, with connectivity not allowed across the established cutting planes. Interior holes in the components representing white matter were filled, resulting in a single filled volume for each cortical hemisphere. The resulting surface was covered with a polygonal tessellation and smoothed to reduce metric distortions. A refinement procedure was then applied to obtain a representation of the gray/white boundary, and the resulting surface was subsequently deformed outwards to obtain an explicit representation of the pial surface. Once generated, the cortical surface model was manually reviewed and edited for anatomical accuracy. Minimal manual editing was performed in accordance with standard, objective editing rules. Each subject's cortical surface was mapped to spherical atlas space, using a diffeomorphic registration procedure based on folding patterns (6). The surface alignment method used is not anchored to specific anatomical landmarks (e.g., fundus of the central sulcus). Rather, it uses the entire pattern of surface curvature at every vertex across the cortex to register individual subjects to atlas space (8). Then, for each subject, the standardized atlas surface tessellation was transformed into subject space based on the inverse of the subject to atlas mapping. Vertex-wise estimates of areal expansion or compression from atlas space to subject space, for each subject, were then obtained using standard FreeSurfer functions (9). The thickness of the gray matter can be computed at any point in the cortex as the shortest distance between the gray/white and pial surfaces. Finally, the vertex-wise maps were then smoothed using iterative nearest-neighbor averaging.

The optimal size of smoothing was determined empirically by reanalyzing the data with various levels of smoothing to investigate the effect of smoothing on the heritability estimates. We used three levels of smoothing from small to large corresponding to 176, 705, and 2,819 iterations, respectively, of nearest-neighbor smoothing on the standardized atlas tessellation. The 2,819-iteration was found to be of the smallest order that can yield sufficiently high heritability values to enable accurate estimation of genetic correlations, which are critical for the stability of the subsequent cluster analysis.

**Twin Analysis.** Based on our previous findings of minimal common environmental influences on surface area (10, 11) and cortical thickness (12, 13), we used a twin model that estimated contributions of additive genetic effects (A) and individual-specific environmental effects (E) to the variance in cortical thickness at each vertex. The variance–covariance patterns were examined by fitting models with Mx, a maximum-likelihood-based structural equation modeling program (14). We sought to map shared genetic effects on cortical-thickness measure between each pair of cortical locations. To accomplish this aim, univariate AE models are easily extended to the bivariate case (15). In addition to genetic and environmental sources of variance, genetic and environmental sources of covariance can also be examined in the bivariate model. In the present study, we used bivariate models to compute genetic correlations of cortical-thickness measures between each pair of vertices on the cortex. A phenotypic correlation measures shared variance; a genetic correlation measures shared genetic variance. More specifically, a phenotypic correlation is defined as the total covariance (genetic plus environmental) of two variables divided by the square root of the product of the total variance of variable 1 and the total variance of variable 2. After decomposing the sources of variance in the bivariate model, we computed genetic correlations. These are defined as the genetic covariance divided by the square root of the product of the genetic variance of variable 1 and the genetic variance of variable 2.

Before the model fitting, the cortical thickness data were adjusted for age and site effects and then normalized to adjust for global effects (i.e., the average cortical thickness was subtracted from the vertex-wise data at each cortical location). The thickness measure at each location was then standardized to z-scores across all subjects.

**Fuzzy Cluster Analysis.** Clustering methods partition the dataset into clusters based on the chosen proximity relations. We calculated pair-wise genetic correlations of thickness measures between

every two vertices on the entire cortex to generate interregional genetic correlation matrices for the left and right hemispheres simultaneously. To reduce computation time and make the cluster analysis feasible, we subsampled the standardized cortical-surface tessellation from the original 163,842–2,562 vertices per hemisphere. We then transformed the genetic correlation matrix into the distance matrix by subtracting each genetic correlation in the genetic correlation matrix from 1. The value of this distance measure ranges between 0 and 2, indicating that two objects are closely related or very different, respectively. In fuzzy clustering, objects can belong to more than one cluster and with different degrees of membership to the different clusters: between 0 (absolutely doesn't belong) and 1 (absolutely belongs). Thus, the memberships of objects at the overlapping boundaries can express the ambiguity of the cluster assignment.

The clustering procedure was performed by the *cluster* package implemented in R (www.r-project.org/). Fuzzy clustering aims to minimize the objective function

$$\sum_{v=1}^{k} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} u_{iv}^{r} u_{jv}^{r} d(i,j)}{2\sum_{j=1}^{n} u_{jv}^{r}},$$

where $n$ is the number of observations, $k$ is the number of clusters (from 2 to 12, e.g.), $r$ is the membership exponent, $u$ is the cluster membership, and $d(i, j)$ is the dissimilarity between observations $i$ and $j$ (16). The cluster memberships $u$ are nonnegative and sum to one for a given data point. To investigate the stability of the clustering in relation to initialization, we randomly initialized the algorithm for 100 runs and picked the cluster solution that maximized the likelihood function.

**Silhouette Coefficient.** Two cluster properties are usually evaluated: cohesion, which determines how closely related the objects in a cluster are, and separation, which determines how distinct or well-separated a cluster is from other clusters. Quantitative indices called silhouette coefficients combining both cohesion and separation are commonly used to approximately determine the correct number of clusters (16, 17). The silhouette coefficients can be computed by

$$s_i = (b_i - a_i)/\max(a_i, b_i),$$

where $s_i$ is the silhouette coefficient for the $i$th object; $a_i$ is the average distance between the $i$th object and all other objects in the same cluster; calculate the average distance between the $i$th object to all of the objects in a given cluster and $b_i$ is the minimum value with respect to all clusters. An overall measure of the goodness of a cluster can be obtained by computing the average silhouette coefficient of all objects. The natural number of clusters in a dataset can be determined by looking at the number of clusters at which there is a peak in the plot of the silhouette coefficients when it is plotted against the number of clusters.

**Spectral Cluster Analysis.** To validate the stability of the fuzzy clustering results, we used an entirely different clustering algorithm called spectral clustering (18, 19). Instead of directly using distance matrices like in fuzzy clustering, spectral clustering methods transform distance matrices into affinity matrices (ref. 18); for reviews, see ref. 20). Each element of an affinity matrix $A$, as defined by Ng et al., can be computed by

$$A_{ij} = \exp\left(-\left\|s_i - s_j\right\|^2/2\sigma^2\right).$$

The affinity matrix $A_{ij}$ measures the affinity between data points $s_i$ and $s_j$. In our study, the distance between $s_i$ and $s_j$ is equal to 1 minus genetic correlations. $\sigma$ is a scaling parameter. The affinity matrix is then normalized $L = D^{-1/2}AD^{-1/2}$ where $D$ is a diagonal matrix with $D_{ii} = \sum_j A_{ij}$. The dataset is now represented in an eigenspace using the top eigenvectors of the $D$ matrix (in this study, we used the top 30 eigenvectors) (18). Projecting the data into the eigenspace of the affinity matrix allows us to capture the primary distribution of the original data points. We would like to ensure that the selected eigenvectors are sufficiently relevant to inform the variation of data distribution. We have validated the relevance of the top 30 eigenvectors using an eigenvector selection algorithm to confirm that the selected eigenvectors were informative to separate data into groups (19). Including only top and relevant eigenvectors makes clustering results less susceptible to noise in the data (19).

Next, we used Gaussian mixture models to cluster the data into different groups using the new data representation in the eigenspace. Gaussian mixture models are widely used in statistics for clustering data as a model-based clustering algorithm (17). The number of clusters was approximated by evaluating the number of Gaussian components using Bayesian Information Criterion (BIC) (19). In the context of high-dimensional data, fitting Gaussian mixture models is challenging due to the large number parameters relative to the small sample size. Thus, regularization was used to improve estimation of the covariance matrices in Gaussian mixture models. Moreover, because the penalty term in the BIC will dominate the likelihood function because of the large number of parameters, we instead used a 10-fold cross-validation procedure. We used nine-tenths of the data as a training set to fit the Gaussian mixture models and then evaluated the fit of the model via the penalized likelihood on the one-tenth of the data held out as a testing set. This procedure was repeated until all data points have served as part of the testing set. The results indicated ~11–15 clusters existing in the dataset depending on the methods of regularization.

**Determining the Number of Clusters.** Here, we sought to determine the most appropriate number of clusters to explain patterns in the genetic correlation data based on the silhouette coefficient, which evaluates how appropriate a cluster solution is based on maximizing within-cluster cohesion and between-cluster separation (17). The silhouette plot, based on the clustering in the unadjusted thickness data before controlling for global effects, demonstrated a plateau starting around the 11–12 cluster solution (Fig. S3). The adjustment for global cortical thickness effects allows us to examine region-specific genetic effects. For data adjusted for average cortical thickness, the coefficient began to reach a plateau around the 12–14 cluster solutions (Fig. S4). However, it continued to increase beyond even the 20-cluster solution. We suspect that the continuous increase in the silhouette coefficient after the 20-cluster solution may be due to lower heritability estimates for each cortical location after adjusting for global effects. Some locations may have low heritability estimates, and therefore less reliable genetic correlation estimates.

To validate the number of clusters, we also used spectral clustering, which uses a fundamentally different clustering algorithm. Spectral clustering relies on the eigen-structure of the similarity matrix rather than on the original similarity matrix to partition points into clusters (18). An important advantage of the spectral algorithm is its robustness to atypical cluster features and noise in the data (18). Spectral clustering was performed for a simultaneous cluster-number estimation and data clustering based on Gaussian mixture models. Results yielded 11–15 clusters, confirming that the true number of clusters is unlikely to be more than 20. The genetic cluster maps derived from these approaches revealed very similar features (Fig. S2), providing further evidence of the converging results.

**Genetic Similarity Matrix and Dendrogram.** To visualize the genetic proximity between clusters, we plotted the pair-wise genetic correlation matrix. The rows and columns of the matrix were sorted by the cluster labels. Thus, all data points belonging to the same clusters are grouped together, and the matrix has roughly a block diagonal structure. Each cell in the genetic similarity matrix is the weighted mean of genetic correlations within and between clusters. The patterns displayed in the similarity matrix can reveal the relationship between clusters. The dendrogram was then produced by hierarchical cluster analysis.


1. Kremen WS, et al. (2006) Genes, environment, and time: The Vietnam Era Twin Study of Aging (VETSA). *Twin Res Hum Genet* 9(6):1009–1022.
2. Goldberg J, Curran B, Vitek ME, Henderson WG, Boyko EJ (2002) The Vietnam Era Twin Registry. *Twin Res* 5(5):476–481.
3. Eisen S, Neuman R, Goldberg J, Rice J, True W (1989) Determining zygosity in the Vietnam Era Twin Registry: An approach using questionnaires. *Clin Genet* 35(6): 423–432.
4. Anonymous (1999-2004) *National Health and Nutrition Examination Survey (NHANES III): Trends in Health and Aging* (National Center for Health Statistics, Hyattsville, MD).
5. Dale AM, Fischl B, Sereno MI (1999) Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* 9(2):179–194.
6. Fischl B, Sereno MI, Dale AM (1999) Cortical surface-based analysis. II. Inflation, flattening, and a surface-based coordinate system. *Neuroimage* 9(2):195–207.
7. Dale AM, Sereno MI (1993) Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: A linear approach. *J Cogn Neurosci* 5(2):162–176.
8. Fischl B, Sereno MI, Tootell RBH, Dale AM (1999) High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum Brain Mapp* 8(4): 272–284.
9. Rimol LM, et al.; Alzheimer's Disease Neuroimaging Initiative (2010) Sex-dependent association of common variants of microcephaly genes with brain structure. *Proc Natl Acad Sci USA* 107(1):384–388.
10. Eyler LT, et al. (2011) Genetic and environmental contributions to regional cortical surface area in humans: a magnetic resonance imaging twin study. *Cereb Cortex* 21(10):2313–2321.
11. Chen CH, et al. (2011) Genetic influences on cortical regionalization in the human brain. *Neuron* 72(4):537–544.
12. Rimol LM, et al. (2010) Cortical thickness is influenced by regionally specific genetic factors. *Biol Psychiatry* 67(5):493–499.
13. Kremen WS, et al. (2010) Genetic and environmental influences on the size of specific brain regions in midlife: The VETSA MRI study. *Neuroimage* 49(2):1213–1223.
14. Neale MC, Boker SM, Xie G, Maes HH (2003) *Mz: Statistical Modeling* (Department of Psychiatry, Medical College of Virginia, Richmond, VA), 6th Ed.
15. Neale MC, Cardon LR (1992) *Methodology for Genetic Studies of Twins and Families* (Kluwer, Dordrecht, The Netherlands).
16. Kaufman L, Rousseeuw P (1990) *Finding Groups in Data: An Introduction to Cluster Analysis* , Wiley, New York).
17. Tan P-N, Steinbach M, Kumar V (2006) *Introduction to Data Mining* (Addison-Wesley, Reading, MA), 1st Ed.
18. Ng A, Jordan M, Weiss Y (2002) On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems 14 (MIT Press, Cambridge, MA), pp 849–856.
19. Xiang T, Gong S (2008) Spectral clustering with eigenvector selection. *Pattern Recognit* 41(3):1012–1029.
20. Luxburg U (2007) A tutorial on spectral clustering. *Stat Comput* 17(4):395–416.


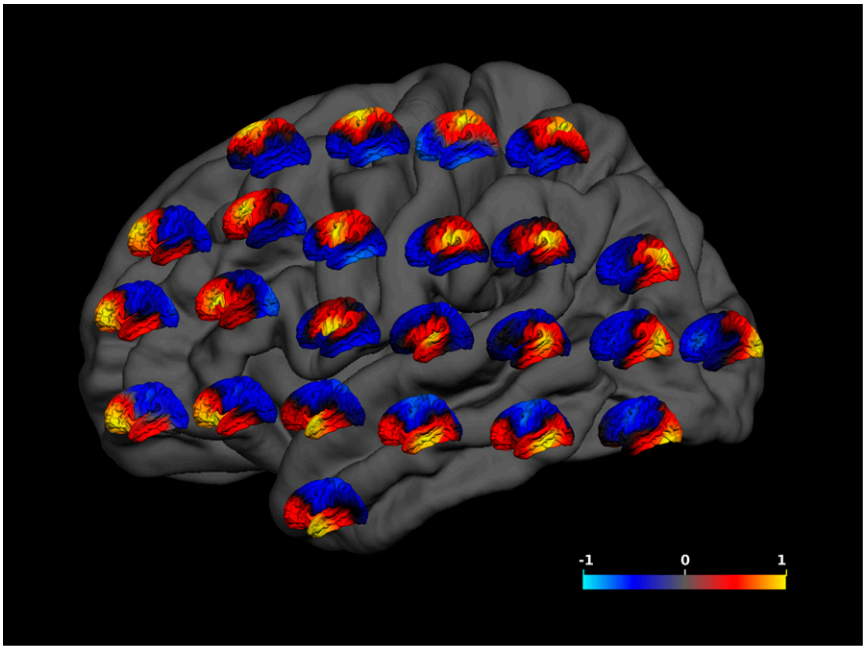


**Fig. S1.** A grid placement of genetic correlation maps from 24 different seeds. We also performed seed-point analyses to map genetic correlations of cortical-thickness measures between selected seed regions and all other cortical locations after adjusting for mean thickness averaged across the entire cortex. Genetic correlation maps are a simple way to visualize the genetic patterning because the color codes directly reflect the strength of genetic correlations between the seed regions and all other cortical points. For the selection of seed points, we used a grid of regularly spaced seeds distributed across the entire lateral aspect of one cortical hemisphere. The location of each small genetic correlation map on the gray brain map represents the location of the seed for that correlation map. Color scale indicates the strength of genetic correlations between the surface area at the seed region and at all other locations on the cortical surface; these correlations range from positive to negative.

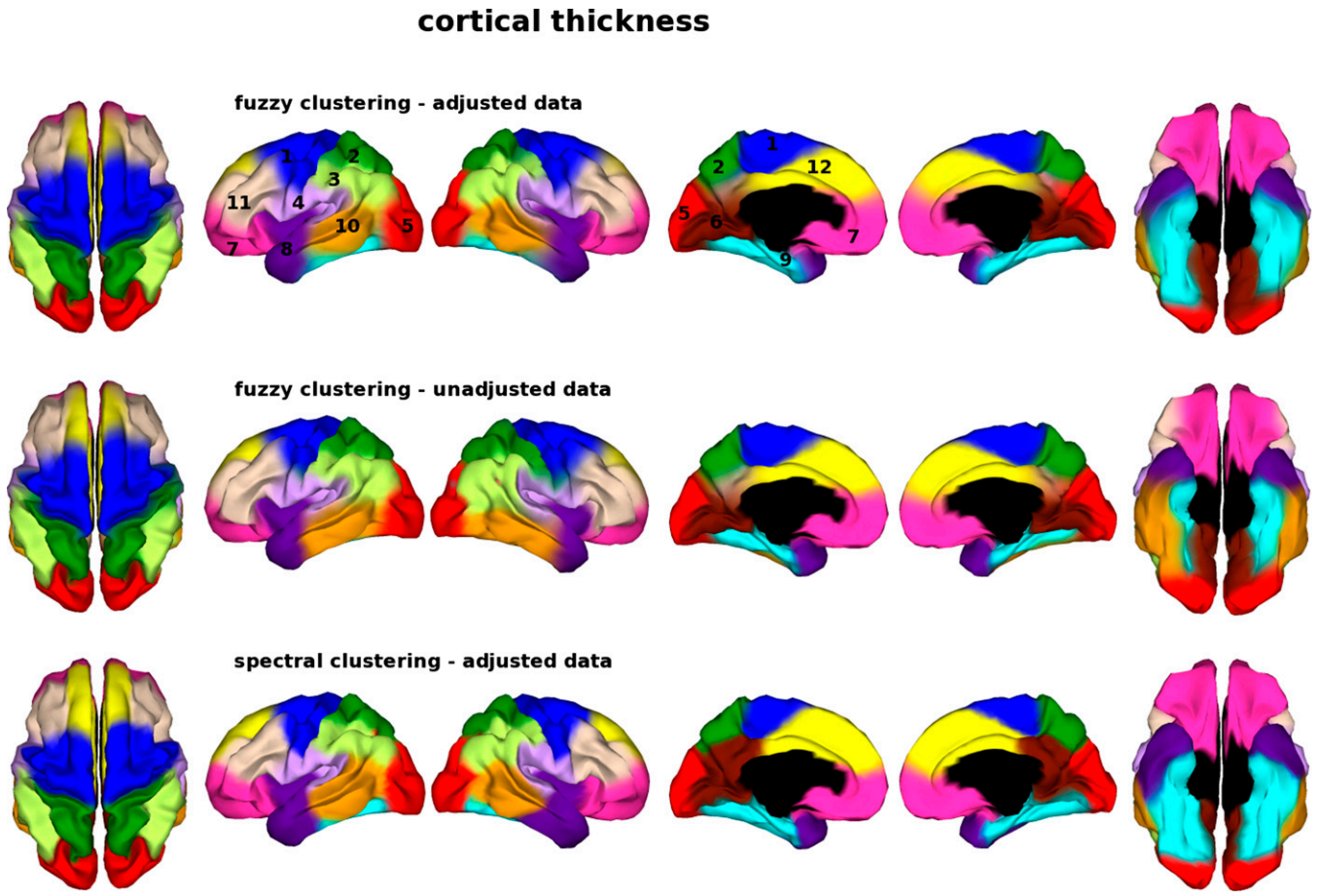


**Fig. S2.** Thickness clustering derived from different clustering approaches. (*Top*) Fuzzy clustering in the adjusted data. (*Middle*) Fuzzy clustering in the unadjusted data. (*Bottom*) Spectral clustering in the adjusted data.

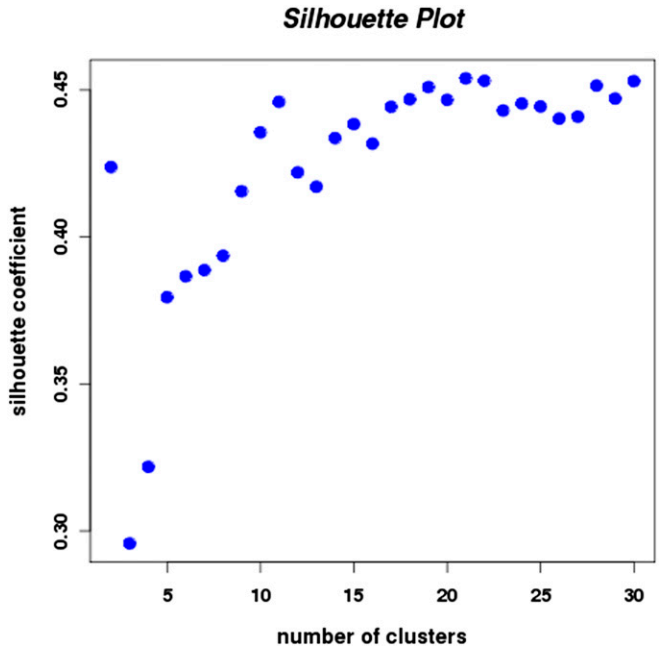


**Fig. S3.** Silhouette Plot of the genetic clustering based on unadjusted data. The plot shows the silhouette coefficients against the number of clusters. When the number of clusters reached 11–12, the silhouette coefficients started to reach a plateau.

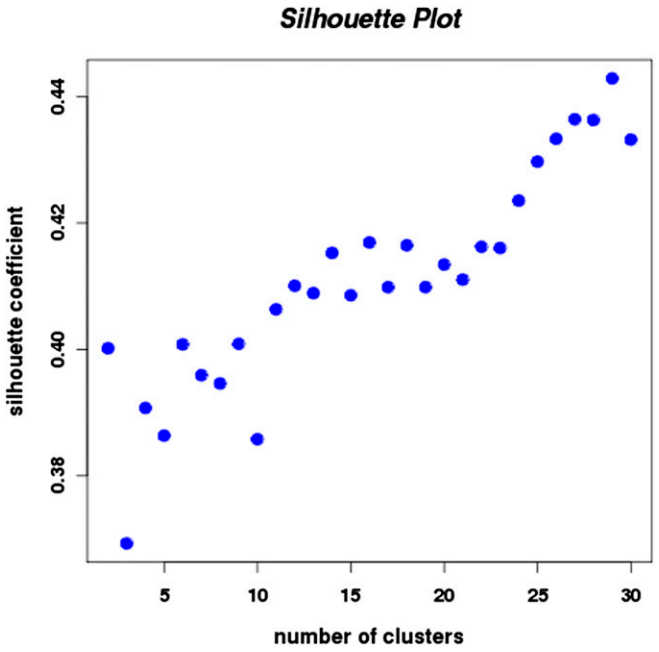


**Fig. S4.** Silhouette Plot of the genetic clustering based on adjusted data. The plot shows the silhouette coefficients against the number of clusters. When the number of clusters reached 12–14, the silhouette coefficients started to reach a plateau.

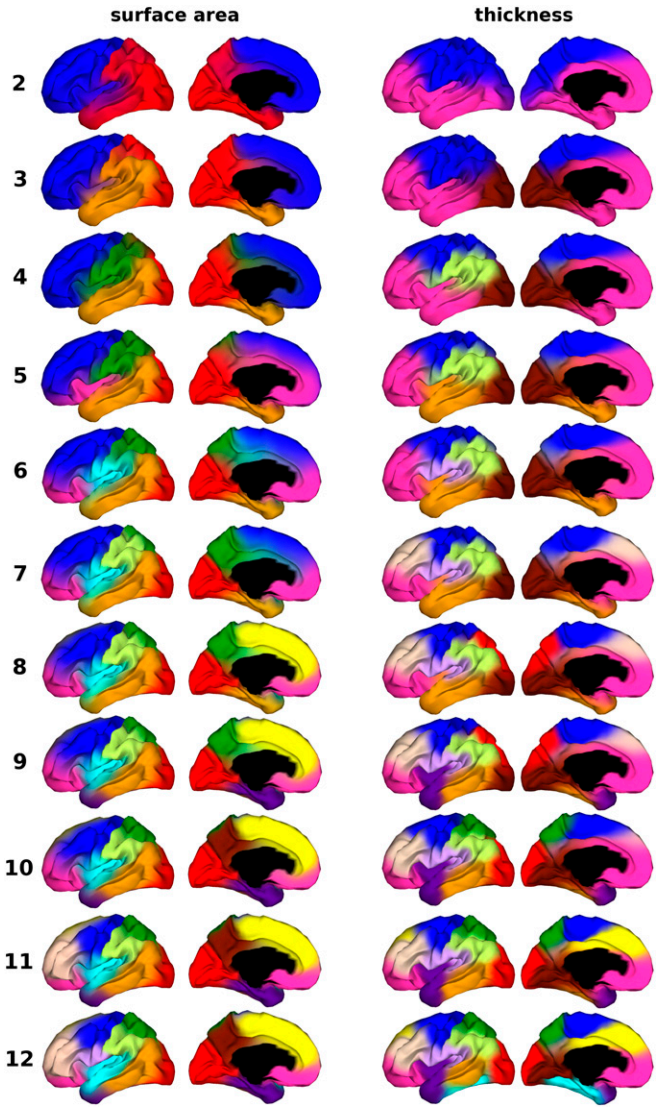


**Fig. S5.** The order of emergence of the 12 clusters. The two columns on the *Left* show the cluster maps sequentially from 2 to12 clusters for surface area. The two columns on the *Right* show the cluster maps sequentially from 2 to12 clusters for cortical thickness.